Wei Wu · Hani Choudhry *Editors*

Next Generation Sequencing in Cancer Research, Volume 2

From Basepairs to Bedsides

Springer

Next Generation Sequencing in Cancer Research, Volume 2

Wei Wu • Hani Choudhry
Editors

Next Generation Sequencing in Cancer Research, Volume 2

From Basepairs to Bedsides

Springer

Editors
Wei Wu
Department of Pathology and Laboratory Medicine, Southern Alberta Cancer Research Institute
University of Calgary
Calgary, AB, Canada

Hani Choudhry
Faculty of Science
Biochemistry Department
Center of Innovation in Personalized Medicine
King Fahd Center for Medical Research
King Abdulaziz University
Jeddah, Saudi Arabia

ISBN 978-3-319-15810-5 ISBN 978-3-319-15811-2 (eBook)
DOI 10.1007/978-3-319-15811-2

Library of Congress Control Number: 2013943582

Springer Cham Heidelberg New York Dordrecht London

Printed on acid-free paper

Springer International Publishing AG Switzerland is part of Springer Science+Business Media (www.springer.com)

Preface

Cancer is a complex and heterogeneous disease with alterations of the genome that accumulate over time or, in some cases, occur as a one-time, catastrophic shattering and rearrangement of chromosomes (chromothripsis) after an exposure to radiation or genotoxic chemicals. Cancer is now a leading noncommunicable cause of death worldwide. Overall, for example, there were 14.1 million new cases and 8.2 million deaths in 2012 (GLOBOCAN 2012), and this number is projected to continue to rise—particularly in developing countries. We are pursuing two equally important tasks in the fight against cancer: understanding mechanisms of carcinogenesis and developing remedies to treat individual cancer patients.

During the last century, great progress has been made in understanding the framework of cellular and molecular mechanisms of initiation and development of cancer, and the process of its metastasis. Looking back, the cancer research journey began in 1914 with the chromosomal abnormality theory proposed by Theodor Boveri, and it progressed to the identification in 1960 of the abnormal Philadelphia chromosome in chronic myelogenous leukemia by Peter Nowell and David Hungerford. The concept that "cancer is a chromosomal disorder" grew into a genetic framework for the development of cancer. Alfred Knudson's evidence for a multiple-hit hypothesis for mutations in tumor suppressor genes and the identification of proto-oncogenes by J. Michael Bishop and Harold E. Varmus in the early 1970s and the cloning of k-RAS oncogene and RB1 tumor suppressor gene in the 1980s broadened the concept that "cancer is a disease of genetic and epigenetic aberrations" and established a solid foundation for molecular cancer biology.

The late 1990s was a turning point in cancer research. The strategy changed dramatically from piecemeal methods to screen for cancer genes to the decoding of the cancer genome with high-throughput technology. Initially, this was done with Sanger sequencing, and later with massively parallel sequencing. As the Human Genome Project was moving to completion in 2003, the cancer genome project was initiated with the aim to identify somatically acquired, sequence variants and mutations and, hence, to identify the genes that are critical in the development of human cancers. Subsequently, in 2008, the International Cancer Genome Consortium (ICGC) was developed to provide a collaborative and comprehensive picture of all

the mutations, including copy number changes, insertions, and deletions, in 50 types of cancers. To date, numerous cancer genomes and epigenomes have been sequenced for a range of cancer types and are helping us to gain an unprecedented understanding of molecular mechanisms underlying the complexity of tumor biology. We have defined the cancer initiatome as the collection of all perturbations in genome space that lead to the emergence of malignant transformations; the driver genomic changes now extend beyond the ~2 % protein-coding gene content of the genome and they reside in the noncoding RNA molecules (e.g., piRNAs, microRNAs, long noncoding RNAs) from the actively transcribed regions of the genome. Therefore, our contemporary thinking is that cancer is "a disease of genome alterations."

A deep understanding of cancer biology is now revolutionizing the clinical management of cancer patients. The overall cancer survival rate is improving from early detection and diagnosis to early treatment, from monotherapy to multimodule therapy (chemotherapy, radiotherapy, hormonal therapy, immunotherapy, and so forth) or combinatory treatments, and from general cytotoxic therapy to targeted molecular treatment. As a result, targeted therapy has indeed improved treatment for certain cancers using drugs such as Gleevec (imatinib mesylate) for chronic myelogenous leukemia, Erlotinib for non-small cell lung cancer with EGFR mutations, Herceptin (trastuzumab) for a subset of breast cancer with HER2/neu gene amplification, and recent BRAF inhibitors for metastatic melanoma. An unprecedented number of more genomically derived drugs are currently under clinical trials. Hence, living with chronic cancer disease while maintaining a high quality of life is not uncommon.

Worldwide efforts to beat cancer have never stopped. The result is that the cancer genome, epigenome, and transcriptome can now be read at the single nucleotide level, using massively parallel sequencing technology with a short turnaround time at an acceptable cost. Precision cancer medicine has been born, and the demand is now for individualized cancer therapy to effectively treat this genomically heterogeneous disease.

This book is the second in a series of "Next generation sequencing technology in cancer research—from basepairs to bedsides." Our goal continues to be filling the gap between cancer genome research and clinical management of the individual cancer patient. Our aims are to present the principles of next-generation sequencing (NGS) technologies and massively parallel DNA sequencing and their application of the whole-genome sequences (WGS), whole exome-seq (WES), RNA-seq, miRNA-seq, and ChIP-seq in cancer research programs, and to apply the newly discovered driver genetic alterations for prevention, early diagnosis, and genome-oriented precision cancer treatment. Therefore, we have again invited international cancer researchers and physician-scientists, all of whom are working in multidisciplinary programs, to contribute their achievements in cancer genomic research with the use of NGS technologies. They are eager to translate their new and novel findings from the cancer genomes of individual patients to an orchestrated cancer management team comprised of physicians, genomicists, bioinformatians, clinical researchers, and bioethicists in order to develop precision cancer treatments for individual patients.

We bring together the implementation of a wide range of NGS technologies, including single-cell sequencing, in the clinical setting: discovery and validation of cancer biomarkers; standardization of NGS data production; NGS data reporting systems for clinicians; novel anticancer therapies development from NGS data; and conducting clinical trials of newly investigated cancer drugs. Several chapters discuss the issue of formalin-fixed and paraffin-embedded (FFPE) specimens as input materials for NGS. Moreover, decoding of viral genomes in cancer and the epigenetic genome is also covered. Intriguingly, the authors are providing pipelines for the discovery of novel therapeutic targets, using the actionable and druggable mutations from the cancer gene regulatory networks. Lastly, basic bioinformatic analysis is included in almost every chapter. With the authors' optimistic and enthusiastic translation of cancer genome knowledge into clinical practice, we expect to improve diagnostic, prognostic, and therapeutic outcomes for individual patients.

We intended our book to be a comprehensive guide to contemporary cancer genome research with experimental and computational biology with application in clinics. It provides compelling evidence to signal a new future for health care and a new standard for cancer care. It will be of interest to a broad readership—including medical students, cancer biologists, bioinformaticians, and oncologists.

Successful completion of this book would not have been possible without conversations and assistance from many more people than we can individually acknowledge. Our thanks to all of the authors who worked diligently to produce their enthusiastic contributions and meet the goals of the volume. We are grateful to Dr. Fred Biddle for his encouragement during the course of the book preparation and beyond. Our thanks also go to the Springer staff who have been constructive partners in the publication of this frontier cancer genome research and have ensured that the series is produced in an efficient and timely fashion. Our heartfelt gratitude goes to our own families, who continue to patiently support us as we put forward our efforts for this publication.

Calgary, AB, Canada — Wei Wu
Jeddah, Saudi Arabia — Hani Choudhry

Contents

Contributors

Shiekh Tanveer Ahmad Department of Pathology and Laboratory Medicine, Southern Alberta Cancer Research Institute, University of Calgary, Calgary, AB, Canada T2N 4N1

Ashwag Albukhari Biochemistry Department, Faculty of Science, King Abdulaziz University, Jeddah, Saudi Arabia

Department of Oncology, University of Oxford, Oxford, UK

Ahmad Awada Department of Medicine, Institut Jules Bordet, Brussels, Belgium

Joffre Baker Genomic Health, Inc, Redwood City, CA, USA

Nigel C. Bennett UQ Centre for Clinical Research, The University of Queensland, Herston, QLD, Australia

Rachele Bigi Department of Experimental Medicine, University of Rome "La Sapienza", Rome, Italy

Fawzi F. Bokhari Post Graduate Training and Research Centre, Medical Services General Directorate, Ministry of Defence, Armed Forces Hospitals, Taif Region, Saudi Arabia

Marc De Braekeleer Laboratoire d'Histologie, Embryologie et Cytogénétique, Faculté de Médecine et des Sciences de la Santé, Université de Brest, Brest, France

Institut National de la Santé et de la Recherche Médicale (INSERM), Brest, France

Service de Cytogénétique et Biologie de la Reproduction, Hôpital Morvan, CHRU Brest, Brest, France

Etienne De Braekeleer Division of Stem Cells and Cancer, German Cancer Research Center (DKFZ) & Heidelberg Institute for Stem Cell Technology and Experimental Medicine GmbH (HI-STEM), Heidelberg, Germany

Emilio Bria Department of Medicine, Medical Oncology, University of Verona, Verona, Italy

Rocco Cappellesso Department of Medicine (DIMED), Surgical Pathology Unit, University of Padua, Padua, Italy

Jennifer A. Chan Department of Pathology and Laboratory Medicine, Southern Alberta Cancer Research Institute, University of Calgary, Calgary, AB, Canada

Elizabeth Chao Department of Clinical Diagnostics, Ambry Genetics, Aliso Viejo, CA, USA

Department of Pediatrics, Division of Genetics and Metabolism, University of California, Irvine, CA, USA

Jian Chen Department of Gastroenterology, Hepatology, and Nutrition, The University of Texas MD Anderson Cancer Center, Houston, TX, USA

Chao Cheng Department of Genetics, Geisel School of Medicine at Dartmouth, Hanover, NH, USA

Institute for Quantitative Biomedical Sciences, Geisel School of Medicine at Dartmouth, Lebanon, NH, USA

Norris Cotton Cancer Center, Geisel School of Medicine at Dartmouth, Lebanon, NH, USA

Charles Chesnelong South Alberta Cancer Research Institute (SACRI), Cumming medical School, University of Calgary, Calgary, AB, Canada

Chandra Sekhar Reddy Chilamakuri Department of Tumor Biology, Norwegian Radium Hospital, Oslo University Hospital, Norwegian Radium Hospital, Oslo, Norway

Norwegian Cancer Genomics Consortium (cancergenomics.no), Oslo, Norway

Hani Choudhry Faculty of Science, Biochemistry Department, Center of Innovation in Personalized Medicine, King Fahd Center for Medical Research, King Abdulaziz University, Jeddah, Saudi Arabia

Andrew Dalley UQ Centre for Clinical Research, The University of Queensland, Herston, QLD, Australia

Aarti N. Desai Persistent Labs, Persistent Systems Ltd., Pune, Maharashtra, India

Rodrigo Dienstmann Sage Bionetworks, Fred Hutchinson Cancer Research Center, Seattle, WA, USA

Dirk P. Dittmer Department of Microbiology and Immunology, Lineberger Comprehensive Cancer Center, Center for AIDS Research (CfAR), Carolina Center for Genome Sciences, University of North Carolina at Chapel Hill, Chapel Hill, NC, USA

Jill S. Dolinsky Department of Clinical Diagnostics, Ambry Genetics, Aliso Viejo, CA, USA

Nathalie Douet-Guilbert Laboratoire d'Histologie, Embryologie et Cytogénétique, Faculté de Médecine et des Sciences de la Santé, Université de Brest, Brest, France

Institut National de la Santé et de la Recherche Médicale (INSERM), Brest, France

Service de Cytogénétique et Biologie de la Reproduction, Hôpital Morvan, CHRU Brest, Brest, France

Alan Mitchell Durham Instituto de Matemática e Estatística, Universidade de Sao Paulo, Sao Paulo, SP, Brazil

Aaron M. Elliott Department of Research and Development, Ambry Genetics, Aliso Viejo, CA, USA

Camile S. Farah UQ Centre for Clinical Research, The University of Queensland, Herston, QLD, Australia

The Australian Centre for Oral Oncology Research & Education, School of Dentistry, University of Western Australia, Nedlands, WA, Australia

Matteo Fassan Department of Medicine (DIMED), Surgical Pathology Unit, University of Padua, Padua, Italy

Ambrogio Fassina Department of Medicine (DIMED), Surgical Pathology Unit, University of Padua, Padua, Italy

Heather Fecteau Department of Clinical Cancer Genetics, Texas Health Presbyterian Hospital, Dallas, TX, USA

Tonje Fiskaa Department of Medical Biology, Faculty of Health Sciences, University of Tromsø, Tromsø, Norway

Fei Gao Science & Technology Department, BGI-Shenzhen, Shenzhen, China

Section of Comparative Paediatrics and Nutrition, Department of Veterinary Clinical and Animal Sciences, Faculty of Medical and Health Sciences, University of Copenhagen, Copenhagen, Denmark

Chia-Ling Gau Department of Clinical Diagnostics, Ambry Genetics, Aliso Viejo, CA, USA

Daniel Heim Institute of Pathology, Charité Universitätsmedizin Berlin, Berlin, Germany

Abhay Jere Persistent Labs, Persistent Systems Ltd., Pune, Maharashtra, India

Maryam Jessri UQ Centre for Clinical Research, The University of Queensland, Herston, QLD, Australia

Steinar D. Johansen Department of Medical Biology, Faculty of Health Sciences, University of Tromsø, Tromsø, Norway

Marine Genomics group, Faculty of Biosciences and Aquaculture, University of Nordland, Bodø, Norway

Frederick Klauschen Institute of Pathology, Charité Universitätsmedizin Berlin, Berlin, Germany

Erik Knutsen Department of Medical Biology, Faculty of Health Sciences, University of Tromsø, Tromsø, Norway

Farzaneh Kordbacheh UQ Centre for Clinical Research, The University of Queensland, Herston, QLD, Australia

Holly LaDuca Department of Clinical Diagnostics, Ambry Genetics, Aliso Viejo, CA, USA

Biaoru Li Department of Pediatrics, Medical College at GA, Augusta, GA, USA

Shaoke Lou Department of Genetics, Geisel School of Medicine at Dartmouth, Hanover, NH, USA

Rajyalakshmi Luthra Department of Hematopathology, University of Texas MD Anderson Cancer Center, Houston, TX, USA

Yan Ma Genomic Health, Inc, Redwood city, CA, USA

Leonardo A. Meza-Zepeda Department of Tumor Biology, Oslo University Hospital, Norwegian Radium Hospital, Oslo, Norway

Norwegian Cancer Genomics Consortium (cancergenomics.no), Oslo, Norway

Lopa Mishra Department of Gastroenterology, Hepatology, and Nutrition, The University of Texas MD Anderson Cancer Center, Houston, TX, USA

Karthikraj Natarajan Science & Technology Department, BGI-Shenzhen, Shenzhen, China

Nadine Norton Department of Cancer Biology, Mayo Clinic, Jacksonville, FL, USA

Liliane Santana Oliveira Albert Einstein Research and Education Institute, Hospital Israelita Albert Einstein, Sao Paulo, SP, Brazil

Tuya Pal Department of Cancer Epidemiology and Internal Medicine, Moffitt Cancer Center, Tampa, FL, USA

Maria Perander Department of Medical Biology, Faculty of Health Sciences, University of Tromsø, Tromsø, Norway

Tina Pesaran Department of Clinical Diagnostics, Ambry Genetics, Aliso Viejo, CA, USA

Kunbin Qu Genomic Health, Inc, Redwood city, CA, USA

Manar Samman Leeds Institute of Cancer and Pathology, St James' University Hospital, Wellcome Trust Brenner Building, Leeds, UK

King Fahad Medical City, Riyadh, Saudi Arabia

Marcia Sanders Department of Microbiology and Immunology, Lineberger Comprehensive Cancer Center, Center for AIDS Research (CfAR), Carolina Center for Genome Sciences, The University of North Carolina at Chapel Hill, Chapel Hill, NC, USA

Program in Global Oncology, Lineberger Comprehensive Cancer Center, and Center for AIDS Research (CfAR), The University of North Carolina at Chapel Hill, Chapel Hill, NC, USA

Aldo Scarpa ARC-NET Research Center and Department of Pathology and Diagnostics, University of Verona, Verona, Italy

Neeraj Sethi Leeds Institute of Cancer and Pathology, St James' University Hospital, Wellcome Trust Brenner Building, Leeds, UK

Patricia Severino Albert Einstein Research and Education Institute, Hospital Israelita Albert Einstein, Sao Paulo, SP, Brazil

Rajesh R. Singh Department of Hematopathology, University of Texas MD Anderson Cancer Center, Houston, TX, USA

Virginia Speare Department of Clinical Diagnostics, Ambry Genetics, Aliso Viejo, CA, USA

Albrecht Stenzinger Institute of Pathology, University of Heidelberg, Berlin, Germany

Xiaoping Su Department of Bioinformatics and Computational Biology, The University of Texas MD Anderson Cancer Center, Houston, TX, USA

Jordan Texier Department of Microbiology and Immunology, Lineberger Comprehensive Cancer Center, Center for AIDS Research (CfAR), Carolina Center for Genome Sciences, The University of North Carolina at Chapel Hill, Chapel Hill, NC, USA

Program in Global Oncology, Lineberger Comprehensive Cancer Center, and Center for AIDS Research (CfAR), The University of North Carolina at Chapel Hill, Chapel Hill, NC, USA

François Thomas Thomas Conseil SPRL, Brussels, Belgium

Cheryl L. Thompson Department of Family Medicine and Community Health, Case Comprehensive Cancer Center, Case Western Reserve University, Cleveland, OH, USA

Department of Epidemiology and Biostatistics, Case Comprehensive Cancer Center, Case Western Reserve University, Cleveland, OH, USA

Matthew H. Ung Department of Genetics, Geisel School of Medicine at Dartmouth, Hanover, NH, USA

Vinay Varadan Department of General Medical Sciences (Oncology), Case Comprehensive Cancer Center, Case Western Reserve University, Cleveland, OH, USA

Frederick S. Varn Department of Genetics, Geisel School of Medicine at Dartmouth, Hanover, NH, USA

Wei Wu Department of Pathology and Laboratory Medicine, Southern Alberta Cancer Research Institute, University of Calgary, Calgary, AB, Canada

Jie Xiong The University of North Carolina at Chapel Hill, Chapel Hill, NC, USA

Department of Statistics and Operations Research, The University of North Carolina at Chapel Hill, Chapel Hill, NC, USA

Hongyan Xu Department of Biostatistics and Epidemiology, Medical College of Georgia, Georgia Regents University, Augusta, GA, USA

Dongmei Yang Department of Microbiology and Immunology, Lineberger Comprehensive Cancer Center, Center for AIDS Research (CfAR), Carolina Center for Genome Sciences, The University of North Carolina at Chapel Hill, Chapel Hill, NC, USA

Program in Global Oncology, Lineberger Comprehensive Cancer Center, and Center for AIDS Research (CfAR), The University of North Carolina at Chapel Hill, Chapel Hill, NC, USA

Xiaomeng Zhang School of Computer Engineering, Nanyang Technological University, Singapore, Singapore

Jie Zheng School of Computer Engineering, Nanyang Technological University, Singapore, Singapore

Single-Cell Next-Generation Sequencing and Its Applications in Cancer Biology

Biaoru Li, Xiaomeng Zhang, and Jie Zheng

Abstract A complete set of DNA with its transcripts is defined as genome, which includes both the genes and the noncoding sequences of the DNA/RNA. After making advances in decoding different genomes across species, genomic techniques such as SNP microarrays and gene expression microarray have been synchronously developed to analyze the genomic functions. Now, scientists are able to take the study of genomics into deep consideration of biological evolution and mechanism of different diseases. However, there are still challenges with the genomic technology. Some tissues of human and animals, such as tumor tissues, contain multiple heterogeneous cells, making analysis extremely difficult. Additionally, some specimens have very few cells, such as circulating tumor cells. To fully study DNA genomic changes and its expression changes in cancer, single-cell genomic techniques have been broadly applied to fields such as cytogenomic diagnosis for specimens on glass slides, tumor cells in circulating blood, measurement of sensitivity and specificity of genomic analysis at tumor tissue level, mechanism of differentiation of cancer stem cell, etc. Recently, next-generation sequencing (NGS) has become an important tool in single-cell genomic analysis. Here, we systemically introduce single-cell NGS from single-cell sampling, single-cell NGS, and single-cell NGS-related bioinformatics into its application for tumor biology. This chapter also describes some advantages of single-cell NGS and addresses some challenges of single-cell NGS for genomics analysis due to the specimen features.

B. Li, M.D., Ph.D. (✉)
Department of Pediatrics, Medical College at GA, Augusta, GA 30912, USA
e-mail: BLI@gru.edu; brli1@juno.com

X. Zhang • J. Zheng
School of Computer Engineering, Nanyang Technological University,
Singapore 639798, Singapore

W. Wu, H. Choudhry (eds.), *Next Generation Sequencing in Cancer Research, Volume 2*, DOI 10.1007/978-3-319-15811-2_1

1 Introduction

DNA (deoxyribonucleic acid) composed of four bases and its double helical strand structure was first demonstrated by James D. Watson and Francis Crick in 1953 [1]. Since then, genes at the DNA and mRNA levels were broadly studied for their functions such as normal evolution of species and mechanism of human diseases as described by Drs. Er and Chang in 2012 [2]. During the early period of research, DNA sequencing techniques played important roles for studying gene structures and gene expression. In 1977, Frederick Sanger launched DNA sequencing technology that relied on DNA chain-termination method (Sanger sequencing) [3] and Walter Gilbert studied chemical modification and cleavage at specific bases of DNA as an early sequencing technology [4]. Sanger sequencing is described as the first-generation DNA sequencing due to its high efficiency and low radioactivity as delineated by Dr. Pareek in 2011 [5]. Following great accomplishments from the human genome project in 2002–2003, massively parallel sequencing systems called as next-generation sequencing (NGS) were brought about the world. In 2005, the 454 sequencing system provided massively parallel sequencing reading platform as reported by Margulies et al. in 2005 [6]; Solexa developed Genome Analyzer system as portrayed by Warren et al. in 2006 [7]; and Agencourt supplied SOLiD platform as explained by Mardis in 2008 [8]. All three NGS systems have similar features including high throughput and accuracy although there are differences such as the read lengths. Recently, the founder companies were bought by other companies. For instance, SOLiD system was purchased by Applied Biosystems in 2006; in 2007, 454 sequencing system was bought by Roche and Solexa system was picked up by Illumina as reviewed by Dr. Liu in 2012 [9]. The three systems exhibit their advantages including their read length, accuracy, and applications as presented in Table 1. NGS system has also been developed into compact model for small size of sample, such as Ion Personal Genome Machine (PGM) and MiSEQ. These two systems were extended by Ion Torrent and Illumina for their advantages in fast running and its cheap costs as shown in Table 2. Moreover, accompanied with increasing new modifications in NGS, a third-generation sequencing such as Single-Molecule

Table 1 NGS system comparison

Systems	454 GS FLX	HiSeq system	SOLiD system
Sequencing mechanism	Pyrosequencing	Sequencing by synthesis	Ligation and two-base coding
Read length	700 bp	50SE, 50PE, 101PE	50 + 35 bp
Reads	1 M	3 G	1200–1400 M
Output data/Run	0.7 Gb	600 Gb	120 Gb
Time/Run	24 h	3–10 days	7 days for SE
Advantage	Read length, fast	High throughput	Accuracy
Accuracy	99.90 %	98 %	99.94 %
Disadvantage	Low throughput	Short read assembly	Short read assembly
Cost/million bases	$10	$0.07	$0.13

Table 2 Compact NGS

Compact NGS	MiSeq	PGM
Sequencing method	Sequencing by synthesis	Semiconductor technology with a simple sequencing chemistry
Read length	Up to 2×300 bp	200–400 bp
Output	540 MB to 15 GB	30 MB to 2 GB
Sequencing time	4 h for 1×36 single read, 27 h for 2×300 bp end read	2.3–4.4 h for 200 bp reads 3.7–7.3 h for 400 bp reads
Sample preparation time	About 2 h	8 samples in parallel, less than 6 h
Input amount	Nanogram (Nextera)	μg

Real-Time (SMRT) has increased appliances in genomic studies. SMRT does not need PCR amplification and the nucleotides real-time signal of SMRT in enzymatic reaction can be captured by fluorescent (Pacbio) or electric current (Nanopore). Technically, NGS of whole-genomic DNA is called as DNA-Seq consisting of whole-genomics sequencing (WGS) and whole-exome sequencing (WES); NGS of whole mRNA is named as RNA-Seq; NGS of whole microRNA is said as miRNA-Seq and so on.

The research and development (R&D) of single-cell genomics in tumor biology has the advantage of requiring few cells and a cellular environment of mixed cells. For instance, development of clonal cell (such as cancer stem cell in cancer) occurred with subtle heterogeneity at an early period including few mutations and chromosomal rearrangements finally leading to massive cell proliferation and differentiation in a mixed tissue due to switch of the tumor cell program with enriched genomic changes. In the tumorigenesis, very few cells are available in the early period while mixed-cell tumor tissues arise in late tumor development according to Dr. Breivik's study in 2005 [10]. All these reasons require R&D of single-cell genomic techniques to study the tumorigenesis. In addition, genomic diagnosis for a given type of cells in mixed-cell tumor tissue can only adopt very small numbers of cells such as clinic biopsy specimens or single-cell isolated from laser capture microscopy of tumor tissues. The single-cell technique with downstream genomics needs to be applied itself from cells on slides in molecular pathology and cytogenetic. Moreover, it is necessary for biomarker discovery of tumor cells in circulating blood as described by Liberko et al. in 2013 [11]. Several years earlier, genomics of identification and quantification have been developed into single-cell genomic level including Array-CGH and SNP-microarray for DNA genomics and mRNA microarrays, subtractive cloning and differential display (DD) for mRNA genomic profiles as illustrated by Ning et al. in 2014 [12]. Technically, each single-cell genomic analysis and diagnosis has its own disadvantages and advantages. After NGS was applied in 2007 and developed into single-cell genomic technique in 2010, single-cell NGS techniques have allowed physicians and scientists to use the important tools for single-cell diagnosis as explained by Ebenezer et al. in 2012 [13].

In order to distinctly advocate single-cell NGS, here we will first introduce the single-cell techniques and then present single-cell NGS techniques with downstream

single-cell NGS bioinformatics. Finally, we will briefly review applications to the study of cancer biology by using the single-cell NGS techniques. In conclusion part, we will also discuss advantages and disadvantages of applying different single-cell genomic techniques.

2 Single-Cell Technique

Tumor specimens of animal and human tissue often contain multiple cells. Different DNA changes and different gene expression profiles in a given type of cells coexist in the same specimen of animal and human tissue. Theoretically, important findings of genomic-DNA SNP profile or mRNA expression profiles will be unclear for a certain type of cells if we make use of tissue-level genomic profile. Therefore, pure or representative single cells will provide the most precise analysis possible of these subtle gene expression patterns in the given type of cells. Here, in order to explicitly discuss single-cell NGS, two techniques, or single-cell sampling and DNA/mRNA amplification from a single cell will be first introduced.

2.1 Single-Cell Sampling

As shown in Table 3, flow-cytometric cell sorting (FACS) and laser-based microdissection of tumor tissues provide ways to isolate single cells for DNA genomics change and gene expression profiling in a given type of cells. In FACS system, cells labeled with fluorescent signals in solution can be isolated based on a specific biomarker such as a tumor antigen attached to an antibody labeled by a fluorescent signal. At present, FACS can specifically separate targeted cells and collect the single cell into 96 wells for downstream genomics (AmpliGrid by Advalytix) as reported by Brück et al. in 2010 [14]. Although FACS and multicolor FACS can

Table 3 Single-cell sampling

Methods	Advantages	Disadvantages
Laser-capture microdissection	Microenvironment and local data	Theoretical damage to the target cell
Laser-assisted mechanical microdissection	Microenvironment and local data	Laborious
Laser-catapult microdissection	Very little contamination with microenvironment and local data	Special slides
Flow-cytometric cell sorting	Auto- and rapid separation into 96-well plate	Limit in some cells such as neuron without microenvironment data

isolate and sort homogeneous cells, even single cell, three challenges limit their applications: (a) FACS cannot be subject to some types of cells such as neurons; (b) intracellular biomarker cannot be well defined and sorted by FACS; (c) the tumor microenvironment of a cell cannot be evaluated by FACS. The microdissection technique can avoid the aforementioned three limitations. In 1976, the use of lasers in tissue microdissection has been reported by Meier-Ruge et al. [15]. In contrast to single-cell FACS, microdissection allows both rapid in vivo localization and ability to analyze the cellular microenvironment as depicted by Schutze et al. in 1998 [16]. At present, three microdissection systems have been broadly developed as reported by Li in 2005 [17]: (1) laser-assisted mechanical tissue microdissection, (2) laser pressure catapult microdissection, and (3) laser capture microdissection (LCM). Laser-assisted mechanical tissue microdissection can focus on small target cell areas, reducing the chance of contamination with neighboring cells as portrayed by Emmert-Buck et al. in 1996 [18]. Although the concept of using a laser to dissect out individual cells is quite simple, the technique is laborious. Laser pressure catapult microdissection concentrates on an interesting region with a high-energy cutting laser. Following a low-power laser sets the depth of the tissue section, a pressure wave then separates the targeted tissue from the slide and catapults it into a receptacle. The high precision of the thin beam laser is sufficient to isolate subcellular targets such as chromosomes. The absence of physical contact between the surrounding tissues and the collection apparatus results in a much lower incidence of contamination. In laser capture microdissection, a thin ethylene vinyl acetate film is mounted on the tissue section. After an infrared laser heats and melts a cell of interest, the resolidified plastic film binds directly to this cell and catches it as reported by Fend et al. in 1999 [19]. Now, all of three systems are commercially available for laboratory studies in animals, plants, and human beings.

2.2 Genomic Amplification from Single Cells

In a human diploid cell, the quantity of DNA is a constant or 6.6 pg of each diploid single cell (three billion base pairs multiply two for diploid and multiply 660 for molecular weight of each base pair), although about 5 pg per human cell is harvested in real experiment. After more than 10 years effort, genomics DNA isolation and amplification from single cells are very mature called as whole-genome amplification (WGA) (Table 4). Now, three companies [Genomeplex kit (Sigma), Picoplex kit (Rubicon), and Genomiphi kit (GE)] are commercially available for genomics DNA isolation and amplification for single cells as, respectively, delineated by Fiegler et al. in 2007 [20], Kurihara et al. in 2011 [21], and Pan et al. in 2008 [22]. All three products work very well for genomic DNA amplification although there are some subtle differences such as base pair length and PCR amplification techniques (see Table 4) and although some scientists prefer to perform MDA (Multiple Displacement Amplification) from the product to process DNA of single cell.

Table 4 Single-cell genomic amplification

Genomic types	Methods	Advantages	Primers	Amplification products	Commercial available
DNA amplification	Primer-extension preamplification (PEP)	Stable genomic amplification	Degenerate oligonucleotide primer (DOP)	400–1,500 bp	Sigma (Genomeplex)
	Two cycling	Stable genomic amplification	Degenerate oligonucleotide primer (DOP)	400–500 bp	Rubicon (picoplex)
	Phi	>1 kb	Multiple strand displacement (MSP)	Variable for repeat	GE (Genomiphi)
mRNA amplification	Tang's	Longer fragment	UP1 for olig-T and UP2 with OligoA	Longer until 3 kb	Combined kit with design
	Smart-Seq	Easy and commercial available	UP1 for olig-T and UP2 with CCC switch	Nanogram start	SMARTer® Ultra™ Low RNA Kit
	STRT	Longer fragment with barcoding sequencing	UP1 for olig-T and UP2 with CCC switch and barcoding seq	Longer until 2 kb	Combined kit with design
	Cell-Seq	Linear amplification	UP1 for olig-T and barcoding seq and UP2 with multiplex cell seq	Sensitivity	Combined kit with design

The quantity of mRNA in a single cell is greatly different, 1.0–20 pg (about 5×10^5–10×10^7 molecules) based on the cell size, cell function, and cell differentiating stage as described by Ambion in 2004 [23]. Although some scientists try to isolate RNA from single cells, most of scientists prefer to use a crude cell lysate without purifying procedures as reported by Klebe et al. in 1996 [24]. This protocol has two important advantages. First, it ruptures the cells and releases the RNA directly into a cell lysis buffer without loss of RNA. Moreover, the heating step to rupture cells inactivates endogenous RNase for protecting RNA from degradation. Theoretically, mRNA amplification should be applied in single-cell genomic technique. Now, four mRNA amplifications strategies have been developed into single-cell RNA-Seq. Their performances with their amplification mechanism, primers design, and PCR product sizes are listed in Table 4: Smart-Seq (switching mechanism at the 5′ end of the RNA transcript), STRT techniques (single-cell tagged reverse transcription), CEL-seq, and Tang's method, as respectively reported by Ramsköld et al. in 2012 [25], Lobo et al. in 2009 [26], Hashimshony et al. in 2012 [27], and Tang et al. in 2009 [28]. Here, two basic mRNA amplification principles will be first launched: mRNA amplification (aRNA) and PCR-based cDNA amplification. The aRNA procedure begins with total RNA or poly(A)+RNA that is reversely transcribed using an oligo (dT) primer containing a T7 RNA polymerase promoter sequence. After first-strand synthesis, the reaction is treated with RNase H to fragment the mRNA. These fragments serve as primers during a second-strand synthesis reaction that produces a double-stranded DNA template for transcription. rRNA, mRNA fragments, and primers are removed before using the cDNA template to produce linearly amplified aRNA. The amplification yields can reach 1,000- to 5,000-fold following two rounds of in vitro transcription. RNA amplification is commercially available and has been increasingly reported in gene expression studies as described by Eberwine in 1996 [29]. PCR-based amplification has two protocols: specific profile and global profile applications. Specific profile methods such as RT-PCR or multiplex RT-PCR reactions are sensitive at the single-cell level, especially in nested PCR. Because the genes studied using these methods are preselected, it can only be applied to known genes. Global PCR-based approaches have been developed in genomic analysis. Two approaches are commercially available, homomeric tailings and 3′-(3-primer-end) amplification (TPEA). The homomeric tailings as designed by Toellner et al. in 1996 [30] use terminal deoxynucleotide transferase-generated homomeric 3′ tails to the first-strand cDNA. After RT-PCR and 3′ tailing addition and PCR amplification, it has been applied to the analysis of single-cell global gene expression. Even though homomeric tailings can be used effectively in global profile analysis, many of the cDNA copies are not full length and shorter cDNAs are preferentially amplified. 3′-end-amplification (TPEA) as reported by Dixon et al. in 1998 [31] is a randomized amplification of mRNA using an oligo-dT primer together with a 5′ primer containing a random pentamer. It can enable the detection of both high- and low-abundance mRNA transcripts from single cells.

3 Single-Cell Next-Generation Sequencing

3.1 Single-Cell DNA-Seq

Routine genomic DNA performance including next-generation sequencing and SNP microarrays requires sufficient and high-quality DNA. For single-cell genome analysis, as previously discussed, a special process termed whole-genome amplification (WGA) is added as illustrated in Fig. 1a. The WGA process can amplify the whole DNA population producing large amounts of DNA from a single cell whose quantity is comparable to routine genomic DNA. Due to the exponential amplification, three challenges will be created in the amplification process: amplified sequence bias during WGA, genetic material contamination caused by heterogeneous amplification, and genomic dropouts caused by tiny DNA materials as described by Gole and Gore in 2013 [32]. To overcome the three obstacles, each step of the process must be performed under quality control (QC) with Good Management Practice (GMP) compliance.

After performing single-cell sampling and WGA from target cells, high-throughput sequencing using WGA DNA is carried out using routine genomic DNA-Seq, which is briefly elaborated as follows. After the genomic DNA library is prepared, genomic DNA is fragmented and purified for enzymatic processes such as DNA end repair, A-tailing, adaptor ligation, DNA fragment size selection, and DNA fragment amplification. Following library amplification, the library is quantified

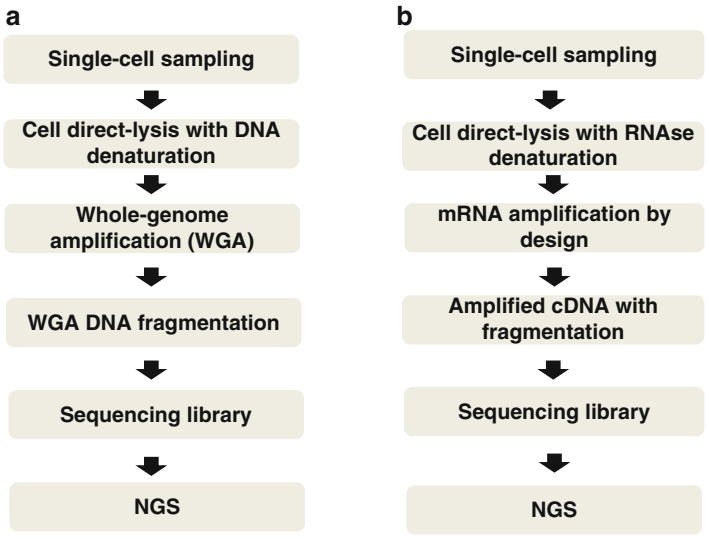


Fig. 1 The diagram of next-generation sequencing procedure: (**a**) Single-cell DNA-Seq workflow: after single-cell sampling and whole-genome amplification and fragmentation, library is quantified and is submitted to the sequencer; (**b**) single-cell RNA-Seq workflow: after single-cell sampling and whole-genome amplification by design and fragmented, library is quantified and is submitted to the sequencer

using real-time PCR and a predetermined amount of DNA library is submitted to the sequencer. The exact protocol of all of these steps is described in different NGS systems by Landau et al. in 2014 [33].

3.2 Single-Cell RNA-Seq

Routine RNA-Seq provides accurate quantification of mRNA expression levels with entire transcript lengths. Routine sampling for RNA-Seq is largely based on traditional molecular biological protocols including the basic steps of poly-(A)+RNA isolation, fragmentation, reverse transcription, and amplification before the actual sequencing takes place. The selection of poly-(A)+RNA is usually performed in order to suppress rRNA and tRNA. The fragmentation step is carried out in order to produce many short RNA or DNA fragments that represent the original transcript.

Following the basic principles of transcripts amplification discussed previously, four strategies of mRNA amplification for single-cell RNA-Seq have been developed into the single-cell level: Smart-Seq (switching mechanism at the 5′ end of the RNA transcript), STRT (single-cell tagged reverse transcription), CEL-seq, and Tang's amplification. After performing single-cell sampling and transcripts amplification and fragmentation, high-throughput RNA-Seq is performed using routine RNA-Seq. As DNA-seq performance, after a genomic RNA library is prepared, genomic fragment is purified for enzymatic processes such as end repair, A-tailing, adaptor ligation, library fragment size selection, and library fragment amplification. The library of accurate quantity is also measured by real-time PCR and then accurate amount of library is submitted to the sequencer as shown in Fig. 1b. The detailed protocol of NGS is described in different NGS platforms by Panagopoulos et al. in 2014 [34].

4 Single-Cell NGS-Related Bioinformatics

Next-generation sequencing (NGS) is a radical breakthrough at whole-genome level, offering unprecedented data depth not found in previous Sanger sequencing technology. A number of NGS platforms are developed based on different sequencing technologies, the details of which are beyond the scope of the work. Here, we simply highlight that all NGS platforms perform a common task that is to sequence millions of small fragments of DNA in parallel. Consequently, each of several billion bases in the target species or disease genome is sequenced multiple times, leading to a high level of data depth and accuracy. By making use of appropriate bioinformatics analysis tools, these fragments of data can be pieced together whereby individual reads are mapped to a species-specific reference genome. The mapped genome is highly sought after as it may shed light on the unexpected DNA variation or the quantity of RNA expression. In this section, we will focus our discussion on bioinformatics analysis related to single-cell NGS, to be more limited, single-cell DNA and RNA sequencing.

4.1 Single-Cell DNA-Seq Bioinformatics

As discussed in Sect. 3.1, single-cell DNA-Seq faces three obstacles (amplified sequence bias, genetic material contamination, and genomic dropouts). These obstacles make the single-cell DNA sequencing data inaccuracy. The impaired single-cell sequence data cannot be analyzed by most bioinformatics tools developed for bulk cell sequencing. To tackle this problem, some new bioinformatics tools have been designed for analyzing single-cell sequencing following the eruption of single-cell sequencing data. In this section, we will describe the applications of single-cell bioinformatics in analyzing single-cell WGS or WES.

Theoretically, single-cell sequencing data open up an opportunity to study genealogy of an individual tumor cell. The genealogy of the tumor cell unveils the complete picture from the earliest signs of mutation until accumulated heterogeneous tumor. If the mutation pedigree is constructed in a systematic manner, any unrelated lineage can be easily identified. In early model, Navin and his colleagues performed copy number variation analysis on breast tumors using low coverage single nucleus sequencing as reported in 2011 [35]. Their study aimed to explain clonal evolution of the tumors. They constructed a phylogenetic tree based on sample cell numbers and subpopulations based on the distances in the tree between the samples. Following Navin's analysis, Hou et al. used exome sequencing data from 58 single cells of an essential thrombocythemia (ET) tumor and Li et al. utilized exome sequencing data from 66 single cell samples of a bladder transitional cell carcinoma, respectively, to perform mutation to study subgroup of the samples in 2012 [36, 37]. All of these studies clearly illustrated clonal evolution using single-cell sequencing. In 2014, Kim and his colleagues continued working on the model of mutation pedigree including temporal and lineage relationships among DNA sequence mutation sites. They applied their algorithm in an 18-sites map as a lineage [38] which Dr. Hou had previously identified as lineage family in their single-cell sequencing dataset so that Dr. Kim proposed a new method to construct evolutionary mutation tree, which could indicate the temporal order relationship between mutation sites. They also proposed a method for estimating the proportion of time starting from the earliest mutation event and from the emergence of most recent common ancestor, respectively, toward the end of mutation. In conclusion, many new bioinformatics tools designed can be used to analyze single-cell DNA sequencing data as illustrated in Fig. 2.

4.2 Single-Cell RNA-Seq Bioinformatics

On top of the usual RNA-seq processes, single-cell RNA-Seq performance requires two additional processes: single-cell sampling and RNA amplification. Although four techniques of RNA amplification available from company products have been developed for RNA-seq, sensitivity and specificity of genomic expression after

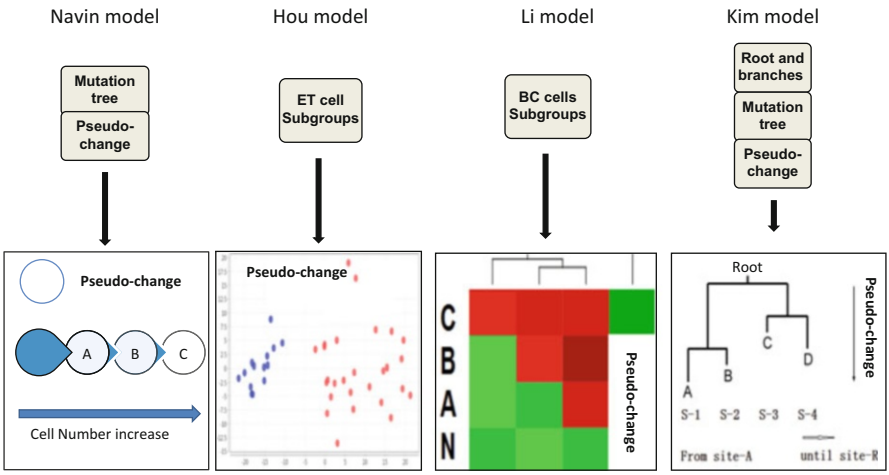


Fig. 2 Four models of single-cell NGS to detect DNA informative change and pseudo-change. Navin model is to study clonal evolution of mutation tree from A to C according to enhancement of tumor cell number from A to C related with mutation order pattern and pseudo-change based on unrelated information; Hou model is to use Principle Component Analysis (PCA) as model: PCA-1 as *x*-axis and PCA-2 as *y*-axis, *blue spots* from normal cells and *red spots* from tumor cells, all exome sequencing data from 58 single cells of an essential thrombocythemia (ET) tumor to study ET cell mutation subgroups; Li model is to utilize mutation pattern in heat map to study mutation subgroup from 66 single cell samples of a bladder transitional cell carcinoma, in which N is control from normal cell, A is mutation pattern explained as earliest cell, B is second, and C is third described as continuance pattern with column indicating different genes, *red* is higher frequency mutation and *green* is lower mutation frequency; Kim model is working on mutation pedigree among DNA sequence mutation sites in which they have 18 sites defined as branches (total 18 sites from site-A to site-R although the figure show only from S-1 to S-4). They can map root to branches including mutation tree and pseudo-change

RNA amplification have been carefully analyzed by our scientists. Our single-cell RNA-Seq technique from CD8 cell of tumor-infiltrating lymphocytes (TILs) demonstrated that fragments were 250–450 bp after fragmentation, amplification, and adapter addition. There were 11.6 million reads mapped in raw sequencing reads (19.6 million). The numbers of mapped genes, mapped transcripts, and mapped exons were 31,332, 41,210, and 85,786, respectively. All QC results illuminated that RNA-seq techniques could be used for single-cell genomic performance. Analysis of the mapped genes showed that the number of genes mapped by RNA-seq (6,767 genes) was much higher than that of differential display (288 libraries) among similar specimens which we had previously developed. The single-cell RNA-Seq can detect gene splicing using different subtype by using TGF-beta analysis. The results using Q-RT-PCR assays demonstrated that sensitivity was 76 % and specificity was 55 % from the single-cell RNA-Seq technique although some gene expression was still missing (2/8 genes). Therefore, the results support that RNA-Seq technique is feasible to analyze single-cell mRNA specimens as described by Xu et al. in 2013 [39].

5 Single-Cell NGS Application

5.1 Pathological Diagnosis

The pathologic diagnosis of tumors relies on cell morphology, tumor cell arrangement, and its infiltrating into normal tissue. The diagnosis of cytogenetics depends on chromosome structure with its number and arrangement change. Following the development of single-cell techniques, a new term, "Single-Cell Diagnosis," has arisen in disease diagnosis in which single-cell genomic diagnosis is involved in molecular pathology and genetics, especially for tumor diagnosis. Now, single-cell genomic diagnosis can be applied for many clinical specimens, such as surgical specimens, biopsy specimen, and tumor cell from circulating blood. Single-cell genomic analysis and diagnosis have much more advantages than other diagnosis. For instances, along with genomic analysis from tumor cells, genomic data can convert pathological changes of tumors into biomarker discovery; in pace with genomic analysis, single-cell genomic analysis can link tumor diagnosis into targeted therapeutics so single-cell genomic diagnosis can be used for personalized therapy; in addition, single-cell genomic analysis and diagnosis can be developed for several other applications such as study of mechanism of tumorigenesis as explained by Macaulay et al. in 2014 [40]. Single-cell NGS of cancer diseases is one of earliest applications for next-generation sequencing. Because single-cell NGS plays a very important role in cancer biomarker discovery and personalized therapy, now, Genomeplex kit from Sigma Inc, Picoplex kit from Rubicon Inc., and Genomiphi from GE all participate in the research and development of single-cell NGS related to genomic analysis and diagnosis such as single cell from slides or single cell from circulating blood of tumor disease.

5.2 Biomarker Discovery

Early diagnosis and treatment is an important impact to reduce the mortality of tumor disease. Currently, some of screening tools (CT, X-ray, mammography, and invasive needle or surgical evaluation for cancer disease) are not sensitive enough for early detection of the diseases, thus some of tumors cannot be treated at an early stage. Theoretically, if some special biomarkers can define each type of tumor cells, it should be the best way for early diagnosis although it is difficult to define now. Genomic technologies have allowed scientists to discover some special biomarkers from thousands of gene expression profiles and evaluate functions of special biomarkers to obtain a global view of tumor cells. After tumor cells are defined on slides or after tumor cells are harvested from circulating blood, single-cell genomic diagnosis is a rational module to define the tumor biomarkers. Single-cell RNA-Seq has been begun to apply for biomarker discovery including their therapeutic targeting. Now, Single-Molecule Real-Time (SMRT), third-generation sequencing has been successfully applied for biomarker discovery from glioblastomas as delineated by Meldrum et al. in 2011 [41].

5.3 Therapeutic Targeting Identification

Recent development of cancer research has enabled scientists to understand the difference of certain type of cancers to respond to chemotherapy analyzed by single-nucleotide polymorphisms (SNP) and genome-wide association studies (GWAS). GWAS analysis, one of genomic medicine, emphasizes different responses of drugs in a certain SNP, called as pharmacogenetics. Because SNP is the information archive but most of the FDA compounds and drugs are directed at phenotype alteration (such as RNA or proteins), not direct to DNA archives, the phenotype products of genotype change have also a great impact on the genomic medicine. Now gene expression profiles related network are used to uncover genomic expression signature (GES, previously called as therapeutics targeting identification, TI) to discover sensitive drugs, broadly called as pharmacogenomics. According to both concepts, drug discovery based on either GWAS or genomic expression signature related network is increasingly developed in treatment of drug-resistant tumor diseases as discussed below.

5.3.1 GWAS Related with Therapeutic Targeting and Personalized Therapy

Cancer stem cells (CSCs) and drug-resistant tumor cells mixed in tumor tissues play an important function in the tumor development and progression. CSCs drive the metastatic spread of cancer and are able to resist conventional therapies so that the disease is difficult to be completely eradicated. If a specific mutant or fusion protein, which results in tumor development or resistance of conventional therapy, can be uncovered by GWAS analysis, a specific targeting compound or Ab to target this mutant or fusion protein will offer a new therapeutic tool to treat drug-resistant tumor cells. According to this concept, several special antibodies and compounds to these mutant or fusion proteins have been routinely used to treat drug-resistant tumors called as molecular therapy or targeted therapy (or one kind of personalized therapy) as illustrated by Guan et al. in 2012 [42]. Single-cell DNA genomics can definitely uncover mutant and fusion proteins by GWAS analysis. Now single-cell NGS-related GWAS analysis is being developed in the tumor cells from slides or from circulating blood of tumor disease.

5.3.2 Network Related with Personalized Therapy

In clinical fields, besides GWAS-related personalized therapy as discussed in Sect. 5.3.1, genomic (or proteomics) expression profile, a second module of personalized medicine of special therapeutic strategies, is going to extend into different diseases. The personalized medicine is directly tailored for physicians to prevent and care individual patient relying on personal genomic expression profiles. It is often called as “the right treatment for the right person at the right time.” All examples of successful personalized treatments require a rational clinical genomic

expression analysis based on R&D of clinical genomic expression diagnosis, and we have successfully established a bioinformatics module from genomic expression profile for personalized therapy in 2008 [43]. The module included mRNA genomic expression profile mined from a specimen, genomic expression signature discovered by quantitative network, and sensitive drugs uncovered from drug bank. Now, after Single-Molecule Real-Time (SMRT), a third-generation sequencing, is brought into the new fields, single-cell genomic diagnosis (such as single cell from slides or single cell from circulating blood) related with discovery of genomic expression signature will make great contribution for the personalized therapy.

5.3.3 Network Related with Personalized Immunotherapy

Personalized immunotherapy is a major breakthrough in cancer immunotherapy including genetically engineered T cells by chimeric-antigen-receptor to kill own tumor cells, using own tumor cells to develop a personalized vaccine to kill own tumors, and activating T-cells quiescent network using own T-cells to kill own cancer cells. CD8 cells from tumor infiltrating lymphocytes (TILs) can directly and specifically recognize and kill own tumor cells after they are activated and expanded ex vivo. If the cells, which have been attached to tumor cells and will recognize own specific tumor antigen, are harvested by single-cell technique, the single-cell genomic profiles will play an important role in a personalized immunotherapy. We have studied single-cell genomic profiles from TILs for more than 10 years. According to concepts of immunology and tumor immunotherapy, CD8 cell of TILs has two obvious advantages: (a) the CD8 T-cells have function of MHC class I to access tumor cells; (b) the CD8 T-cell is specifically recognizing tumor antigen to kill tumor cells. If we uncover genomic profiles related to the specific CD8 cell from TILs which has been specifically accessing tumor cells obtained by single-cell technique, the genomic profiles can decode CD8 cell quiescence. Under culturing the TILs ex vivo combined with dequiescence and with specific function activity by network analysis in silico, the cultured TILs have much stronger function to kill tumor cells. As we all know, CD8 T-cell is an earliest cell model to be developed by single-cell genomic technique. In order to develop personalized immunotherapy to treat tumor diseases, we have developed single-cell genomic techniques from single-cell differential display, single-cell microarray until single-cell NGS as reported by Zhang et al. in 2009 [44]. Now, single-cell NGS-related quantitative network is being developed in personalized immunotherapy to treat advanced tumor diseases.

5.4 Tumorigenesis Related to Cancer Stem Cell

As discussed earlier, single-cell NGS can identify the earliest mutations and set phylogenetic tree of tumor cells. All of these pedigree trees can address clonal evolution. The earliest mutation site is located at the root and then gradual extension

from the root to other sites in the tree. Eventually, the trees can be used to estimate the earliest mutation event of the tumor to the most recent common ancestor (MRCA) of the cells. Because very early CSCs have very few cells, the CSCs definitely require single-cell NGS to mine genomic change related with pedigree tree as described by Jiao et al. in 2014 [45]. If genomic profiles are discovered to the CSCs tumorigenesis, a new generation of therapeutic strategies including GWAS-based molecular therapy, personalized therapy, and personalized immunotherapy as all discussed earlier will appear in the treatment of tumor diseases.

6 Conclusion

Single-cell techniques with downstream genomic analysis have emerged in application of single-cell specimens from glass slides or circulating tumor cell and mixed cells tumor tissue. Recently, next-generation sequencing (NGS) has become an important tool in single-cell level. According to current R&D of single-cell NGS as given in Table 5, single-cell RNA-Seq has same significant advantages as routine RNA-Seq. As single-cell RNA-Seq is adopted to analyze transcriptome profiles, reported results include quantitative mRNA expression, RNA splicing, and new transcripts. Moreover, if RNA-Seq data using BWA platform mapping with GATK/Samtool analysis, which compare DNA reference and SNP reference of genome, are utilized to analyze the genomic profiles, they also can uncover SNP, deletion, and insertion in the exome region, so that results of single-cell RNA-Seq are much better

Table 5 Comparison of single-cell genomic techniques

Genomic types	Methods	Advantages	Disadvantages
mRNA transcriptome	Single-cell NGS	Genomic expression with splicing and exome SNP, deletion, and insertion	Bias, dropouts, and contamination
	Single-cell microarray	Genomic expression with good model for normalization	Bias, dropouts, and contamination
	Single-cell differential display	Genomic expression with very good specificity	Bias, dropouts, and contamination with lower sensitivity
DNA genomic change	Single-cell NGS	DNA level change with genetic tree discovery and new SNP discovery	Bias, dropouts, and contamination
	Single-cell SNP microarray	DNA level change with good bioinformatics support	Bias, dropouts, and contamination with limiting known SNP
	Single-cell ACGH	Chromosome level change with good SOP for clinical diagnosis	Resolution level only for chromosome and large deletion and insertion

than those from single-cell microarray. As most of single-cell genomic techniques, single-cell NGS still has three challenges: bias produced by amplification, genetic material contamination caused by heterogeneous amplification, and genomic drop-outs caused by tiny DNA materials. In order to avoid the three problems, single-cell performance definitely requires a GMP regulation with QC monitor. Technically, several single cells such as 5–10 are minimal cell numbers for DNA-Seq and several single cells are optimal selection for RNA-Seq due to cell dropout from single-cell sampling process. The new genomic technique and its analysis will be developed into diagnosis of molecular pathology and cytogenetics of cancer diseases, discovery of differentiation biomarkers of cancer stem cells, and inducing therapy for cancer stem cells; furthermore, clinical application of molecular therapy, personalized therapy, and personalized immunotherapy for cancer patients.

Acknowledgments Under the support of Dr. H.D. Preisler, we have set up the method to analyze single-cell genomic profiles of CD3, CD4, and CD8 from TIL and tumor cell from solid tumors. The work is supported by both National Cancer Institute IRG-91-022-09, USA, for Dr. Biaoru Li and AcRF Tier 2 grant ARC39/13 (MOE2013-T2-1-079) and AcRF Tier 1 seed fund on Complexity RGC2/13 (M4011101), Ministry of Education, Singapore for Dr. Jie Zheng. During the 10-year effort, Qianqing Ding, Hongliang Hu, Yunbo Xu et al. gave the work great contributions in culture TIL cell and establishment of local galaxy analysis system. Nancy S. Debry and Shen Li contribute the chapter modification. Mention of trade names or commercial products in this article is solely for the purpose of providing specific information and does not imply recommendation.

Competing interests statement: The authors declare competing financial interests.

References

1. Watson JD, Crick FH. The structure of DNA. Cold Spring Harb Symp Quant Biol. 1953;18:123–31.
2. Er TK, Chang JG. High-resolution melting: applications in genetic disorders. Clin Chim Acta. 2012;414:197–201. doi:10.1016/j.cca.2012.09.012.
3. Sanger F, Nicklen S, et al. DNA sequencing with chain-terminating inhibitors. Proc Natl Acad Sci U S A. 1977;74(12):5463–7.
4. Maxam AM. Gilbert W (1992) A new method for sequencing DNA. Biotechnology. 1977;24:99–103.
5. Pareek CS, Smoczynski R, et al. Sequencing technologies and genome sequencing. J Appl Genet. 2011;52(4):413–35. doi:10.1007/s13353-011-0057.
6. Margulies M, Egholm M, et al. Genome sequencing in microfabricated high-density picolitre reactors. Nature. 2005;437(7057):376–80.
7. Warren RL, Sutton GG, et al. Assembling millions of short DNA sequences using SSAKE. Bioinformatics. 2006;23(4):500–1.
8. Mardis ER. The impact of next-generation sequencing technology on genetics. Trends Genet. 2008;24(3):133–41.
9. Liu L, Li Y, et al. Comparison of next-generation sequencing systems. J Biomed Biotechnol. 2012;25:1364–75. doi:10.1155/2012/251364.
10. Breivik J. The evolutionary origin of genetic instability in cancer development. Semin Cancer Biol. 2005;15(1):51–60.

11. Liberko M, Kolostova K, et al. Essentials of circulating tumor cells for clinical research and practice. Rev Oncol Hematol. 2013;88(2):338–56. doi:10.1016/j.critrevonc.2013.05.002.
12. Ning L, Liu G, et al. Current challenges in the bioinformatics of single cell genomics. Front Oncol. 2014;4:7. doi:10.3389/fonc.2014.00007.
13. Ebenezer V, Medlin LK, et al. Molecular detection, quantification, and diversity evaluation of microalgae. Biotechnology (NY). 2012;14(2):129–42. doi:10.1007/s10126-011-9427-y.
14. Brück S, Evers H, et al. Single cells for forensic DNA analysis–from evidence material to test tube. J Forensic Sci. 2010;56(1):176–80. doi:10.1111/j.1556-4029.2010.01553.
15. Meier-Ruge W, Bielser W, et al. The laser in the Lowry technique for microdissection of freeze-dried tissue slices. Histochem J. 1976;8(4):387–401.
16. Schutze K, LAHR G, et al. Identification of expressed genes by laser-mediated manipulation of single cells. Nat Biotechnol. 1998;16(8):737–42.
17. Li B. A strategy to identify genomic expression at single-cell level or a small number of cells. J Biotechnol. 2005;8(1):71–81. doi:10.2225/vol8-issue1-fulltext-3.
18. Emmert-Buck MR, Bonner RF, et al. Laser capture microdissection. Science. 1996;274(5289): 998–1001.
19. Fend F, Emmert-Buck MR, et al. Immuno-LCM: laser capture microdissection of immunostained frozen sections for mRNA analysis. Am J Pathol. 1999;154(6):1857–66.
20. Fiegler H, Geigl JB, et al. High resolution array-CGH analysis of single cells. Nucleic Acids Res. 2007;35(3):e15.
21. Kurihara T, Kamberov E, et al. Rubicon PicoPlex-NGS Kits available for sequencing single cells using the Illumina Genome Analyzer. J Biomol Tech. 2011;22(Suppl):S51.
22. Pan X, Urban AE, et al. A procedure for highly specific, sensitive, and unbiased whole-genome amplification. Proc Natl Acad Sci U S A. 2008;105(40):15499–504. doi:10.1073/pnas.0808028105.
23. Ambion Catalog. Macromolecular components of *E. coli* and Hela cells. 2004. p. 192
24. Klebe RJ, Rodriguez SA, et al. RT-PCR without RNA isolation. Biotechniques. 1996;21(6): 1094–100.
25. Ramsköld D, Luo S, et al. Full-length mRNA-Seq from single-cell levels of RNA and individual circulating tumor cells. Nat Biotechnol. 2012;30(8):777–82.
26. Lobo MK. Molecular profiling of striatonigral and striatopallidal medium spiny neurons past, present, and future. Int Rev Neurobiol. 2009;89:1–35. doi:10.1016/S0074-7742(09)89001-6.
27. Hashimshony T, Wagner F, et al. CEL-Seq: single-cell RNA-Seq by multiplexed linear amplification. Cell Rep. 2012;2(3):666–73. doi:10.1016/j.celrep.2012.08.003.
28. Tang F, Barbacioru C, et al. mRNA-Seq whole-transcriptome analysis of a single cell. Nat Methods. 2009;6(5):377–82. doi:10.1038/NMETH.1315.
29. Eberwine J. Amplification of mRNA populations using aRNA generated from immobilized oligo(dT)-T7 primed cDNA. Biotechniques. 1996;20(4):584–91.
30. Toellner KM, Gulbranson-Judge A, et al. Immunoglobulin switch transcript production in vivo related to the site and time of antigen-specific B cell activation. J Exp Med. 1996;183(5): 2303–12.
31. Dixon AK, Richardson PJ. Expression profiling of single cells using 3 primer end amplification (TPEA) PCR. Nucleic Acids Res. 1998;26(19):4426–31.
32. Gole J, Gore A. Massively parallel polymerase cloning and genome sequencing of single cells using nanoliter microwells. Nat Biotechnol. 2013;31(12):1126–32. doi:10.1038/nbt.2720.
33. Landau YE, Lichter-Konecki U, et al. Genomics in newborn screening. J Pediatr. 2014;164(1):14–9. doi:10.1016/j.jpeds.2013.07.028.
34. Panagopoulos I, Thorsen J, et al. Sequential combination of karyotyping and RNA-sequencing in the search for cancer-specific fusion genes. J Biochem Cell Biol. 2014;S1357–2725(14):00176–9. doi:10.1016/j.biocel.2014.05.018.
35. Navin N, Kendall J, et al. Tumor evolution inferred by single-cell sequencing. Nature. 2011;472(7341):90–4. doi:10.1038/nature09807.
36. Hou Y, Song L, et al. Single-cell exome sequencing and monoclonal evolution of a JAK2-negative myeloproliferative neoplasm. Cell. 2012;148(5):873–85. doi:10.1016/j.cell.2012.02.028.

37. Li Y, Xu X, et al. Single-cell sequencing analysis characterizes common and cell-lineage-specific mutations in a muscle-invasive bladder cancer. Gigascience. 2012;1(1):12. doi:10.1186/2047-217X-1-12.
38. Kim KI, Simon R. Using single cell sequencing data to model the evolutionary history of a tumor. BMC Bioinformatics. 2014;15:27. doi:10.1186/1471-2105-15-27.
39. Xu YB, Hu HL, et al. Feasibility of whole RNA sequencing from single-cell mRNA amplification. Genet Res Int. 2013;2013:724124. doi:10.1155/2013/724124.
40. Macaulay IC, Voet T. Single cell genomics: advances and future perspectives. PLoS Genet. 2014;10(1):e1004126. doi:10.1371/journal.pgen.1004126.
41. Meldrum C, Doyle MA, et al. Next-generation sequencing for cancer diagnostics: a practical perspective. Clin Biochem Rev. 2011;32(4):177–95.
42. Guan YF, Li GR, et al. Application of next-generation sequencing in clinical oncology to advance personalized treatment of cancer. Chin J Cancer. 2012;31(10):463–70. doi:10.5732/cjc.012.10216.
43. Li B, Senzer N, et al. Approach to individual cancer target identification. In: 11th annual meeting of the American Society of Gene Therapy. 2008;8(45):1004
44. Zhang W, Ding JQ, et al. Genomic expression analysis by single-cell mRNA differential display of quiescent CD8 T cells from tumour-infiltrating lymphocytes obtained from in vivo liver tumors. Immunology. 2009;127(1):83–90. doi:10.1111/j.1365-2567.2008.02926.
45. Jiao W, Vembus S, et al. Inferring clonal evolution of tumors from single nucleotide somatic mutations. BMC Bioinformatics. 2014;15:35. doi:10.1186/1471-2105-15-35.

Utility of Next-Generation Sequencing in Cancer Drug Development and Clinical Trials

François Thomas and Ahmad Awada

Abstract Next-generation sequencing (NGS) has great potential to tailor the treatment of patients to their cancer genome alterations. Case reports, retrospective analysis of phase I trials, open studies of targeted therapies in population enriched in particular genotypes, and series of breast and lung cancer patients have shown encouraging clinical outcome for the matching of drugs to specific molecular alterations.

Ongoing clinical trials are testing how NGS of tumors can guide individualization of treatment and whether the integration of the NGS into patient care can translate into superior patient outcome. The use of NGS comes with multiple challenges such as access to tumor material, data interpretation, and adaptation of regulatory frameworks for drugs targeting small population and for complex molecular diagnostics. Analytical validation of sequencing platforms and gene panels, access to multiple therapies addressing new targets and development of blood-based tests will support the expanding role of NGS in drug development and clinical trials.

1 Introduction

The recent and often rapid registration of anticancer drugs that target specific proteins from mutated cancer genes on the basis of superior activity in the only patients that have the corresponding alterations supports the value of precision medicine. In that context, next-generation sequencing (NGS) has great potential for providing data on cancer genes for targeting drug prescription to patient's genetic abnormality and for guiding the development of new drugs. This review summarizes the current data on NGS in the settings of clinical trials and drug development in oncology.

F. Thomas (✉)
Thomas Conseil SPRL, 46 Avenue des Villas, Brussels 1060, Belgium
e-mail: thomasconseil.be@gmail.com

A. Awada
Department of Medicine, Institut Jules Bordet, 121 Bd de Waterloo, Brussels 1000, Belgium

W. Wu, H. Choudhry (eds.), *Next Generation Sequencing in Cancer Research, Volume 2*, DOI 10.1007/978-3-319-15811-2_2

2 Next-Generation Sequencing in the Clinical Setting

2.1 Process and Interpretation

The ability of NGS to interrogate multiple gene sequences is likely to replace more standard technologies based on multiplex polymerase chain reaction (PCR) for tumor genotyping. The amount of information presently available on cancer genomes [1] and the need to better ascertain cancer genes to predict the activity of targeted therapies support the clinical development of NGS which comes with challenges. Core biopsies of tumors should provide sufficient quantity of cancer cells. Dilution by stromal cells and tumor heterogeneity [2] requires sequencing with high coverage. Technologies for clinical applications must meet the attributes of diagnostic tests [3] and come with robust protocols from sample preparation to analysis. Validation of analytical performance metrics [4] is needed. Orthogonal technologies such as Sanger sequencing and high sensitivity PCR or comparison of sequence with known reference material should be used for that purpose [5]. Validation testing has demonstrated false positive variant calls for certain genes with low variant frequency in some gene panels [6]. Quality control metrics need to be determined for each step of the process including sample library preparation, fragment amplification, sequencing, and data analysis [7]. These requirements increase the cost of laboratory developed tests (LDT) when performed in clinical laboratory improvement amendment (CLIA) certified laboratories in the USA, a setting required when the molecular data are used for taking clinical decisions [3]. Certification and accreditation under CLIA regulation ensure that clinical labs meet certain quality standards. LDT do not require premarket evaluation or clinical validity.

The multiple steps are associated with relatively long "turn around time" from tissue collection to data interpretation. High throughput, automation of preparation steps, and center experience are likely to reduce the time to results below 2–3 weeks, a timeframe acceptable for clinical use.

Determining the clinical correlation between genomic variations and alterations and phenotypes is a major issue for practical use [8]. Clinical interpretation is difficult in an area where both technologies and scientific data are in very rapid flux. Well-characterized effects of genetic alterations are presently assessed by validated technologies that guide the treatment of registered targeted therapies [9] (Table 1), and potential therapeutic consequences are far less established for newly discovered mutations. The tyrosine kinase inhibitors (TKI) vemurafenib or dabrafenib registered for the treatment of *BRAF* V600E mutated metastatic melanoma have demonstrated activity in other *BRAF*-mutated tumors such as lung cancer [10] but not in colorectal cancer where the activation of the epidermal growth factor receptor (EGFR) pathway [11] is responsible for TKI resistance. Different mutations may activate a tyrosine kinase but available TKI may only work on some of them as it is the case for EGFR mutations [12]. Reports to clinicians need to present the strength of evidence for drugs and drug candidates available in clinical trials on top of the mutation effect on gene functions and pathways. The interpretation of adequately

Table 1 US FDA-approved molecularly targeted drugs for solid tumors

Drug	Primary molecular target(s)	Main indications
Ado-trastuzumab emtansine	HER2	**HER-2** positive metastatic breast cancer (MBC)
Afatinib	EGFR, HER2, HER4	**EGFR** mutated non-small cell lung cancer (NSCLC)
Axitinib	PDGFR, VEGFR	Renal cell carcinoma (RCC)
Cabozantinib	MET, RET, VEGFR, KIT, FLT3, TRKB, AXL	Medullary thyroid cancer (MTC)
Ceritinib	ALK	**ALK** positive NSCLC
Cetuximab	EGFR	**RAS** wild type colorectal cancer (CRC), advanced head and neck squamous cell carcinoma (HNSCC)
Crizotinib	ALK, ROS1, MET	**ALK** positive NSCLC
Dabrafenib	BRAF	V600E **BRAF** melanoma
Dasatinib	BCR-ABL, SRC	Chronic myeloid leukemia (CML)
Erlotinib	EGFR	**EGFR** mutated NSCLC
Everolimus	mTOR	RCC, MBC
Gefitinib	EGFR	**EGFR** mutated NSCLC
Imatinib	BCR-ABL, KIT, PDGFR	CML, **KIT** positive gastrointestinal stromal tumor (GIST)
Lapatinib	HER2, EGFR	**HER-2** positive MBC
Panitumumab	EGFR	**RAS** wild type colorectal cancer
Pazopanib	VEGFR, PDGFR, KIT	RCC
Pertuzumab	HER2	**HER-2** positive MBC
Regorafenib	VEGFR, RET, KIT, PDGFR, RAF	Colorectal cancer; GIST
Sorafenib	RAF, VEGFR, PDGFR	Hepatocellular, thyroid carcinomas, RCC
Sunitinib	VEGFR, PDGFR, KIT, RET, FLT3	RCC, GIST and neuroendocrine tumors of the pancreas
Temsirolimus	mTOR	Mantle cell lymphoma
Trametinib	MEK	V600E **BRAF** melanoma
Trastuzumab	HER2	**HER-2** positive breast cancer
Vandetanib	VEGFR, EGFR, RET	MTC
Vemurafenib	BRAF	V600E **BRAF** melanoma
Vismodegib	Smoothened (Hedgehog pathway)	Basal cell carcinoma of the skin

Sources: http://fda.gov-cancer.gov
Texts in bold represent molecular marker for prescription

processed data is an issue due to insufficient experience outside of very specialized centers. Some centers have set ad-hoc boards with members bringing complementary scientific skills to interpret data and suggest clinical strategies [13, 14], CLIA laboratories such as Foundation Medicine Inc. are providing NGS to a broad range of physicians such as US community oncologists who order 60 % of the test. The consequences of such data flow in terms of treatment consequences and patient outcome remain to be established by prospective and registry studies (Table 2).

Table 2 Examples of prospective trials of predictive biomarkers for targeted therapies

Study	Setting	Testing	Design (and early results)	Reference (clinicaltrials.gov)
BATTLE1[a] (completed)	Pretreated metastatic NSCLC ($n=255$)	Five genes analyzed	• Four targeted Rx regimens unselected on biomarker followed by biomarker-driven allocation (adaptive design)	Cancer Discov. 2011;1:45
BATTLE2[a] (ongoing)	Refractory NSCLC ($n=200+200$)	NGS analysis	• Allocation to four targeted therapies arms to treat EGFR wild type NSCLC based on molecular results • Two stage study: randomization to identify biomarker for best individual treatment followed by adaptative randomization based on biomarker from phase 1	Clin Cancer Res. 2012;18:638 NCT 01248247
CUSTOM[a] (ongoing)	Advanced lung and thymic tumors ($n=600$)	Initially 12 genes, genotyping, FISH, 197 gene exons by NGS	Study of molecular alterations useful to assign treatment with five molecular therapies in three advanced thoracic tumors (15 treatment arms)	ASCO 2013; Abst. 7513 NCT 013060145
FMI registry[b] (ongoing)	Prospective observational study (500 patients) of Foundation One test	NGS of (314+29 gene) panels	Impact of the test on subsequent treatment pattern and clinician report outcomes	NCT 01851213
FOCUS4[a] (ongoing)	• Maintenance therapy after first line CT of CRC • Testing of four molecularly defined groups of new targeted drugs for CRC	Genotyping, IHC (so far)	Adaptive phase 2/3 trial with a randomization against a control for each biomarker/treatment group	J Clin Oncol. 2013;31:4562
IMPACT2[b] (ongoing)	Metastatic cancer; up to three prior therapies (1,360 patients screened for 300 eligible)	NGS of 315+28 cancer genes	Randomization between targeted therapy (off-label use or clinical trial) and treatment not selected on genetic profiling (phase 2)	NCT 02152254 ASCO Educational Book 2014:61–69
ISPY-2[a] (ongoing)	Neoadjuvant treatment of breast cancer ($n=800$)	71 genes	Eight experimental drugs on top of standard therapy with prespecified genomic signature at entry (adaptive phase 2/3 design)	NCT 01042379

Lung-MAP[a]	Advanced squamous cell carcinoma of the lung (second line after platin-based chemotherapy) 500/1,000 patients screened/year	Gene panel of Foundation Medicine (314 + 29 gene) panel	• Value of five targeted therapies compared to docetaxel (or erlotinib) • Allocation by biomarker in five matched substudies in a phase 2/phase 3 adaptative trial on top of substudy of screened patients not eligible for biomarker-driven substudies	NCT 02154490 ASCO Educational Book 2014:71
MATCH 1 (completed[b]; ongoing[c])	Patients with metastatic breast, colorectal and gynecological tumors, a biopsiable lesion and candidates for phase I/II clinical trials ($n = 485$)	• NGS of gene panel • Mass spect genotyping	Percentage of acceptable core or fine needle biopsies, successful analysis and percentage of actionable results	NCT 01703585 ASCO 2013; Abst. 11002
MATCH2[c] (ongoing)	Umbrella protocol with multiple molecularly based phase 2 studies (30 patients per arm)	NGS of 200 genes	Phase 2 of 20–25 targeted agents approved in other indications with evidence of activity against a known target in a particular tumor type	ASCO Educational Book 2014:71
MOSCAT01[b] (completed)[c]/ (ongoing)	Previously treated patients with metastatic cancer ($n = 900$)	CGH array and gene (30) sequencing	• Progression free survival (PFS) with a targeted treatment selected by molecular profiling compared to the PFS for the most recent treatment • Filter for entry into phase 1 and 2 trials	NCI 01566019 ASCO 2013; Abst. 2512
SAFIR01 trial[b] (completed)	Metastatic breast cancer patients ($n = 427$) with metastases amenable to biopsies	Array CGH and PI3KCA/AKT mutations (Sanger sequencing)	• Feasibility of CGH array (67 %) and sequencing (70 %) • Frequency of targetable alterations (46 %) • Percentage of patients (13 %) receiving matched targeted therapy (due to insufficient access to relevant drugs)	Lancet Oncol. 2014;15:267
SAFIR02 breast (to be started)[b,c]	Metastatic non-HER2-positive breast cancer (phase 3; $n = 400$)	• CGH gene panel • NGS	Randomization between targeted therapy based on genomic alteration or maintenance CT in patients not progressing after 6–8 cycles of CT	ASCO Educational Book 2014;71

(continued)

Table 2 (continued)

Study	Setting	Testing	Design (and early results)	Reference (clinicaltrials.gov)
SAFIR02 lung[a] (ongoing)	NSCLC maintenance therapy after first line (phase 2; $n=650$)	• CGH gene panel • NGS	Randomization between six targeted therapies based on genomic alteration or standard maintenance after 4–6 cycles of CT	NCT 02117167
SHIVA[b] (ongoing)	Recurrent/metastatic tumors ($n=1,000$)	Not specified	Comparison of standard therapy to targeted treatment in patients with actionable mutations	NCT 01771458
SIGNATURE[c] (ongoing)	• Molecular phase 2 testing • At least eight drugs • Basket trials of multiple indications	Molecular profiling in CLIA labs demonstrating pathway validation	• Patients with gene alterations targetable by eight different drugs that have not shown activity or lack thereof in a particular indication • Seventy patients per trial	www.signaturetrial.com NCT01833169 NCT01831726 NCT01885195 NCT01981187 NCT02002689
WINTHER[b,c] (ongoing)	Advanced malignancies ($n=200$)	Gene sequencing, gene expression analysis	• Patients with actionable DNA aberrations receive existing targeted therapies or are included in phase I trials • Patients without actionable aberrations receive a therapy based on a RNA profiling algorithm	www.WINconsortium.org NCT 01856296

[a]Search for biomarker and adaptive design to validate predictive test and drug activity
[b]Feasibility and value of multiplex analysis
[c]Screening for entry into phase 1/2 trials matching targeted therapies with molecular aberrations

Fixed costs are significant in terms of investment and personnel, and come on top of variable tumor procurement and reagent costs. They can partly be addressed by infrastructure sharing and outsourcing. Reimbursement of NGS under test specific codes will require demonstration of clinical utility [15] including the ability to replace multiple single gene tests for the prescription of targeted therapies.

2.2 The Different Types of Genome Sequencing

Sequencing of cancer specific gene panels is becoming common in academic cancer centers and molecular diagnostic service companies. The first gene panels have been based on allele-based (genotyping) technologies. They cover a relatively narrow numbers of genes and actionable mutations (for example 19 genes and 238 mutations studied by mass spectroscopy genotyping technology in Sequenom Oncocarta V1.0 [16]). Actionable mutations are mutations with potential clinical consequences of prognostic value (to modulate treatment intensity) or of predictive value (for the response to a drug). Actionable mutations with potential predictive value can be targeted by drugs or be in a pathway in which key other members can be targeted. Good examples are available for melanoma, breast and non-small cell lung cancers [17].

The number of studied genes for solid or hematological malignancies in NGS panels may vary from about 50–400, or may be smaller with a particular focus on specific tumor types. In theory, there is no need to sequence the germ-line DNA to provide a control for known cancer genes. FFPE tumor material is adequate. Limited coverage breadth allows for high coverage depth to detect lower frequency somatic variant. Gene panels are supported by the relatively small number of cancer genes, being either oncogenes or tumor suppressor genes identified so far [1, 18]. Detection of rearrangement is possible with some panels that provide intron baits to capture known rearrangement [5, 19].

Whole-genome sequencing (WGS) remains expensive as high redundancy is needed to provide sufficient coverage due to dilution by normal cells and genetic heterogeneity. The large body of data is difficult to store and to interpret in the clinical setting. Furthermore, WGS should be performed on high-quality DNA that cannot be obtained from FFPE tissues. WGS is unlikely to be used for clinical application in the short to mid-term, despite some pilot studies [13]. WGS of germline DNA raises the question whether, how, and which information on disease susceptibility genes of unclear consequences should be reported to patients [20]. Whole-exome sequencing (WES) requires parallel sequencing of the germ-line DNA, but can be performed on FFPE tissues. WES will likely be used for potential clinical applications in centers of excellence. The Boston groups recently identified 15 relevant alterations in 16 patients enrolled in a prospective study [21].

Transcriptome sequencing provides data on gene rearrangement and splice variants on top of RNA abundance of mutated genes and drug targets [22]. RNA expression also provides information on the microenvironment that plays an important role in invasion/metastasis, and resistance to anticancer drugs.

3 Early Experience with Predictive Multiple Gene Testing

3.1 Clinical Data Supporting Potential Value

Earlier generation genotyping has been tested on a large scale in multicentric studies (Table 2). The use of those platforms has provided very important information on the feasibility of genomic projects and has driven a larger acceptance of collecting tumor material in the context of clinical trials. In the SAFIR 01 trial [23] that enrolled metastatic breast cancer (MBC) patients with metastases amenable to biopsies, about 46 % of patients had targetable alterations, but only 13 % receives matched targeted therapy emphasizing the needs to use more sensitive detection technologies and to access a sufficient number of targeted drugs.

Several large genotyping studies support the value of patient triage to enrol patients in phase I or II protocols. The MD Anderson reported its phase I experience in 1,144 patients [24]. Patients who enrolled into a trial of a drug targeting a genetic abnormality in their tumor had a higher response rate than with their previous treatment, longer time to treatment failure and survival than patients who enrolled into trials of agents for which molecular matching was not possible. The marked difference of outcome in the context of phase I studies should facilitate the enrolment of patients into phase I as more patients have access to molecular profiling and benefit from trial participation.

Phase II trials have tested different drugs with a treatment allocation based on molecular profiling. Table 2 describes the study main characteristics (disease, line of treatment, biomarkers, drugs, end points). Nevertheless, protocols may evolve [25] incorporating new data emerging from outside of the trial, technologies and drug candidates in rapid flux. BATTLE in non-small cell lung cancer (NSCLC) used an adaptative design. Randomization to 4 drugs in the first 97 patients was followed by an assignment based on multiplex genotyping in the next 158 patients [26]. This first prospective trial has been followed by others in NSCLC due to the number of actionable mutations in that disease [16, 17, 24].

Individual patients have benefited from sequencing information on actionable mutations and disease pathways. A case report showed superior sensitivity of NGS over cytogenetic techniques to identify the gene rearrangement typical of acute promyelocytic leukemia, a rare leukemia that can be cured by targeted therapy [27]. In other cases [28], unpredicted and extraordinary activity of a targeted agent has been explained by specific gene mutations. These cases support what has been recently named phenotype to genotype "n of 1 studies" of individual cases. The US National Cancer Institute (NCI) is studying such outliers that provide hypotheses for prospective clinical studies [29].

The largest experience in performing NGS of cancer gene panels is that of Foundation Medicine Inc. (www.foundationmedicine.com) in collaboration with several academic centers. Genomic DNA can be extracted from FFPE tumors with precise guidelines for eligibility [5]. The technology based on academic work [30] has been optimized and validated to detect more than 5 % of mutant allele frequency

of base substitutions and more than 10 % of indels with 99 % accuracy [5]. The company is rapidly increasing the number of captured sequences (from 183 then 236 and now 343 genes). The platform has potential to discover new actionable mutations, study the profile of primary vs metastatic tumors or allocate genotype directed treatments [5, 19, 31]. Foundation Medicine has entered multiple agreements with pharmaceutical companies to use the test in clinical trials of targeted agents.

Iterative measurement can be provided by NGS of circulating tumor (ct) DNA. The low amount of tumor DNA in plasma DNA presently supports the sequencing with high redundancy of a limited set of genomic alterations [32]. ctDNA in plasma can be quantified and allowed a more sensitive readout than circulating tumor cells in a series of patients with metastatic breast cancer [33]. Access to ctDNA may provide an integrated view of the cancer genome landscape [34] in the context of tumor heterogeneity [2]. Serial analysis of ctDNA can track early genomic evolution [35] of metastatic cancer in response to therapy. Quantifications of allele fractions in plasma identified increased representation of mutant alleles with emergence of drug resistance [36] and support the potential of plasma DNA sequencing to study clonal evolution. Of note, ctDNA has potential to detect minimum residual disease and early relapse [37]. Analysis of circulating tumor cells (CTC) will be complementary with the improvements of technologies to isolate CTC. As of today, ctDNA has the greatest attribute of ease of collection and high throughput analysis [37].

3.2 Challenges for Clinical Studies

Data on potentially actionable mutations are mostly valuable when targeted agents are available (Table 2). Those agents may be marketed and their use supported by safety data but cost and reimbursement are issues for off-label use (Table 1). Off-label prescription is common (30 %) in the treatment of patients with metastatic cancer in the USA [38] and often reimbursed if supported by accepted (e.g., National Comprehensive Cancer Network, NCNN) guidelines. Observational studies using quality assured clinical registries are needed to evaluate both activity and safety in new indications. Temporary recommendation for use [39] can provide a framework to insure monitoring of activity and safety, and access to reimbursement.

Access to tumor material can be an issue for metastatic sites. Studies [40] have reported the risk of major complications (need for hospitalization or surgery) of 0.8 % after core biopsies of thoracic and abdominal tumors. Improvements in both DNA purification and sequencing technologies ("single cell" sequencing [7]) may in the future allow the analysis of cells from fine needle aspirates.

Tumor subclones (within a tumor and between primary and metastases) create challenges for predictive biomarkers. Spatial and temporal variability of validated biomarkers such as *HER-2*, hormone receptors, *EGFR* activating mutations and *KRAS* mutations is well known and supports deep sequencing of metachronous

metastases on top of primary tumors, despite the cost and morbidity of imaging-guided core biopsies. Analysis of the primary tumor in patients with metastases can presently provide sufficient information, when the material has been recently and appropriately stored and the patient not treated by drugs that may select clonal evolution. Repeated biopsies may be needed to predict and understand resistance to targeted therapies and will benefit from the development of blood-based tests [32–34]. *KRAS* mutant alleles can be detected in the serum before clinical resistance to anti-*EGFR* antibodies in colorectal cancer [41].

The integration of the multistep process from genomic profiling to patient management [8, 14, 17] raises organization issues [42]. The clinical experience with genomic studies has mainly been restricted so far to large academic centers. Smaller centers may either use CLIA laboratories with validated LDTs. The difficulty of clinical interpretation and the access to a large pool of patients for clinical trials support the collaboration of the different providers. The developments of shared platforms as supported by the Institut National du Cancer in France [43]. Multicentric early clinical trials are needed.

4 Design of Clinical Trials Incorporating NGS

Several designs have been proposed to investigate the related role of genetic predictive biomarkers and drug candidates (Tables 2 and 3). Availability of well-characterized tumor material is mandatory [40] for patient enrolment. Biopsies require institutional review board (IRB) approval and patient consent. The designs of the trials are based on the information available on the predictive value of the test, and the availability and activity of drug candidates on different targets. Studies provide different answers on drug effect, biomarker effect, biomarker by treatment effect, and the strategic value of complex gene analysis [44].

4.1 Protocols to Support Drug Activity in Biomarker-Defined Populations

The development of therapies targeting a particular genetic alteration supports the inclusion of patients bearing that molecular alteration in enrichment trials [9, 45, 46]. Only patients with a certain molecular-defined tumor type are enrolled. This strategy increases the power of the trial to demonstrate drug activity. The consequences of selection are the needs to define and develop early the predictive marker and to screen a larger population to select adequate patients. Fortunately, the larger treatment effect expected from the enrichment on biomarkers with strong credentials from early trials reduces the number of patients to be enrolled in a randomized trial. Strong activity in single arm trial may also support accelerated US approval, as for crizotinib in NSCLC with *ALK* rearrangements [16]. Enrichment trials have some

Table 3 Biomarker clinical strategies

Biomarker	Randomization	Treatment allocation	Biomarker	Consequences
"Histology agnostic" or basket clinical trial	None	The only patients with positive biomarker receive new treatment Patients are stratified by site of origin	Biomarker needs to be well defined	• Basket trials adapted to rare mutations in multiple cancer types • Adaptive design to close cohorts of inactive drug
Enrichment design	None (phase 1/2)	The only patients with positive biomarker receive new treatment (single arm phase II) or are randomized between several new treatments (including combinations) in adaptative phase II, or are randomized between new treatment and standard of care (phase III)	Biomarker needs to be well defined and to have high predictive value	• No possibility to study drug effect in patients with negative biomarker • Examples: phase 3 of trastuzumab (HER2), EGFR TKI (EGFR mutations); vemurafenib (BRAF), crizotinib (ALK)
Phase 3 for drug (retrospective analysis by biomarker)	New drug (or combination) versus control (standard) treatment	Classical randomized trial not affected by biomarker	• No predictive biomarker at time of study design • Retrospective analysis raises issues of sample availability, and potential bias	• Used to demonstrate the effect of (*K*)*RAS* mutations on the activity of anti-EGFR antibodies in CRC • Retrospective nature requires strong biological rationale and at least two positive studies for validation
Randomize all phase 3 trial with stratification on biomarker	New drug (or combination) versus control (standard) treatment	• Randomization not affected by biomarker	• Predictive biomarker with promising but uncertain predictive value • Stratification by biomarker allows to study biomarker effect in all patients and/or subgroups and its predictive and prognostic values	• Predictive value • Probability of biomarker effect impacts statistical plan and trial size
Biomarker strategy design (to assess the value of multiplex analysis)	Biomarker guided treatment versus non-guided control therapy	• Patients with positive biomarker receive new agent(s) (or off-label) • Patients with negative biomarker(s) and patients in non-guided control group receive standard therapy (various combination of patient flow)	• Validate complex biomarker strategy • Effect dependent on the availability and activity of targeted drugs	Useful to test complex biomarker leading to an array of targetable alterations

drawbacks such as limits to refine biomarker [e.g., format and cutoff for defining the biomarker-positive population, as illustrated by the value of fluorescence in situ hybridization (FISH) and immunohistochemistry (IHC) threshold for HER-2 positivity for the prescription of trastuzumab], and to restrict information on drug effect and its label to the biomarker-positive group.

Aberration-specific but histology-independent trials are referred as basket trials. Such trials enrol patients with specific markers and not specific histologies. The activity may not be observed in all tumor types as the effect of molecular aberration is possibly disease specific. This basket design should ensure a sufficiently broad inclusion of patients with different histologies followed by enrichment in tumor types for which early signs of antitumor activity have been shown. A good example of this design is provided by the Signature phase 2 studies of Novartis that link eight targeted therapies (http://www.signaturetrial.com) to different pathways (alterations in more than 30 genes). Seventy patients previously profiled in CLIA laboratories will be treated in each trial by a drug for which safety data exist and phase 2 dose is defined. Protocols exclude indications for which drug activity or lack of activity has been demonstrated, or presently studied in phase 3 trials.

For more frequent tumors with unvalidated predictive markers, testing of multiple agents in patients that had extensive molecular characterization, allows to validate predictive markers in conjunction with the administration of drugs targeting specific pathways. Adaptative design strategy can match patients in a second phase (BATTLE1 trial in NSCLC [26]) and close early treatment arms with limited activity or add new experimental drugs (I-SPY-2 trial in high-risk breast cancer treated by neo-adjuvant chemotherapy [46]). The neo-adjuvant setting is of value to test drug combinations as illustrated by the registration of pertuzumab in addition to trastuzumab in HER-*2* positive breast cancers [47, 48]. Two drugs, carboplatin and neritinib (HER-2 TKI) have graduated phase 2 in I-SPY-2 cohorts of triple negative and HER-2 positive breast cancer, respectively.

Analysis of treatment activity by gene alteration is becoming standard for early stage trial. Gene panels are extremely useful in this regard and explain the partnership of pharmaceutical companies with CLIA providers of tests such as Foundation Medicine and Illumina. Retrospective analysis by marker is important in the analysis of phase 2 trials of drug candidates with the risks of false positives with multiple biomarker testing (need to control the alpha error rate) and selection bias (adjustment for confounders is useful) and false negatives with small trial size. The retrospective analysis by marker of randomized trials between two treatment strategies provides valuable information if assays can be performed in all or most patients. At least two positive studies, a strong biological rationale, access to quality material for most patients, a validated test are required to infer causality. These criteria were met to demonstrate the activity and restrict the license of anti-EGFR antibodies for the treatment of colorectal cancer patients with wild-type (*K*)*RAS*.

In "Randomize all patients stratified by the biomarker" phase 3 trial, a new treatment is compared to standard treatment in all patients and the drug effect is analyzed by marker presence. This design is used when the evidence for the predictive

marker is not sufficiently compelling to rule out a clinically meaningful effect in biomarker negative patients. This design is usually not recommended if the effect of the drug is likely restricted to the biomarker positive patients and/or if the biomarker has low prevalence, in which cases an enrichment strategy is preferred. There are several ways to analyze such trials [49]: (1) separately in each biomarker positive and negative populations, (2) in the overall and in the biomarker positive patient population (sequentially or in parallel), (3) in the biomarker positive population and only in the negative population, if the drug is active in the biomarker positive group (sequential analysis). The statistical plan impacts the size of the trial (allocation of the overall false positive error rate) and risk of recommending an ineffective treatment for the biomarker negative subgroup.

A particular design (US National Cancer Institute initiative) is the study of exceptional and excellent responders to a drug candidate and for whom adequate tumor tissue is available for WES [28, 50].

In conclusion, aberration-specific histology-independent trials, testing of multiple agents in patients having a particular disease with an adaptive strategy, retrospective analysis by biomarker of phase 2 trials, analysis of exceptional responders in otherwise negative trials, enrichment trials for biomarkers with very strong credentials are different and new strategies for early drug development. NGS of gene panels are very promising in most of the settings. For late stage (phase 3) trial, sponsors can use enrichment trials for biomarkers validated in phase 2, stratification by biomarker in randomize all trials for biomarkers with reasonable credentials, or perform conventional randomized phase 3. Retrospective analysis of molecular alterations (e.g., by NGS) may be valuable to generate correlative data on biomarker to be confirmed in another clinical trial.

4.2 Protocols to Support the Value of Complex Biomarkers

So far, early experience or retrospective analysis of the use of large (Foundation Medicine) gene panels [4, 51, 52] suggest that about 70–80 % of patients will harbor at least one genetic alteration linked to potential treatment options and that 20–30 % of patients will receive genotype-directed treatments, mostly in the context of clinical trials, with few patients (less than 10 % of the total) achieving objective tumor regression.

Several trials compare the outcome of treatments guided or not by biomarkers (Table 2). The outcome will likely show a superiority of gene analysis, if a sufficient number of active targeted therapies are available off-label (see Table 1 for a list of registered drugs) or through clinical trials, as it was the case in a large retrospective series of NSCLC patients [53]. The number of patients to be enrolled takes into account assumptions on the percentage of patients with targetable aberrations, the percentage of aberrations for which targeted therapies are already registered, and the percentage of aberrations for which drugs can be offered off-label or through clinical trials. One out of 4/5 screened patients may be eligible for randomization [54].

5 Development of Drug and Diagnostic Devices

5.1 Molecular Tests for Drug Prescription

Pharmaceutical companies have reluctantly embraced disease segmentation due to initial concerns on market potential. The commercial success of orphan drugs, the high clinical failure rate of anticancer drug candidates in broadly defined population and the rapid registration of targeted drugs in biomarker-defined patient population with high medical needs has dramatically changed the picture. Phase 3 that used predictive biomarkers are able to produce the highest relative improvement of overall survival and progression-free survival (PFS) [55]. The new US Food and Drug Administration (FDA) legislation of breakthrough therapy [56] has further accelerated the US development of innovative agents with early sign of clinical activity through industry communication with FDA on expedite development programs. Demonstrating cost effectiveness through health technology assessment is becoming an important bottleneck beyond registration. High activity supporting high pricing is unlikely to be demonstrated in non-selected patient populations. Regulatory authorities [57] are willing to consider new paradigms in order to facilitate access to active drugs for cancer patients with high medical needs, including access to non-registered drugs [39, 58] and accelerated or full registration on relatively small open trials if the drug candidate is associated with high quality response rates in cancers with specific gene alterations [56, 57].

These regulatory changes are likely to have a significant impact on the development of precision medicine for the treatment of molecularly defined tumor subsets. The recent simultaneous registrations of drugs and of their companion diagnostics in the USA have emphasized the challenges involved in development execution which include the number of patients to be screened when few patients are eligible, the cost of diagnostic test development, the needed infrastructure to collect and process tumor samples and regulatory requirements. The latter in the USA includes the desirability to co-develop under investigational new drug (IND) and IDE (the equivalent of IND for diagnostic devices) regulations, the drug and the diagnostic [59]. Such predictive test is considered of significant clinical risk [3] and requires premarket approval that comes with expensive analytical and clinical development to support the test performance for selecting patients for the corresponding drug. CE marking for marketing devices in EU only requires a manufacturer declaration that its product complies with regulations and there is not yet a requirement that a companion diagnostic be approved by European Medicine Agency (EMA) before or after a corresponding drug is approved. For example, vemurafenib label in melanoma for patients with *BRAF* V600E mutation specifies in the USA but not in EU the needs for detection with an approved (FDA) test. A diverse range of proprietary and "in-house" tests that are consistent with the marketing authorization are used in EU for that purpose. Nevertheless, the proposed new legal framework for CE marking for class C (high individual risk) in vitro diagnostic will support the checking by a notified body of the analytical test and clinical performance and of utility [60].

Return on diagnostic investment is and will be supported by large volume of screening tests, as payers require a positive predictive test prior to approving reimbursement for certain expensive drugs.

5.2 Registration of Multigene Tests

FDA is planning to regulate high risk LDT that guides clinical decisions for severe diseases and particularly in the context of personalized medicine [61]. FDA is concerned that compliance with CLIA regulations alone does not ensure that diagnostic devices are safe and effective. CLIA labs do not have to assess quality manufacturing of LDT, and to ensure their analytical validations before clinical use, and mostly to assure that LDT have been properly clinically validated. Interpretation of this guidance will restrict the number of providers and support the use for drug development of the only high risk tests that passed regulatory oversight. Furthermore, multigene LDT with the same intended use as an approved companion diagnostic (e.g., for the testing of *BRAF*, *RAS*, *EGFR*, …) will need premarket review.

NGS of gene panels raise the issues of validation of tests and molecular diagnostic instruments that have initially been developed for research purposes [57]. The FDA approved in 2013 the first diagnostics using NGS (to detect multiple mutations in the cystic fibrosis gene) as well as the Illumina MiSeqDx for development of NGS diagnostics, setting the stage for multiple developments [62]. Assay validation is key and should document sequence accuracy, variant accuracy, false positive and variant discrepancy rates [4]. Such requirements may apply to a relatively short list of targeted genes [63] but not to broad scope analysis such as large gene panels or WES because of the difficulty to develop measures of acceptable performance [4]. Main barriers for validation include the absence of bioresources with a large collection of annotated variants, prohibitive costs, higher sensitivity of NGS compared to Sanger sequencing, and the artifacts caused by formalin fixation in FFPE tumors. The latter requires validation in clinical specimens in an end-to-end setting [5, 63]. Even if a commercial panel has been extensively validated by its provider, the diagnostic laboratory must perform and document an in-house verification procedure including third party bio-informatics resources [63].

The validation by a CLIA laboratory of its panel of 287-cancer-related genes [5] provides a good example of procedures for the detection of base substitution, indel and copy number changes, measures of detection performance, comparison with other variant calling “approaches” and test platforms, and reproducibility. Analysis of 2,221 FFPE specimens provided an overall evaluation of the assay performance to detect genetic alterations, possibly targetable across multiple tumor types.

Leading US cancer centers are using NGS of gene panels instead of approved tests in order to avoid multiple testing. Targeted drugs so far have been developed and approved with single gene specific tests and the potential of NGS of broad gene panels to replace gold standards is presently unclear due to longer turnaround time, FDA enforcement on high risk and complex tests [57, 64] and reimbursement

barriers. Nevertheless, FDA is willing to work with diagnostic sponsors to assure that NGS will represent the status of the biomarker in the same way as the standard technology, as it realizes that a test per drug will rapidly become impractical with multiple development efforts converging on the same target and multiple targets recognized within a disease entity. Clinical samples with associated outcome will be key for that validation [64]. At the present time, it is likely that companies will still develop specific gene tests selected out of larger panels used in the earlier phase of clinical development.

Reimbursement of complex molecular diagnostic tests is still unclear in most countries. US insurers often cover LDT performed in CLIA laboratories when they are endorsed by established medical organizations such as the American Association of Clinical Oncology (ASCO) or NCCN. They support the use of gene expression profiling to guide the adjuvant treatment of breast cancer while cost impacts the lower and variable acceptance across European countries (from less than 20 % to up to 80 %). EU oncologists quote lack of reimbursement (51 %), price (31 %) but not lack of evidence (19 %) as the main reasons for non-utilization [65]. Cost efficiency for NGS will depend on the number of sample gene tests needed for a particular indication and to a lesser extent on the benefit of multiplexing over multiple single-gene tests in terms of sample quantity and to the lesser tolerance to suboptimal quality DNA. In that context, publicly funded laboratories may serve as regional hub for expert and cost efficient molecular testing in countries like the UK (Cancer Research UK; http://www.cancer.org.uk) or France [43]. The results of the characterization of six genes alterations in 10,000 NSCLC patients [66] illustrate the value of such coordination. Those platforms may solve issues of access and reimbursement.

6 Conclusion

Tumor profiling is changing the development of drugs that target specific molecular alterations. NGS by providing information on multiple genes is very promising in this setting, but comes with challenges in terms of analytical validation and medical interpretation.

The multiplicity of genomic alterations discovered by NGS supports clinical trials of multiple drugs and their combinations, whereas the development and commercialization of those agents and NGS-based tests require new regulatory paradigms. Numerous studies are ongoing using recent designs such as enrichment and stratification on biomarkers, adaptative change in treatment allocation, or study of exceptional responders. The recent introduction of Breakthrough Therapy status in the USA has already supported the registration of four drugs in 2014. A first NGS platform was authorized by FDA for diagnostic use in 2013. Coming regulation of high risk LDT will insure higher quality and demonstration of clinical performance.

Recent data on the analysis of plasma DNA are very encouraging to detect and understand resistance to targeted therapies, and to provide an overall and iterative view of the cancer genome in the context of tumor heterogeneity and clonal evolution. As of today, cancer gene panels are likely to be the most useful tests for development.

References

1. Vogelstein B, Papadopoulos N, Velculescu V, et al. Cancer genome landscapes. Science. 2013;339:1546–58.
2. Fischer R, Pusztai L, Swanton C. Cancer heterogeneity: implications for targeted therapeutics. Br J Cancer. 2013;108:479–85.
3. Gibbs J. Regulating molecular diagnostic assays: developing a new regulatory structure for a new technology. Exp Rev Mol Diagn. 2011;11:367–81.
4. Gargis A, Kalman L, Berry M, et al. Assuring the quality of next generation sequencing in clinical laboratory practice. Nat Biotechnol. 2012;30:1033–6.
5. Frampton G, Fichtenholtz A, Otto G, et al. Development and validation of a clinical cancer genomic profiling test based on massively parallel DNA sequencing. Nat Biotechnol. 2013;31:1023–31.
6. Zhang L, Chen L, Sah S, et al. Profiling cancer gene mutations in clinical formalin-fixed paraffin-embedded colorectal tumor specimens using targeted next generation sequencing. Oncologist. 2014;19:336–43.
7. Schrijver I, Aziz N, Farkas D, et al. Opportunities and challenges associated with clinical diagnostic genome sequencing. J Mol Diagn. 2012;14:525–40.
8. Van Hallen E, Wagle N, Levy M. Clinical analysis and interpretation of cancer genome data. J Clin Oncol. 2013;31:1825–31.
9. Sleyfer S, Bogaerts J, Siu L. Designing transformative clinical trials in the cancer genome area. J Clin Oncol. 2013;31:1834–41.
10. Planchard D, Mazieres J, Riely G, et al. Interim results of phase II study of Dabrafenib in BRAF V600E mutation-positive non-small cell lung cancer. J Clin Oncol. 2013;31(Suppl):Abst. 8009.
11. Prahallad A, Sun C, Huang S, et al. Unresponsiveness of colon cancer to BRAF (V600E) inhibition through feedback activation of EGFR. Nature. 2012;483:100–3.
12. Ohashi K, Maruvka Y, Michor F, Pao W. Epidermal growth factor receptor tyrosine kinase inhibitor-resistant disease. J Clin Oncol. 2013;8:1070–80.
13. Roychowdhury S, Iyer M, Robinson D, et al. Personalized oncology through integrative high throughput sequencing. A pilot study. Sci Transl Med. 2011;3:111ra121.
14. Tran B, Brown A, Bedard P, et al. Feasability of real time next generation sequencing of cancer genes linked to drug response. Results from a clinical trial. Int J Cancer. 2012;132:1547–55.
15. Parkinson D, Mc Cormack R, Keating S. Evidence of clinical utility: an unmet need in molecular diagnostics for patients with cancer. Clin Cancer Res. 2014;20:1428–44.
16. Li T, Kung HJ, Mack P, Gandara D. Genotyping and genomic profiling of non-small-cell lung cancer: implications for current and future therapies. J Clin Oncol. 2013;31:1039–49.
17. Garraway L. Genomics-driven oncology. Framework for an emerging paradigm. J Clin Oncol. 2013;31:1806–14.
18. Ciriello G, Miller M, Aksoy B, et al. Emerging landscape of oncogenic signature across human cancers. Nat Genet. 2013;45:1127–33.
19. Lipson D, Capelletti M, Yelinsky R, et al. Identification of new ALK and RET gene fusions from colorectal and lung cancer biopsies. Nat Med. 2012;18:302–84.
20. Lolkema M, Gadellaa-Van Hooijdonk C, Bredenoord A, et al. Ethical, legal and counseling challenges surrounding the return of genetic analysis in oncology. J Clin Oncol. 2013;31: 1842–8.
21. Van Allen E, Wagle N, Stojanov P, et al. Whole exome sequencing and clinical interpretation of formalin-fixed, paraffin-embedded tumor samples to guide precision cancer medicine. Nat Med. 2014;20:682–8.
22. Kothari V, Wei I, Shankar S, et al. Outlier kinase expression by RNA sequencing as targets for precision therapy. Cancer Discov. 2013;3:280–93.
23. André F, Bachelot T, Commo F, et al. Comparative genomic hybridization array and DNA sequencing to direct treatment of metastatic breast cancer: a multicenter prospective trial (SAFIR01 UNICANCER). Lancet Oncol. 2014;15:267–74.

24. Tsimberidou A-M, Iskander N, Hong D, et al. Personalized medicine in a phase I clinical trial program: the MD Anderson Cancer Center initiative. Clin Cancer Res. 2012;18:6373–83.
25. Kaplan R, Maugham T, Crook A, et al. Evaluating many treatments and biomarkers in oncology: a new design. J Clin Oncol. 2013;31:4562–8.
26. Kim E, Herbst R, Wistubd I, et al. The BATTLE trial: personalizing therapy for lung cancer. Cancer Discov. 2011;1:44–53.
27. Welsh JS, Westervelt P, Ding L, et al. Use of whole genome sequencing to diagnose a cryptic fusion oncogene. JAMA. 2011;305:1577–84.
28. Iyer G, Hanrahan A, Milowsky M, et al. Genome sequencing identifies a basis for everolimus sensitivity. Science. 2012;388:221.
29. Kaiser J. Rare cancer successes spawn exceptional research efforts. Science. 2013;340:263.
30. Wagle N, Berger M, Davis M, et al. High throughput detection of actionable genomic alterations in clinical tumor samples by targeted, massively parallel sequencing. Cancer Discov. 2012;2:82–93.
31. Vignot S, Frampton G, Soria J-C, et al. Next generation sequencing reveals high concordance of recurrent somatic alterations between primary tumor and metastases from patients with non-small-cell lung cancer. J Clin Oncol. 2013;31:2167–72.
32. Forshew T, Murtoza M, Parkinson C, et al. Noninvasive identification and monitoring of cancer mutations by targeted deep sequencing of plasma DNA. Sci Transl Med. 2012;4:136ra68.
33. Dawson S-J, Tsui D, Murtoza M, et al. Analysis of circulating tumor DNA to monitor metastatic breast cancer. N Engl J Med. 2013;368:1199–209.
34. Diaz L, Bardelli A. Liquid biopsies: genotyping circulating tumor DNA. J Clin Oncol. 2014;32:579–86.
35. Aparicio S, Caldas C. The implications of clonal genome evolution for cancer medicine. N Engl J Med. 2013;368:842–51.
36. Murtoza M, Dawson S-J, Tsui D, et al. Non-invasive analysis of acquired resistance to cancer therapy by sequencing of plasma DNA. Nature. 2013;497:108–12.
37. Haber D, Velculescu V. Blood-based analysis of cancer: circulating tumor cells and circulating tumor DNA. Cancer Discov. 2014;9:650–1.
38. Conti R, Bernstein A, Villaflor M, et al. Prevalence of off-label use and spending in 2010 among patent protected chemotherapies in a population-based cohort of medical oncologists. J Clin Oncol. 2013;9:1134–9.
39. Emmerich J, Dumarcet N, Lorence A. France's new framework for regulating off-label drug use. N Engl J Med. 2012;367:1279–81.
40. Overman M, Modak J, Kopez S, et al. Use of research biopsies in clinical trials: are risks and benefits adequately disclosed? J Clin Oncol. 2013;31:17–22.
41. Mirale S, Yaeger R, Hobor S, et al. Emergence of KRAS mutations and acquired resistance to anti-EGFR therapy in colorectal cancer. Nature. 2012;486:532–6.
42. Meril-Bernstein F, Farhangfar C, Mendelsohn J, et al. Building a personalized medicine infrastructure at a major cancer center. J Clin Oncol. 2013;31:1849–57.
43. Nowak F, Soria J-C, Calvo F. Tumor molecular profiling for deciding therapy. The French initiative. Nat Rev Clin Oncol. 2012;9:479–86.
44. Tasik P, Zwinderman A, Mol B, Bossuy T. Trial designs for personalizing cancer care: a systematic review and classification. Clin Cancer Res. 2013;19:4578–88.
45. Rodon J, Saura C, Dienstmann R, et al. Molecular prescreening to select patient population in early clinical trials. Nat Rev Clin Oncol. 2012;9:359–66.
46. Berry D, Herbst R, Rubin E. Design strategies for personalized therapy trials. Clin Cancer Res. 2012;18:638–44.
47. Gianni L, Pienkowski T, Im Y-U, et al. Efficacy and safety of neoadjuvant pertuzumab and trastuzumab in women with locally advanced inflammatory, or early HER-2 positive breast cancer (Neosphere): a randomized muticentre, open-label phase 2 trial. Lancet Oncol. 2012;13:25–32.
48. Prowell T, Pazdur R. Pathological complete response and accelerated drug approval in early breast cancer. N Engl J Med. 2012;366:2438–41.

49. Freidlin B, Korn L. Biomarker enrichment strategies: matching trial design to biomarker credentials. Nat Rev Clin Oncol. 2014;11:81–90.
50. Abrams J, Conley B, Mooney B, et al. National Cancer Institute's precision medicine initiatives for the new national clinical trial network. ASCO educational book. Alexandria, VA: American Society of Clinical Oncology; 2014. p. 71–6.
51. Johnson D, Dahlman K, Knol J. Enabling a genetically informed approach to cancer medicine: a retrospective evaluation of the impact of comprehensive tumor profiling using a targeted next generation sequencing panel. Oncologist. 2014;19:611–22.
52. Schwaederle M, Parker B, Schwab R. Molecular tumor board: The University of California San Diego Moores Cancer center experience. Oncologist. 2014;19:631–6.
53. The Clinical Lung Cancer Genome Project and Network Genomic Medicine. A genomics-based classification of human lung tumors. Sci Transl Med. 2013;5:209ra153.
54. Tsimderiou A, Eggermont A, Schilsky R. Precision cancer medicine: the future is now, only better. ASCO educational book. Alexandria, VA: American Society of Clinical Oncology; 2014. p. 61–9.
55. Ocana A, Amir E, Vera-Badillo F, et al. Phase III trials of targeted anticancer therapies: redesigning the concept. Clin Cancer Res. 2013;19:4931–40.
56. Horning S, Haber D, Selig W. Developing standards for breakthrough therapy designation in oncology. Clin Cancer Res. 2013;19:4297–304.
57. US Food and Drug Administration (FDA). Paving the way for personalized medicine. US Food and Drug Administration (FDA). 2013. www.fda.gov.
58. FDA. Expanded access to investigational drugs for treatment use. FDA draft guidance. 2013. www.fda.org.
59. Evaluation of clinical validity and clinical utility of actionable molecular diagnostic tests in adult oncology. 2013. www.cmtpnet.org
60. Pignetti F, Ehmann F, Hemmings R, et al. Cancer drug development and the evolving regulatory framework for comparison diagnostics in the European Union. Clin Cancer Res. 2014;20:1458–68.
61. FDA. Framework for oversight of laboratory developed tests LDTs. FDA notification to congress only. www.fda.gov.
62. Collins F, Hamburg M. First FDA authorization for next generation sequencer. N Engl J Med. 2013;369:2369–71.
63. Salto-Tellez M, Gonzalez de Castro D. Next generation sequencing: a change of paradigm in molecular diagnostic validation. J Pathol. 2014;234:5–10.
64. Mansfield E. FDA perspective on companion diagnostics: an evolving paradigm. Clin Cancer Res. 2014;20:1453–7.
65. Aapro M, de Laurentis M, Mamounas E, et al. Adoption of multigene assays in HR+, HER2- breast cancer patients in Europe. Results of the multidisciplinary application of genomics in clinical practice survey. Ann Oncol. 2014;25 Suppl 1:i5–7.
66. Barlesi F, Blons H, Beau-Faller M, et al. Biomarkers-France results of routine EGFR, HER-2, KRAS, BRAF, PI3KCA mutation detection and EML – ALK gene fusion assessment on the first 10,000 non-small-cell-lung cancer patients. J Clin Oncol. 2013;31(Suppl):Abst. 8000.

Next-Generation Sequencing in the Era of Cancer-Targeted Therapies: Towards the Personalised Medicine

Ashwag Albukhari, Fawzi F. Bokhari, and Hani Choudhry

Abstract During the last two decades, several efforts have resulted in the fruitful development of cancer-targeted therapies, which have been approved by the FDA for the treatment of solid tumours as well as haematological malignancies. However, the rapid emergence of drug resistance remains a fundamental obstacle that limits the efficacy of targeted therapies. Understanding the molecular and biochemical mechanisms underlying the development of acquired drug resistance and the identification of predictive biomarkers of drug response and resistance would direct the selection of appropriate treatment regimens. Several experimental approaches have been utilised to elucidate the mechanisms of acquired drug resistance. In fact, the rapid evolution of the high-throughput genomics and proteomics technologies has provided novel insights into resistant mechanisms. This chapter provides the background of cancer-targeted therapies using the EGFR-targeted therapies as examples. It will further highlight the most common mechanisms of acquired drug resistance and will explain the experimental approaches to studying such mechanisms using the conventional biochemical and molecular techniques along with the high-throughput omics platforms. Although considerable challenges remain, the extraordinary insights into the biology of cancer therapies have led to the development of milestone combinatorial regimens towards more efficient personalised therapeutic options.

A. Albukhari (✉)
Biochemistry Department, Faculty of Science, King Abdulaziz University, Jeddah, Saudi Arabia

Department of Oncology, University of Oxford, Wellington Square, Oxford OX1 2JD, UK
e-mail: aalbukhari@kau.edu.sa; ashwag.albukhari@oncology.ox.ac.uk

F.F. Bokhari
Post Graduate Training and Research Centre, Medical Services General Directorate, Ministry of Defence, Armed Forces Hospitals, Taif Region, Saudi Arabia

H. Choudhry
Faculty of Science, Biochemistry Department, Center of Innovation in Personalized Medicine, King Fahd Center for Medical Research, King Abdulaziz University, Jeddah, Saudi Arabia
e-mail: hchoudhry@kau.edu.sa

W. Wu, H. Choudhry (eds.), *Next Generation Sequencing in Cancer Research, Volume 2*, DOI 10.1007/978-3-319-15811-2_3

1 Introduction

The bottleneck of cancer treatment remains a challenge particularly with patients presenting to clinical settings in late incurable cancer stages. The current treatment regimens to patients presenting with resectable cancer are mainly surgery in combination with pre- and/or post-operative chemotherapy. In some cases, however, patients may need to undergo triplet therapy where radiation therapy is also used in conjunction with the aforementioned treatment options. The unspecific targeting that occurs using these treatment regimens has led to the notion of developing effective targeted therapies. The modern technological revolution has helped to achieve such a goal. Indeed, next-generation sequencing, for example, makes it possible to assess mutation, DNA copy number, rearrangement, RNA editing, specific allele amplification, methylation or transcription with high-throughput robotic platforms. This ultimate technology eases the characterisation for both patients and tumour genome to aid in developing potential effective targeted therapy.

2 Next-Generation Sequencing Technologies

Identification of widespread genomic alterations including mutations, methylation aberrations, chromosomal rearrangements, structural changes and gene expression alteration, which drives tumorigenesis, has been possible with recent advancements in genomics technologies, in particular, next-generation sequencing (NGS). Screening of cancer genome has significantly improved our understanding on mechanisms that derive cancer initiation, progression, maintenance, resistance and clinical management [1–4]. The NGS technologies provide a comprehensive catalogue of genomic and transcriptome sequences within cancer cells. Moreover, genomic analyses can allow detection of interpatient and intratumoral heterogeneity facilitating treatment decisions for personalised cancer therapy hence improving clinical outcomes [5].

NGS technologies are also known as massive parallel sequencing because of their ability to sequence hundreds of millions of DNA fragments simultaneously in parallel. The NGS technologies are commonly divided into two platforms: conventional NGS platforms and desktop NGS platforms. The conventional NGS platforms are used for large-scale sequencing studies, such as whole-genome sequencing (WGS), whole-transcriptome sequencing and whole-exome sequencing (WES), whereas the desktop NGS platforms are used for low-complexity and targeted gene sequencing [6]. The conventional NGS technologies include Roche 454 GS FLX, Illumina GA/HiSeq and Life Technologies SOLiD/5500 sequencing instruments. On the other hand, Roche 454 GS Junior, Ion Proton, Ion Torrent and Illumina MiSeq are examples of the desktop NGS platforms. The NGS technologies utilise different chemistries for sequencing, for instance, Illumina (GA/HiSeq/MiSeq) uses the reversible terminator chemistry; however, Life Technologies (SOLiD excluding Ion Torrent) utilise the DNA ligase enzyme to perform sequencing [6].

The increasing need of small panel sequencing of genes in clinical setting to identify patients with specific mutation has led to the development of many benchtop NGS technologies. Several studies have evaluated the use of NGS technologies as diagnostic tools for Mendelian diseases and cancer [7, 8]. The sequencing throughput of benchtop NGS technologies is commonly ranging from 10 Mb to 1 Gb and reads lengths ranging from 100 to 200 bp which are suitable for clinical applications [6, 9]. The benchtop NGS platforms can be exploited for clinical screening to identify patients for targeted therapies, monitor responses and disease management. Owing to their throughput and quick turnover time, benchtop NGS is becoming a powerful tool for targeted gene sequencing, in clinics.

3 Targeted Therapies for Cancer

Despite the inconceivable passion prompted by the potential of personalised cancer therapy, several challenges need to be solved along the way before this type of regimen can benefit at least the majority of cancer patients. The implementation of the targeted cancer treatment requires a link that connects the characterisation of the patient's genome and changes occurring in their tumour. This connecting link will help in identifying biomarkers that reflect which patients will respond to personalised therapeutics. Throughout the past two decades, promising advances have been achieved with the emergence of molecularly targeted agents to treatment paradigms. These include, for example, therapeutic antibodies, small molecule inhibitors and si/shRNA. In fact, some of these therapeutic tools have advanced in clinical trials, particularly antibodies and kinase inhibitors.

As an established class of pharmaceuticals, monoclonal antibodies (mAbs) have been well tolerated to treat a variety of diseases, although their large size (approximately 150 kDa) and the unavailability of a proper delivery system sometimes obscure their effectiveness. Also, the immunologic impact that occurs due to the administration of non-human Ab limits their potential use for treatment purposes. Obviously, this can be noted when mouse Abs are used where human develops human anti-mouse antibodies (HAMA). DeNardo and colleagues reported that 50–75 % of patients treated with mouse Abs develop HAMA [10]. The development of HAMA led to the notion of reducing the contents of murine mAbs [11]. Enormous efforts have been made to achieve modification, including the use of chimeric mAbs with human preserved region, human mAbs preserving only mouse complementarity-determining regions (CDRs), and a trial of fully human mAbs. The progress of modifying these mAbs resulted in reducing the use of murine mAbs to nil.

From immunoediting of cancer, immunotherapy has progressed notably with special emphasis on monoclonal antibodies. It is now very well evident that cancerous cells express distinguishable extracellular surface markers compared to normal cells. Obviously, these surface markers can potentially be used as a target for specifically designed mAbs. The FDA has approved several mAb products where some of them are used specifically as cancer therapeutics. Indeed, the approval of

the anti-Her2-targeted mAb, trastuzumab, for the treatment of Her2-overexpressing breast cancer by the FDA in the late 1998 was a crucial milestone in the era of cancer-targeted therapies [12].

Since then, molecular investigations have revolutionised our understanding of several cellular mechanisms occurring in cancer. One obvious key player that orchestrates cellular transduction pathways in normal and cancerous cells is the epidermal growth factor receptor (EGFR). The overexpression, amplification or mutation of EGFR was found in several human malignancies including head and neck, breast, lung, colorectal, prostate, pancreas, ovary, brain and bladder carcinomas [13, 14]. Triggering EGFR can initiate several mechanistic pathways that may contribute to the cancerous transformation such as RAS/MAPK and PI3K/AKT pathway [15]. EGFR has also been reported to act as a transcription factor when it is translocated to the nucleus [16, 17].

Because of its crucial role, EGFR has been studied extensively, and several compounds have been used to hinder its activity in cancer cells, including mAbs and small molecule tyrosine kinase inhibitors. The specifically designed mAbs of EGFR bind to its extracellular domain while in an inactive state and, therefore, inhibit ligand-induced tyrosine activation [18–20]. The EGFR small molecule inhibitors, on the other hand, compete reversibly with the adenosine 5′ triphosphate to bind to the intracellular catalytic domain of the EGFR and, consequently, inhibit EGFR auto-phosphorylation and block further downstream signalling. However, due to the lack of high specificity in these small molecule inhibitors and their possible interactions with other targets, mAbs were decidedly favourable. There are several mAbs that antagonise EGFR function. However, two mAb compounds, cetuximab and panitumumab, are widely used as cancer therapies.

3.1 Cetuximab (Anti-EGFR mAbs)

Also known as C225, Erbitux™ is an immunoglobulin G1 (IgG1) human-murine mAb. Erbitux mAbs bind to the EGFR ligand with about 2-log stronger affinity compared to the natural ligands TGF-α and EGF [21–23]. Cetuximab binds to EGFR and promotes receptor degradation without phosphorylation and activation [24]. As a result, the availability of EGFR receptors will be reduced on the surface and subsequently prevents EGFR-associated downstream signalling pathways. It has also been reported that cetuximab binds to the mutant receptor EGFRVIII causing 80 % reductions in its phosphorylation state [25]. Due to downregulation of these essential pathways, cetuximab arrests the cell cycle at G0/G1 and increases the expression of the cell cycle regulator p27KIPI. This resulted in the induction of apoptosis through initiation of pro-apoptotic proteins such as Bax and/or caspase-3 or by suppressing the functions of anti-apoptotic proteins such as Bcl-2 [26, 27]. Another remarkable effect of using cetuximab is the inhibition of pro-angiogenic factor production such as endothelial growth factor and interleukin-8. Inhibiting these factors decreases angiogenesis and the development of distal metastases in orthotopic cancer models [28].

The first cetuximab clinical trials phase I and II illustrated the safety of the compound alone or in combination with chemotherapeutic agents in several malignancies including head and neck, colorectal and non-small cell lung carcinoma [29, 30]. In a randomised phase II clinical trial, cetuximab was tested alone and in combination with irinotecan in patients with refractory metastatic colon cancer. The response rate of combining cetuximab with irinotecan was significantly higher (23 % where $n=218$) compared to cetuximab alone (11 % where $n=111$). The disease control was also higher in the combined group versus cetuximab alone, 56 and 32 %, respectively. These observations were confirmed by another report, which investigated the efficacy and safety of cetuximab as therapeutic agent in metastatic colorectal cancer [31]. Based on these phenomenal observations, cetuximab was approved for the treatment of patients with metastatic colorectal cancer expressing EGFR refractory to irinotecan-based chemotherapy.

Moreover, EGFR was found to be overexpressed in head and neck cancers. A phase III randomised clinical trial using cetuximab was performed on 424 patients. Patients were categorised into two groups, one group receiving radiation therapy alone and the other group treated with radiotherapy supplemented with cetuximab. A significant survival difference was obviously noticed at a median follow-up of 54 months (49 vs. 29 months). This report was the first of its kind to show a significant difference between using the EGFR mAbs compared to current routinely used radiation therapy [32].

3.2 Gefitinib (EGFR Small Molecule Tyrosine Kinase Inhibitor, TKI)

Non-small cell lung carcinoma (NSCLC) counts for 80 % of lung cancer cases and remains the major cause of cancer-related death worldwide. The current treatment regimens for this neoplastic disorder remain to be surgery, chemotherapy and/or radiotherapy—or a combination of these therapeutic tools depending on the stage of the disease and age of the patient. The efficacy of chemotherapy shows the same treatment patterns in a wide range of ages [33]. However, those who are older than 75 years may present with more toxic profiles. Therefore, a mono-chemotherapy is preferred with vinorelbine, gemcitabine or docetaxel instead of platinum doublets, although a phase III clinical trial of using carboplatin (monthly) and paclitaxel (weekly) may be superior to gemcitabine or vinorelbine in this elderly age group [34].

Gefitinib is an orally administered small molecule inhibitor that targets EGFR, which has been used in treating NSCLC. In 2004, it was demonstrated that somatic mutation of EGFR correlated positively with the responsiveness rate of small molecule inhibitor targeting EGFR in treating patients with NSCLC [35, 36]. In Japan, two randomised phase III clinical trials compared the use of gefitinib with first-line chemotherapy in patients presented with NSCLC. In the West Japan Oncology Group, gefitinib-treated patients reported a median progression-free survival (PFS) of 9.2 months compared to patients treated with chemotherapy (6.3 months) [37]. Another group in north-east Japan demonstrated similar results

where patients treated with gefitinib had PFS of 10.8 months compared to 5.4 months for those treated with chemotherapy [38]. Small molecule inhibitors that target EGFR are showing highly promising results and show the potential of using targeted therapy for cancer patients.

4 The Development of Acquired Cancer Drug Resistance

The development of cancer-targeted therapies has resulted in a remarkable relief to patients and introduced new treatment regimens to different malignancies. However, despite the extensive efforts to develop targeted therapies, the efficacy of cancer chemotherapies and targeted therapies is often limited by the rapid emergence of acquired resistance and patient relapse after initial response. Several mechanisms of acquired resistance have been identified in different tumours as a result of different factors including individual factors as well as somatic variations between different tumours. One of the common mechanisms involves alterations in the oncogenic pathway that enables the cancerous cells to remain addictive to the original oncogene and evade the inhibition of the drug target. Moreover, acquisition of drug resistance could be due to bypass mechanisms that activate parallel oncogenic pathway(s), which provide an alternative survival mechanism to the cancerous cells. However, acquired resistance could emerge due to pathway-independent routes as well as the factors arising from the tumour microenvironment. Genetic alterations such as the development of secondary mutation in the drug target and/or mutations of downstream effectors, loss of the cell surface receptor, translocation of the receptor as well as epigenetic modification have also been involved in the emergence of acquired resistance to cancer therapies. Furthermore, cancer cells could develop a phenomenon known as a multidrug resistance in which cells acquire resistance to different structurally and functionally targeted therapies. This could be due to several mechanisms such as limiting drug uptake, enhancing drug efflux or altering cell membrane lipid components and therefore limiting drug accumulation within the cells. Understanding the underlying molecular and biochemical mechanisms behind the development of the acquired drug resistance will aid in designing novel strategies to prevent the development of drug resistance, overcome resistance and improve the therapeutic outcomes.

5 Experimental Approaches to Investigating the Underlying Mechanisms of Cancer Drug Resistance

Extensive efforts have been made during the last decade to investigate the mechanisms of acquired resistance to different targeted therapies including the mAbs and TKIs in different cancers. Several approaches, including the use of both in vitro and in vivo preclinical models as well as profiling of patient tumour samples, have been utilised to characterise the mechanisms of their resistance.

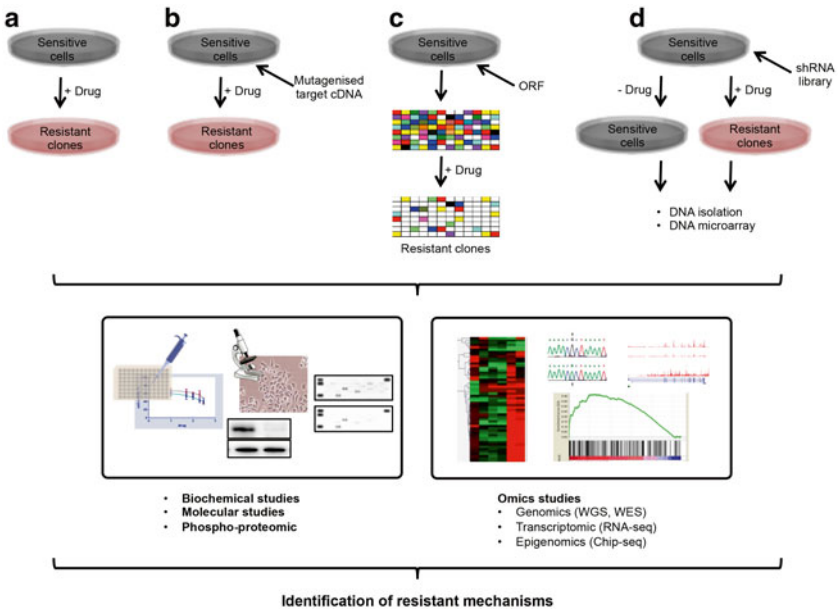


Fig. 1 In vitro development and studies of acquired drug resistance. Development of acquired resistant cell lines by (**a**) continuous exposure of parental sensitive cells to the drug, (**b**) random mutation of the target protein followed by drug selection, (**c**) systematic gain of function screen using open reading frames (ORF) and (**d**) systematic loss of function screen using shRNA libraries. This is followed by functional studies using different biochemical, molecular, proteomics, genomics, transcriptomics and epigenomics approaches to identify the mechanism(s) of acquired drug resistance

5.1 In Vitro Study of the Mechanisms of Acquired Drug Resistance in Tumour Cell Lines

Different methods have been used for the in vitro development of the acquired drug resistance (Fig. 1). One of the most common in vitro approaches involves the development of acquired drug-resistant clones derived from their parental cell lines by the continuous exposure of parental drug-sensitive cells to increasing concentrations of the drug until they become drug refractory and lose their sensitivity. This is followed by the comparison of the parental cells and the emerged drug-resistant subpopulations to identify the genetic, epigenetic, molecular or biochemical alterations that might contribute to their resistance. Mutations or amplification of the drug target is commonly involved in acquired drug resistance. For example, DNA sequencing of the EGFR coding region in both parental DiFi colorectal cancer cell line and the derived cetuximab-resistant clones revealed that a missense mutation (S492R) in the extracellular domain of EGFR prevented the binding of cetuximab to EGFR and conferred resistance to cetuximab [39].

Biochemical studies of the drug target protein or pathway have also been used to identify the mechanisms of acquired drug resistance. In a model of acquired cetuximab-resistant NSCLC, Wheeler et al. [40] have found an increase in steady-state EGFR expression in the acquired resistant NSCLC clones compared to their parental cells that was associated with the deregulation of EGFR internalisation and degradation. Furthermore, investigating the crosstalk between the target protein and other members of the same family of proteins or other related proteins and/or pathways has been used as another approach to identify the mechanisms of acquired resistance in vitro. The identification of such crosstalk could be focused on a particular protein/pathway such as the study by Bianco et al. [41], which revealed that an overexpression of the VEGFR1 receptor in the acquired resistant cell lines to different EGFR inhibitors and the inhibition of VEGFR1 using either the multi-targeted inhibitor vandetanib or by knockdown VEGFR-1 in the resistant cells restored their sensitivity to EGFR inhibitors. A model of acquired trastuzumab-resistant cell lines also revealed the involvement of other receptor tyrosine kinases that interact directly with Her2, trastuzumab target, and induced their resistant trastuzumab treatment [42]. One of the most common methods is the phospho-proteomic profiling of the human receptor tyrosine kinases and their downstream effectors using the commercially available kits such as the human phospho-RTK array kit from R&D Systems. The kit is designed to monitor the relative phosphorylation status of 49 different RTKs and therefore is a useful screening method to detect the activation of other signalling pathways in the resistant clones compared to their parental sensitive cells. Using this system, Wheeler et al. [40] have found an activation of different receptors including Her2, Her3 and cMet in acquired cetuximab-resistant NSCLC and HNSCC cell lines compared to their parental cells.

Another approach is to generate resistant clones by random in vitro mutagenesis screen of the drug target and identification of the mutant variants that induce drug resistance such as the identified *MEK1* mutations that confer resistance to MEK and BRAF inhibitions in BRAF-mutant melanoma cells and tumours obtained from relapsed melanoma patients following treatment with MEK inhibitor [43].

Although such approaches have led to the identification of several underlying mechanisms of acquired resistance to cancer-targeted therapies, such focused studies could be biased, and their interpretation might be straightforward. The use of high-throughput profiling unbiased omics approaches (genomic, transcriptomic, epigenomic and proteomic) could extend such studies to identify novel candidate genes and pathways that are associated with the acquired drug resistance. For example, in a model of acquired tamoxifen-resistant breast cancer cell line, Huber-Keener et al. [44] have compared the expression profiling of the resistant cells to their parental cells using next-generation RNA-seq. This allowed the identification of the deregulated transcripts and their biological functions [44]. Similarly, the transcriptome profiling of the acquired gefitinib-resistant NSCLC cell lines using RNA-seq revealed the involvement of FGF and FGFR1 as a novel pathway of acquired gefitinib resistance [45]. Furthermore, Engelman et al. [46] combined both genome-wide copy number analysis and expression profiling of the acquired gefitinib-resistant NSCLC cell lines. Their study has identified an amplification of the *MET* oncogene, which mediates

Her3 activation and subsequent development of gefitinib resistance. Additionally, analysis of NSCLC patient tumours revealed the acquisition of *MET* amplification in tumours that acquired resistance while receiving gefitinib treatment [46].

Functional genetic screens using systematic large-scale gain or loss of functions provide powerful unbiased tools to identify novel mechanisms of drug resistance in preclinical models. Loss of function using the genome-wide RNA interference (RNAi) screen approach has also been used to identify candidate genes whose suppression could induce drug resistance. This approach involves generation of a genome-wide shRNA library targeting hundreds of genes, transfecting the drug-sensitive cells with the shRNA library followed by drug treatment, isolation of genomic DNA, recovery labelling and hybridisation to DNA microarrays. The use of this technique has led to the identification of the role of PTEN tumour suppressor gene loss in the development of trastuzumab resistance in the Her2-positive breast cancer cell line [47]. Similarly, using the RNAi screen identified that loss of *CDK10* induces activation of MAPK pathway through ETS2-driven transcription of *c-RAF* and loss of oestrogen signalling dependency in breast cancer cells and therefore resistance to tamoxifen treatment [48]. Recently, Huang et al. [49] have identified MED12 as a determinant of response to different cancer drugs including chemotherapies as well as EGFR and ALK inhibitors using the RNAi screening approach. Loss of MED12 induces activation of TGF-beta signalling that mediates resistance to these different therapies through the activation of MEK/ERK pathway.

On the other hand, the gain of function by the expression of open reading frames (ORF) approach has also been used to identify gene(s) whose overexpression confers resistance to cancer-targeted therapies. Genome-wide expression of about 600 kinases and kinase-related ORFs in a *BRAFV600E* melanoma cell line that is sensitive to the RAF kinase inhibitor PLX4720 revealed that COT is a novel kinase that induces resistance to RAF inhibition [50]. Collectively, such systematic functional genetic approaches provide promising insights into the identification of novel genes and/or pathways that are involved in the emergence of acquired resistance to different targeted therapies that will have clinical relevance and lead to the identification of alternative therapeutic strategies to overcome drug resistance.

5.2 *In Vivo Study of the Mechanisms of Acquired Drug Resistance in Animal Models*

Investigating the molecular mechanisms underlying the development of acquired drug resistance to different cancer-targeted therapies using cancer cell lines has resulted in the identification of several resistant mechanisms that have been validated in patient tumours. However, the non-cell-autonomous factors from the tumour microenvironment have been recently recognised as contributors to the development of drug resistance and cancer progression [51]. Therefore, using in vivo models has the advantage of identifying the non-cell-autonomous determinants of acquired drug resistance (Fig. 2). Development of in vivo acquired drug-resistant cells can be

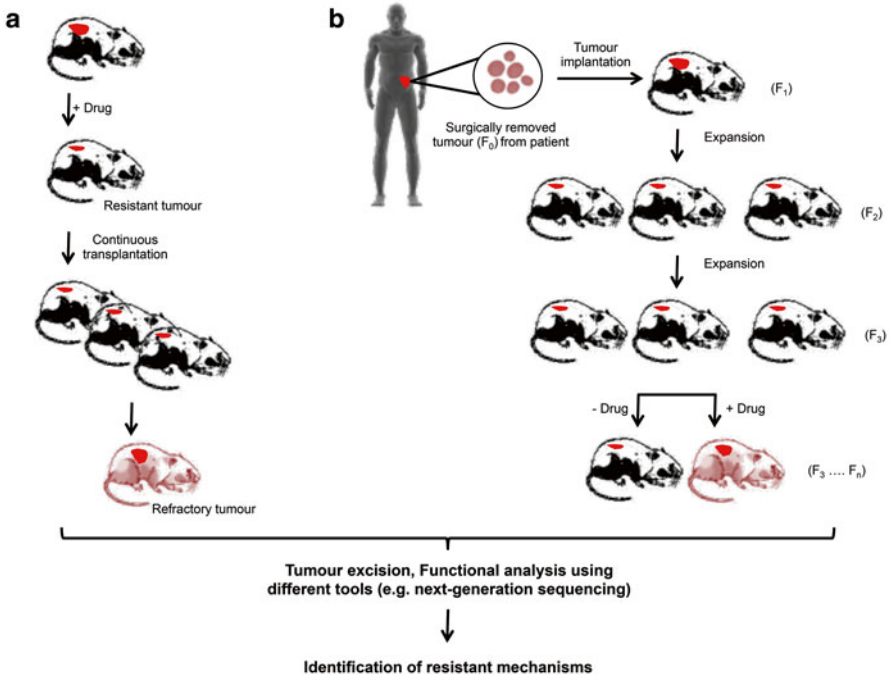


Fig. 2 In vivo development and studies of acquired drug resistance. (**a**) Tumour-bearing mice are treated with drug followed by continuous transplantation into new mice until refractory tumours are generated. (**b**) Establishment of patient-derived tumour xenograft (PDTX) models in immune-compromised mice followed by drug treatment until the development of refractory tumours. Resistant tumours are excised and used for functional analyses as described previously in Fig. 1 to identify the acquired drug-resistant mechanisms

induced by treatment of tumour-bearing mice with a drug until they show a minimum response. This is followed by continuous transplantation of the tumour cells into new sets of mice and treating them with the drug until one tumour becomes refractory and this can be used as a model of drug resistance [52]. The derived in vivo acquired resistant cells can then be used, employing the previously described methods to identify the mechanisms of their resistance. Moreover, genetically engineered mice have also been used to validate the findings obtained from the in vitro studies. For example, the transgenic model that conditionally expresses *PIK3CAH1047R* variant mammary tumour has been used to study the mechanism of resistance to PI3K inhibitors in tumours harbouring this mutational activation on PI3K. Genomic and molecular analyses of the recurrent tumours revealed an amplification of *MET* and *MYC* that maintain tumour survival via PI3K pathway-dependent and PI3K pathway-independent mechanisms, respectively [53]. Using a genetically engineered mouse model of EGFR-mutant NSCLC, Politi et al. [54] have detected both *EGFRT790M* mutation and *MET* amplification in erlotinib-resistant tumours after multiple cycles of erlotinib treatment similar to the

observations that have been found in erlotinib-resistant lung cancer patients [46, 54, 55]. Overall, these in vivo models provided a powerful tool to investigate the mechanisms of acquired drug resistance that have clinical relevance.

Patient-derived tumour xenografts (PDTXs) have recently emerged as preclinical models for cancer drug discovery. PDTX can be obtained by collecting fresh patient tumour tissue, sectioning it into pieces (about 5 mm^3) followed by implantation into immune-compromised rodents such as athymic nude or NOD/SCID mice [56]. The first generation harbouring the patient-derived tumours is called F_0 with subsequent generations called (F_1, F_2 … F_n). Several tumour-specific PDTXs that are biologically stable have been established and used as preclinical models with gene expression pattern, mutational status, metastatic potential and drug responsiveness characteristics similar to the original tumours and thus provide a more reliable model over the standard cell line xenograft models [57]. These models offer a promising tool in the future for studying cancer drug resistance and the validation of in vitro data.

5.3 Studies of Cancer Drug Resistance in Patient Tumours

As mentioned above, both in vitro and in vivo models have provided tools that uncovered several mechanisms behind the development of acquired resistance to different cancer-targeted therapies. However, the main goal of such studies is to confirm such preclinical findings in patient tumours. One of the most common translations of the preclinical studies is to confirm the involvement of a candidate gene, protein or pathway in the clinical tumours from patients that were treated with a particular therapy using different techniques such as western blotting, PCR, FISH analysis and immunohistochemistry to study the changes in gene and protein expressions. Therefore, it is highly recommended to obtain both pre- and posttreatment tumours, which will allow the determination of changes over the time of treatment. However, several studies have been confirmed in archival specimens (paraffin-embedded tissues, fresh frozen tissues or tissue microarrays) that were collected posttreatment only. In this case, the expression pattern could be correlated with the clinical outcomes and clinicopathological features such as tumour size and grade. As mentioned above, the in vitro study of acquired cetuximab-resistant colorectal cancer cell lines has led to the identification of *EGFR* mutation in the resistant cells but not the parental cells. Deep sequencing of DNA obtained pre- and post-cetuximab-treated colorectal cancer patient tumours and revealed that progressed patients acquired this mutation after treatment [39]. Similarly, preclinical studies of the trastuzumab-resistant Her2 breast cancer cells revealed an increase in ADAM10 expression. The ADAM10 expression was associated with decreased clinical response of the Her2-positive breast cancer patients treated with trastuzumab monotherapy and also correlated with poorer relapse-free survival in a cohort of Her2-positive breast cancer patients [58].

In parallel, the advances in the omics and high-throughput technologies including genomic and transcriptomic analyses coupled with high-throughput analyses of

gene functions might accelerate the identification of the predictive biomarkers of response to a particular therapy as well as the identification of novel resistant mechanisms in patient tumours that would guide the oncology treatment decisions. Therefore, collecting tumour samples pre- and post-relapse after treatment with a particular cancer therapeutic treatment has become an interesting and growing area of cancer research. Several prospective clinical trials are being designed to include the collection of multiple tumour specimens during the treatment as well as blood samples. Pre- and posttreated tumour lesions from a melanoma patient, who received RAF inhibitor and developed drug resistance, were subjected to targeted massively parallel sequencing of 138 known cancer genes [59]. Profiling both pre- and post-lesions revealed an activation mutation in *MEK1C121S*, a RAF downstream kinase, in the posttreatment tumour but not in the pretreatment one. Further in vitro validations confirmed that *MEK1C121S* was responsible for the increase in the kinase activity and development of resistance to RAF inhibition. In the phase II clinical trial (TBCRC001), triple-negative breast cancer patients have been randomised to receive cetuximab +/− carboplatin. Gene expression profiling of pre- and post-treated tumours revealed an activation of the EGFR pathway in the majority of the patients who received the combined treatment, and only a minority showed pathway inhibition [60]. Recently, comprehensive molecular profiling of the residual tumours from 74 triple-negative breast cancer patients after neoadjuvant chemotherapy using RNA-seq and digital RNA expression has identified alteration of several targetable genes/pathways in the chemotherapy-resistant lesions. This could provide biomarkers that guide the selection of adjuvant treatments in order to improve the response of triple-negative breast cancer patients to chemotherapy and prevent metastases [61]. Overall, these studies provided genome-wide changes that were associated with tumour relapse and illustrated the utilisation of emerging technologies for assessing the mechanisms of acquired drug resistance.

6 Combinatorial Therapies to Overcome Drug Resistance

The development of acquired cancer drug resistance arises due to several underlying mechanisms. Comprehensive knowledge of the tumour biology, drug pharmacodynamics and pharmacokinetics and the underlying mechanisms of drug escape mechanisms provide rationales for combinatorial treatments to improve the initial response, prolong the duration of response to a particular therapy as well as to provide alternative therapeutic regimens after the relapse of the initial regimen.

Several combinations of targeted therapies or targeted therapies combined with chemotherapies have been approved or are under investigation in phase I, II and III clinical trials. Early elegant genomic studies of imatinib, ABL TKI, in chronic myeloid leukaemia patients identified that a point mutation in the ABL kinase domain is associated with the acquisition of imatinib resistance by inducing a reactivation of BCR-ABL signal transduction [62]. This led to the development of dasatinib, ABL kinase inhibitor, [63] which induced tumour responses in imatinib-resistant

chronic myeloid leukaemia patients in phase I clinical trial [64]. Similarly, studies of the mechanisms of resistance to Her2-targeted therapies in preclinical models and patient tumours have led to the approval of triple combination (pertuzumab + trastuzumab + docetaxel) (NCT00567190) for the treatment of Her2-positive breast cancer patients. Although such approaches have resulted in fruitful drug combinations, a complementary approach of systematic high-throughput unbiased screening strategies would be highly effective in identifying effective drug combinations. This could be achieved by high-throughput screening of drug(s) combinations in large panels of cancer cell lines to identify novel regimens of drug combination that induce synergistic interaction. Moreover, understanding the mechanistic insights of such combinations would help in patient selection using predictive biomarkers.

Computational modelling of complex biochemical pathways and the molecular mechanisms of drug action by utilising the massive data inputs from the next-generation technologies are being used to explain drug resistance and predict potential drug combinations to overcome drug resistance [65–68]. Integration of a large-scale pan-omics approach (genomic, transcriptomic, epigenomic, proteomic, etc.) to analyse tumour specimens from clinical trials coupled with systematic functional studies and computational modelling of gene-gene, protein-protein and genome-environment interactions and the identification of predictive biomarkers of drug response will provide more insights into drug-resistant mechanisms and speed the advent of personalised cancer treatment.

7 Conclusion

Advances in the high-throughput NGS technologies and functional genomics have accelerated our understanding of cancer biology. Furthermore, utilising these techniques has uncovered several novel mechanisms behind the development of acquired resistance to different cancer therapies. However, there are still unexplained observations that need to be investigated in order to improve the efficacy of cancer therapies and enhance the clinical outcomes. Achieving such goals requires collaborations between academia and the pharmaceutical and biotechnological companies as well as funding bodies to develop rationale studies based on strong preclinical data to identify the right treatment for every patient (tailored therapy) at the time of diagnosis, prolong the response and ultimately cure the patient.

References

1. Cancer Genome Atlas N. Comprehensive molecular portraits of human breast tumours. Nature. 2012;490(7418):61–70.
2. Alexandrov LB, Nik-Zainal S, Wedge DC, Aparicio SA, Behjati S, Biankin AV, Bignell GR, Bolli N, Borg A, Borresen-Dale AL, et al. Signatures of mutational processes in human cancer. Nature. 2013;500(7463):415–21.

3. Yu M, Ting DT, Stott SL, Wittner BS, Ozsolak F, Paul S, Ciciliano JC, Smas ME, Winokur D, Gilman AJ, et al. RNA sequencing of pancreatic circulating tumour cells implicates WNT signalling in metastasis. Nature. 2012;487(7408):510–3.
4. Ellis MJ, Ding L, Shen D, Luo J, Suman VJ, Wallis JW, Van Tine BA, Hoog J, Goiffon RJ, Goldstein TC, et al. Whole-genome analysis informs breast cancer response to aromatase inhibition. Nature. 2012;486(7403):353–60.
5. Boehm JS, Hahn WC. Towards systematic functional characterization of cancer genomes. Nat Rev Genet. 2011;12(7):487–98.
6. Mardis ER. Next-generation sequencing platforms. Annu Rev Anal Chem. 2013;6:287–303.
7. Koboldt DC, Steinberg KM, Larson DE, Wilson RK, Mardis ER. The next-generation sequencing revolution and its impact on genomics. Cell. 2013;155(1):27–38.
8. Ku CS, Cooper DN, Roukos DH. Clinical relevance of cancer genome sequencing. World J Gastroenterol. 2013;19(13):2011–8.
9. Metzker ML. Sequencing technologies – the next generation. Nat Rev Genet. 2010;11(1):31–46.
10. DeNardo GL, Bradt BM, Mirick GR, DeNardo S. Human antiglobulin response to foreign antibodies: therapeutic benefit? Cancer Immunol Immunother. 2003;52(5):309–16.
11. Hwang WY, Foote J. Immunogenicity of engineered antibodies. Methods. 2005;36(1):3–10.
12. Slamon DJ, Leyland-Jones B, Shak S, Fuchs H, Paton V, Bajamonde A, Fleming T, Eiermann W, Wolter J, Pegram M, et al. Use of chemotherapy plus a monoclonal antibody against HER2 for metastatic breast cancer that overexpresses HER2. N Engl J Med. 2001;344(11):783–92.
13. Woodburn JR. The epidermal growth factor receptor and its inhibition in cancer therapy. Pharmacol Ther. 1999;82(2–3):241–50.
14. Yarden Y. The EGFR family and its ligands in human cancer. signalling mechanisms and therapeutic opportunities. Eur J Cancer. 2001;37 Suppl 4:S3–8.
15. Yarden Y, Sliwkowski MX. Untangling the ErbB signalling network. Nat Rev Mol Cell Biol. 2001;2(2):127–37.
16. Lin SY, Makino K, Xia W, Matin A, Wen Y, Kwong KY, Bourguignon L, Hung MC. Nuclear localization of EGF receptor and its potential new role as a transcription factor. Nat Cell Biol. 2001;3(9):802–8.
17. Oksvold M, Huitfeldt H, Stang E, Madshus I. Localizing the EGF receptor. Nat Cell Biol. 2002;4(2):22.
18. Ciardiello F, Tortora G. EGFR antagonists in cancer treatment. N Engl J Med. 2008;358(11): 1160–74.
19. Normanno N, Bianco C, De Luca A, Maiello MR, Salomon DS. Target-based agents against ErbB receptors and their ligands: a novel approach to cancer treatment. Endocr Relat Cancer. 2003;10(1):1–21.
20. Normanno N, Maiello MR, De Luca A. Epidermal growth factor receptor tyrosine kinase inhibitors (EGFR-TKIs): simple drugs with a complex mechanism of action? J Cell Physiol. 2003;194(1):13–9.
21. Kim ES, Khuri FR, Herbst RS. Epidermal growth factor receptor biology (IMC-C225). Curr Opin Oncol. 2001;13(6):506–13.
22. Price TJ, Peeters M, Kim TW, Li J, Cascinu S, Ruff P, Suresh AS, Thomas A, Tjulandin S, Zhang K, et al. Panitumumab versus cetuximab in patients with chemotherapy-refractory wild-type KRAS exon 2 metastatic colorectal cancer (ASPECCT): a randomised, multicentre, open-label, non-inferiority phase 3 study. Lancet Oncol. 2014;15(6):569–79.
23. Blick SK, Scott LJ. Cetuximab: a review of its use in squamous cell carcinoma of the head and neck and metastatic colorectal cancer. Drugs. 2007;67(17):2585–607.
24. Steiner P, Joynes C, Bassi R, Wang S, Tonra JR, Hadari YR, Hicklin DJ. Tumor growth inhibition with cetuximab and chemotherapy in non-small cell lung cancer xenografts expressing wild-type and mutated epidermal growth factor receptor. Clin Cancer Res. 2007;13(5): 1540–51.
25. Jutten B, Dubois L, Li Y, Aerts H, Wouters BG, Lambin P, Theys J, Lammering G. Binding of cetuximab to the EGFRvIII deletion mutant and its biological consequences in malignant glioma cells. Radiother Oncol. 2009;92(3):393–8.

26. Mayo LD, Donner DB. A phosphatidylinositol 3-kinase/Akt pathway promotes translocation of Mdm2 from the cytoplasm to the nucleus. Proc Natl Acad Sci U S A. 2001;98(20):11598–603.
27. Gingras AC, Kennedy SG, O'Leary MA, Sonenberg N, Hay N. 4E-BP1, a repressor of mRNA translation, is phosphorylated and inactivated by the Akt(PKB) signaling pathway. Genes Dev. 1998;12(4):502–13.
28. Perrotte P, Matsumoto T, Inoue K, Kuniyasu H, Eve BY, Hicklin DJ, Radinsky R, Dinney CP. Anti-epidermal growth factor receptor antibody C225 inhibits angiogenesis in human transitional cell carcinoma growing orthotopically in nude mice. Clin Cancer Res. 1999;5(2):257–65.
29. Thienelt CD, Bunn Jr PA, Hanna N, Rosenberg A, Needle MN, Long ME, Gustafson DL, Kelly K. Multicenter phase I/II study of cetuximab with paclitaxel and carboplatin in untreated patients with stage IV non-small-cell lung cancer. J Clin Oncol. 2005;23(34):8786–93.
30. Saltz LB, Meropol NJ, Loehrer Sr PJ, Needle MN, Kopit J, Mayer RJ. Phase II trial of cetuximab in patients with refractory colorectal cancer that expresses the epidermal growth factor receptor. J Clin Oncol. 2004;22(7):1201–8.
31. Sobrero AF, Maurel J, Fehrenbacher L, Scheithauer W, Abubakr YA, Lutz MP, Vega-Villegas ME, Eng C, Steinhauer EU, Prausova J, et al. EPIC: phase III trial of cetuximab plus irinotecan after fluoropyrimidine and oxaliplatin failure in patients with metastatic colorectal cancer. J Clin Oncol. 2008;26(14):2311–9.
32. Bonner JA, Harari PM, Giralt J, Azarnia N, Shin DM, Cohen RB, Jones CU, Sur R, Raben D, Jassem J, et al. Radiotherapy plus cetuximab for squamous-cell carcinoma of the head and neck. N Engl J Med. 2006;354(6):567–78.
33. Azzoli CG, Temin S, Aliff T, Baker Jr S, Brahmer J, Johnson DH, Laskin JL, Masters G, Milton D, Nordquist L, et al. 2011 focused update of 2009 American Society of Clinical Oncology Clinical Practice guideline update on chemotherapy for stage IV non-small-cell lung cancer. J Clin Oncol. 2011;29(28):3825–31.
34. Quoix EA, Oster J, Westeel V, Pichon E, Zalcman G, Baudrin L, Lavole A, Dauba J, Lebitasy M, Milleron BJ. Weekly paclitaxel combined with monthly carboplatin versus single-agent therapy in patients age 70 to 89: IFCT-0501 randomized phase III study in advanced non-small cell lung cancer (NSCLC). J Clin Oncol. 2010;28(18):2.
35. Lynch TJ, Bell DW, Sordella R, Gurubhagavatula S, Okimoto RA, Brannigan BW, Harris PL, Haserlat SM, Supko JG, Haluska FG, et al. Activating mutations in the epidermal growth factor receptor underlying responsiveness of non-small-cell lung cancer to gefitinib. N Engl J Med. 2004;350(21):2129–39.
36. Paez JG, Janne PA, Lee JC, Tracy S, Greulich H, Gabriel S, Herman P, Kaye FJ, Lindeman N, Boggon TJ, et al. EGFR mutations in lung cancer: correlation with clinical response to gefitinib therapy. Science. 2004;304(5676):1497–500.
37. Mitsudomi T, Morita S, Yatabe Y, Negoro S, Okamoto I, Tsurutani J, Seto T, Satouchi M, Tada H, Hirashima T, et al. Gefitinib versus cisplatin plus docetaxel in patients with non-small-cell lung cancer harbouring mutations of the epidermal growth factor receptor (WJTOG3405): an open label, randomised phase 3 trial. Lancet Oncol. 2010;11(2):121–8.
38. Maemondo M, Inoue A, Kobayashi K, Sugawara S, Oizumi S, Isobe H, Gemma A, Harada M, Yoshizawa H, Kinoshita I, et al. Gefitinib or chemotherapy for non-small-cell lung cancer with mutated EGFR. N Engl J Med. 2010;362(25):2380–8.
39. Montagut C, Dalmases A, Bellosillo B, Crespo M, Pairet S, Iglesias M, Salido M, Gallen M, Marsters S, Tsai SP, et al. Identification of a mutation in the extracellular domain of the epidermal growth factor receptor conferring cetuximab resistance in colorectal cancer. Nat Med. 2012;18(2):221–3.
40. Wheeler DL, Huang S, Kruser TJ, Nechrebecki MM, Armstrong EA, Benavente S, Gondi V, Hsu KT, Harari PM. Mechanisms of acquired resistance to cetuximab: role of HER (ErbB) family members. Oncogene. 2008;27(28):3944–56.
41. Bianco R, Rosa R, Damiano V, Daniele G, Gelardi T, Garofalo S, Tarallo V, De Falco S, Melisi D, Benelli R, et al. Vascular endothelial growth factor receptor-1 contributes to resistance to anti-epidermal growth factor receptor drugs in human cancer cells. Clin Cancer Res. 2008;14(16):5069–80.

42. Huang X, Gao L, Wang S, McManaman JL, Thor AD, Yang X, Esteva FJ, Liu B. Heterotrimerization of the growth factor receptors erbB2, erbB3, and insulin-like growth factor-i receptor in breast cancer cells resistant to herceptin. Cancer Res. 2010;70(3):1204–14.
43. Emery CM, Vijayendran KG, Zipser MC, Sawyer AM, Niu L, Kim JJ, Hatton C, Chopra R, Oberholzer PA, Karpova MB, et al. MEK1 mutations confer resistance to MEK and B-RAF inhibition. Proc Natl Acad Sci U S A. 2009;106(48):20411–6.
44. Huber-Keener KJ, Liu X, Wang Z, Wang Y, Freeman W, Wu S, Planas-Silva MD, Ren X, Cheng Y, Zhang Y, et al. Differential gene expression in tamoxifen-resistant breast cancer cells revealed by a new analytical model of RNA-Seq data. PLoS One. 2012;7(7):e41333.
45. Ware KE, Hinz TK, Kleczko E, Singleton KR, Marek LA, Helfrich BA, Cummings CT, Graham DK, Astling D, Tan AC, et al. A mechanism of resistance to gefitinib mediated by cellular reprogramming and the acquisition of an FGF2-FGFR1 autocrine growth loop. Oncogenesis. 2013;2:e39.
46. Engelman JA, Zejnullahu K, Mitsudomi T, Song Y, Hyland C, Park JO, Lindeman N, Gale CM, Zhao X, Christensen J, et al. MET amplification leads to gefitinib resistance in lung cancer by activating ERBB3 signaling. Science. 2007;316(5827):1039–43.
47. Berns K, Horlings HM, Hennessy BT, Madiredjo M, Hijmans EM, Beelen K, Linn SC, Gonzalez-Angulo AM, Stemke-Hale K, Hauptmann M, et al. A functional genetic approach identifies the PI3K pathway as a major determinant of trastuzumab resistance in breast cancer. Cancer Cell. 2007;12(4):395–402.
48. Iorns E, Turner NC, Elliott R, Syed N, Garrone O, Gasco M, Tutt ANJ, Crook T, Lord CJ, Ashworth A. Identification of CDK10 as an important determinant of resistance to endocrine therapy for breast cancer. Cancer Cell. 2008;13(2):91–104.
49. Huang S, Holzel M, Knijnenburg T, Schlicker A, Roepman P, McDermott U, Garnett M, Grernrum W, Sun C, Prahallad A, et al. MED12 controls the response to multiple cancer drugs through regulation of TGF-beta receptor signaling. Cell. 2012;151(5):937–50.
50. Johannessen CM, Boehm JS, Kim SY, Thomas SR, Wardwell L, Johnson LA, Emery CM, Stransky N, Cogdill AP, Barretina J, et al. COT drives resistance to RAF inhibition through MAP kinase pathway reactivation. Nature. 2010;468(7326):968–72.
51. Masuda S, Izpisua Belmonte JC. The microenvironment and resistance to personalized cancer therapy. Nat Rev Clin Oncol. 2013;10(2). doi:10.1038/nrclinonc.2012.127-c1.
52. Jones M, Siracky J, Kelland LR, Harrap KR. Acquisition of platinum drug resistance and platinum cross resistance patterns in a panel of human ovarian carcinoma xenografts. Br J Cancer. 1993;67(1):24–9.
53. Liu P, Cheng H, Santiago S, Raeder M, Zhang F, Isabella A, Yang J, Semaan DJ, Chen C, Fox EA, et al. Oncogenic PIK3CA-driven mammary tumors frequently recur via PI3K pathway-dependent and PI3K pathway-independent mechanisms. Nat Med. 2011;17(9):1116–20.
54. Politi K, Fan PD, Shen R, Zakowski M, Varmus H. Erlotinib resistance in mouse models of epidermal growth factor receptor-induced lung adenocarcinoma. Dis Model Mech. 2010;3(1–2):111–9.
55. Pao W, Miller VA, Politi KA, Riely GJ, Somwar R, Zakowski MF, Kris MG, Varmus H. Acquired resistance of lung adenocarcinomas to gefitinib or erlotinib is associated with a second mutation in the EGFR kinase domain. PLoS Med. 2005;2(3):e73.
56. Morton CL, Houghton PJ. Establishment of human tumor xenografts in immunodeficient mice. Nat Protoc. 2007;2(2):247–50.
57. Tentler JJ, Tan AC, Weekes CD, Jimeno A, Leong S, Pitts TM, Arcaroli JJ, Messersmith WA, Eckhardt SG. Patient-derived tumour xenografts as models for oncology drug development. Nat Rev Clin Oncol. 2012;9(6):338–50.
58. Feldinger K, Generali D, Kramer-Marek G, Gijsen M, Ng TB, Wong JH, Strina C, Cappelletti M, Andreis D, Li J-L, et al. ADAM10 mediates trastuzumab resistance and is correlated with survival in HER2 positive breast cancer. Oncotarget. 2014;5(16):6633–46.
59. Wagle N, Emery C, Berger MF, Davis MJ, Sawyer A, Pochanard P, Kehoe SM, Johannessen CM, Macconaill LE, Hahn WC, et al. Dissecting therapeutic resistance to RAF inhibition in melanoma by tumor genomic profiling. J Clin Oncol. 2011;29(22):3085–96.

60. Carey LA, Rugo HS, Marcom PK, Mayer EL, Esteva FJ, Ma CX, Liu MC, Storniolo AM, Rimawi MF, Forero-Torres A, et al. TBCRC 001: randomized phase II study of cetuximab in combination with carboplatin in stage IV triple-negative breast cancer. J Clin Oncol. 2012;30(21):2615–23.
61. Balko JM, Giltnane JM, Wang K, Schwarz LJ, Young CD, Cook RS, Owens P, Sanders ME, Kuba MG, Sanchez V, et al. Molecular profiling of the residual disease of triple-negative breast cancers after neoadjuvant chemotherapy identifies actionable therapeutic targets. Cancer Discov. 2014;4(2):232–45.
62. Gorre ME, Mohammed M, Ellwood K, Hsu N, Paquette R, Rao PN, Sawyers CL. Clinical resistance to STI-571 cancer therapy caused by BCR-ABL gene mutation or amplification. Science. 2001;293(5531):876–80.
63. Shah NP, Tran C, Lee FY, Chen P, Norris D, Sawyers CL. Overriding imatinib resistance with a novel ABL kinase inhibitor. Science. 2004;305(5682):399–401.
64. Kantarjian H, Giles F, Wunderle L, Bhalla K, O'Brien S, Wassmann B, Tanaka C, Manley P, Rae P, Mietlowski W, et al. Nilotinib in imatinib-resistant CML and Philadelphia chromosome-positive ALL. N Engl J Med. 2006;354(24):2542–51.
65. Lehar J, Zimmermann GR, Krueger AS, Molnar RA, Ledell JT, Heilbut AM, Short 3rd GF, Giusti LC, Nolan GP, Magid OA, et al. Chemical combination effects predict connectivity in biological systems. Mol Syst Biol. 2007;3:80.
66. Peifer M, Weiss J, Sos ML, Koker M, Heynck S, Netzer C, Fischer S, Rode H, Rauh D, Rahnenfuhrer J, et al. Analysis of compound synergy in high-throughput cellular screens by population-based lifetime modeling. PLoS One. 2010;5(1):e8919.
67. Lage K, Karlberg EO, Storling ZM, Olason PI, Pedersen AG, Rigina O, Hinsby AM, Tumer Z, Pociot F, Tommerup N, et al. A human phenome-interactome network of protein complexes implicated in genetic disorders. Nat Biotechnol. 2007;25(3):309–16.
68. Hood L, Heath JR, Phelps ME, Lin B. Systems biology and new technologies enable predictive and preventative medicine. Science. 2004;306(5696):640–3.

Mutational Similarities Across Cancers: Implications for Research, Diagnostics, and Personalized Therapy Design

Frederick Klauschen, Albrecht Stenzinger, and Daniel Heim

Abstract Oncology and cancer research are based on the principle that cancers are regarded as organ- and tissue-specific diseases. One of the central aspects of histopathological tumor diagnostics is to determine the tumor's anatomic origin and other morphological features that are the basis for selecting the appropriate therapy. Similarly, research programs are usually also focused on particular cancer entities. However, mutational tumor profiling performed with next-generation-sequencing techniques has made it possible to analyze whether this anatomical tumor classification is valid also on the genetic level. Here, we review recent evidence that substantial similarities exist among tumors across classical anatomic cancer entities on the mutational level. We furthermore discuss the implications of these complex mutational profiles and similarity patterns across cancers for diagnostics, research, and clinical study design and explain why the comprehensive genomic data should be complemented by functional proteomic analyses.

1 Introduction

The organ and tissue specificity of tumors is the basis of histological diagnostics and cancer therapy [1]. On the one side, the fact that tumors are classified based on their anatomic origin has historical reasons, because solid tumors were traditionally associated with certain body regions, which was the major diagnostic information pathologists provided to clinicians in the early days of cancer medicine. On the other hand, the anatomic approach is also supported by the microscopic observation that cells usually undergo a continuous transformation from benign to malignant morphological properties during oncogenesis. Moreover, tumors often retain morphological features of their tissue origin even after the development of distant metastases [2]. But above all, relying on the identification of the anatomic origin

F. Klauschen, M.D., Ph.D., M.Sc. (✉) • D. Heim, M.Sc.
Institute of Pathology, Charité Universitätsmedizin Berlin, Charitéplatz 1, Berlin 10117, Germany
e-mail: frederick.klauschen@charite.de

A. Stenzinger, M.D.
Institute of Pathology, University of Heidelberg, Berlin, Germany

W. Wu, H. Choudhry (eds.), *Next Generation Sequencing in Cancer Research, Volume 2*, DOI 10.1007/978-3-319-15811-2_4

and other histomorphological features, such as the tumor grade, has proven to be an effective means of assessing disease prognosis, stratifying patients, and choosing the appropriate (chemo-)therapy. Continuous improvement of this approach has led to substantial advances in oncology during the last decades and even with the advent of individualized precision oncology [3] has the anatomic cancer classification not lost its importance. Molecular alterations that may be exploited therapeutically, so-called actionable or druggable mutations, are currently regarded as a refinement and not a replacement of the established tumor classification by the World Health Organization (WHO). However, such refinements are currently based only on relatively few and single genetic alterations [4], whereas comprehensive next-generation-sequencing-based [5] mutational profiling now allows to measure practically all mutations in individual tumors tieing in with the notion that cancer is a *genomic* disease, raised already by Theodor Boveri over a century ago when he proposed that cancer is caused by chromosomal derangements [6].

2 Mutational Similarities Across Cancers

The mutational profiling information obtained through next-generation sequencing has made it possible to address the question whether molecular profiles and, in particular, mutational signatures are in line with the conventional anatomic tumor classification and whether the latter needs to continue to be refined by the additional molecular features or replaced by a novel genetics-only-based tumor classification that no longer requires tumors to be classified histologically, i.e., according to their organ- and tissue-origin (Fig. 1). Studies on selected cancer entities have already provided evidence that genetic similarities exist among certain breast, endometrial, and ovarian cancers [7–9]. The comprehensive molecular profiling data now available for thousands of tumors [10] from different cancer entitites through projects such as The Cancer Genome Atlas (TCGA) or the International Cancer Genome Consortium (ICGC) allow for a more systematic evaluation of the mutational patterns across all major cancers and several studies have recently investigated this issue from different perspectives. A study by Alexandrov et al. [11] has found 20 distinct DNA sequence signatures that are defined by different patterns of nucleotide substitutions that recur in 30 different cancer types including various carcinomas, brain tumors as well as hematological malignancies and melanoma. While some gene sequence signature compositions are enriched in certain cancer types and therefore reflect the classical anatomic cancer classification many tumors of the same type differ substantially with respect to these signatures.

Other studies do not consider alterations based on the detailed DNA sequence changes as in the above study by Alexandrov et al., but define signatures based on sets of mutated genes. Ciriello et al. [12], for instance, analyze over 3,000 tumors from 12 cancer types available at TCGA and describe signatures based on genes with somatic mutations, copy number variations and methylation events and observe alterations enriched in either somatic mutations (M) or copy number variations (C) depending on the cancer type. Ovarian and breast cancer, for instance, are

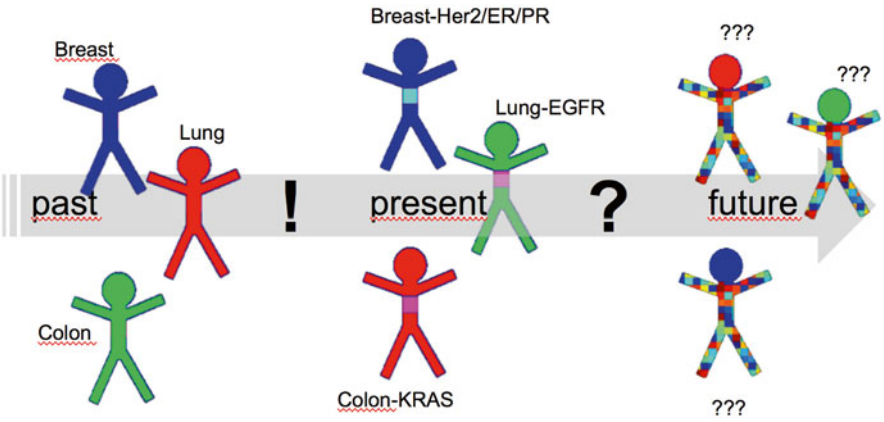


Fig. 1 Transition from a purely anatomic cancer classification (past) to concepts of tumors that rely on the conventional tumor classification but incorporate mutational information (present). Will the increasing molecular characterization of tumors lead to a purely genetic tumor classification (in future)?

characterized mainly by copy number variations tumors, whereas kidney cancer and colorectal cancers belong to the simple somatic mutation group. Based on these findings they propose a novel genetic tumor classification that divides tumors into the two top-level classes M (simple somatic mutations) and C (copy number variation) tumors and the subclasses M1–M8, M9–M14, M15, M16, M17 as well as C1–C6 and C7–C14 indicating different signaling pathway modules associated with the different mutational profiles. Classes M9–M14, for instance, correspond to different sets of genetic alterations in the Wnt and MAPK signaling pathways, whereas classes C9–C11 indicate enrichment in copy number variations affecting MYC-driven proliferation signaling.

In a complementary analysis [13, 14] of TCGA data we studied over 4,700 tumors from 14 cancer types and computed a systematic similarity map on the level of simple somatic mutations (comprising non-silent insertions, deletions, and substitutions) to systematically explore to what extent mutational patterns are in line with the conventional organ- and tissue-based tumor classification. For each of the over 4,700 tumors we searched for the most similar tumor in terms of the mutational profiles using a mutation concordance measure and found that, on average 43 % of all tumors of a given anatomic origin are genetically more similar to tumors arising at a different organ/tissue than to other tumors of the same anatomic site. Accordingly, only about 57 % of the tumors, on average, show cancer-type-specific mutational profiles. Interestingly, mutational profile similarities do not only occur among tumors arising in the same organ, such as adenocarcinoma and squamous cell cancer of the lung (about half of the lung squamous cell carcinomas are highly similar to adenocarcinomas of the lung) or among tumors of the same histological type, such as adenocarcinomas from different organs, where, for instance, about 13 % of the breast adenocarcinomas have their closest mutational relative among the ovarian adenocarcinomas and 3 % of breast tumors closely resemble gastric adenocarcinomas.

Interestingly, such similarity patterns are also present among tumors that neither share the same organ-origin nor the same tissue-origin, but that arise in completely unrelated anatomic structures, such as melanoma, glioblastoma, acute myeloid leukemia and carcinomas. About 22 % of the acute myeloid leukemia cases, for instance, show highest mutational similarities to carinomas arising in ovaries, the breast, kidney, and thyroid as well as sporadically to glioblastoma, melanoma, colon and head and neck cancers. Equivalent observations can be made for glioblastomas that show substantial genetic similarities with carcinomas and even AML and melanomas in about 54 % of the cases (Fig. 2). Of note, the results presented here are

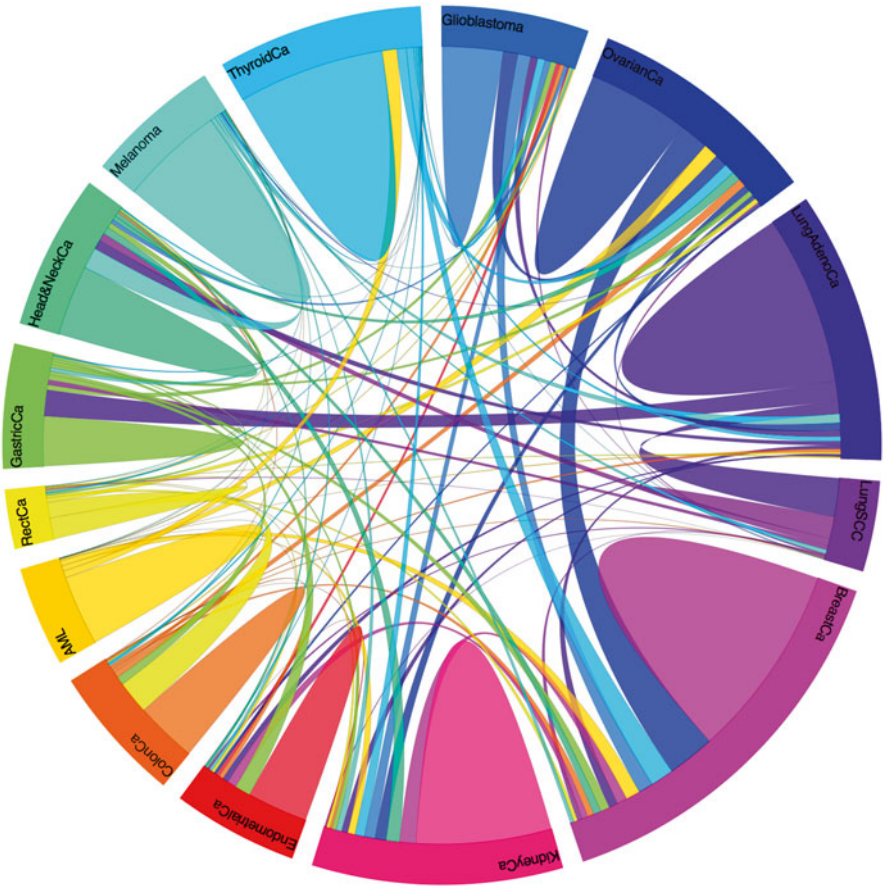


Fig. 2 Mutational similarities of 4,796 tumors from 14 major cancers from the TCGA database for all 24,858 genes. Mutational profiles disagree with the established cancer classification in 43 % of the cases on average. Chord connections between pairs of tumor types indicate the number of tumors of a particular type that are more similar to tumors from the other entity than to tumor of their own anatomic origin on the level of mutational profiles. As an example, about 13 % of the breast cancer cases are genetically more similar to ovarian cancers that to other breast tumors. Hill-like structures indicate tumors that resemble tumor of the same type most and are therefore in line with the conventional tumor classification (55 % on average)

based on all 24,845 genes available in the TCGA data that—except for the fact that silent mutations were excluded—were not weighted with respect to their different oncogenetic relevance. Therefore, including all genes in the similarity analysis might potentially lead to similarity patterns that are pathologically irrelevant. To exclude such confounders, alternative scenarios were compared performing the same similarity analysis for (1) a set of 400 cancer-related genes (used in cancer panel sequencing), (2) the 1,000 most frequently mutated genes as well as (3) genes that were classified as functionally relevant members of cancer-related cellular pathways in the Molecular Signatures Data Base (MsigDB) [15]. While minor quantitative differences are present among the different scenarios, the overall similarty patterns are remarkably stable across the different assumptions indicating the biological relevance of the findings [13]. The conclusion of this study is that the mutational profiles correspond to the established anatomic tumor classification—that is the basis of diagnostics and therapy—only in 57 % of the cases, on average, and that the remaining tumors are genetically more similar to tumors from other anatomic sites.

3 Genetic vs. Histological Tumor Classification

At first glance, the results of the studies reviewed here might suggest that the current histological tumor classification needs to be substantially amended if not replaced by a—yet to be defined—novel genetic tumor classification. However, while it is unlikely that the established tumor classification will remain unaffected by the molecular profiling efforts in the coming years, the success of using histomorphological features in estimating disease prognosis and therefore therapy response should not be underestimated. Ultimately, the answer to this question will hinge on the clinical relevance of genetic tumor classifications. With novel precision medicine approaches that rely on the detection of actionable or druggable mutations indicating the efficacy of a targeted therapy in individual patients, a central question is if mutations known to be druggable in one cancer are also druggable in a cancer of a different anatomic origin. Several cases with sometimes well-established and sometimes anecdotal evidence exists for and against the clinical utility of the same targeted therapy in different cancers that harbor the same druggable mutation. As an example, both patients with breast and gastric cancer, in which the growth factor receptor ERBB2/HER2 is amplified, usually benefit from HER2-inhibitory therapy although to a different extent. A strikingly different example are BRAF V600E mutations in melanoma and colorectal cancer. While mutated BRAF may (at least initially) be effectively treated by the inhibitor Vemurafenib, colorectal cancers with the same mutations are resistant [16]. These examples illustrate that the therapeutic benefit of a targeted therapy against an actionable mutation shown in one cancer cannot necessarily be transfered to another. Nevertheless, because of the potentially enormous benefit of transferring targeted therapies across cancer types, so-called “basket trials” have been designed to include patients based on the presence of druggable mutations in their tumors irrespective of the anatomic origin of the tumor [17–19].

The different responses to targeted therapy of tumors with the same mutations but that arise at different anatomic sites show that identical mutations may have different functional effects in different cancers. This may be due to the often high number of mutations in tumors, where considering just single druggable mutations underestimates the biological complexity of the oncogenic mechanisms. The TCGA study on squamous cell lung cancer, for instance, has reported 360 exon mutations, 165 genomic rearrangements, and 323 segments with copy number variations per tumor on average [20]. Although a substantial amount of these mutations is believed to be functionally irrelevant, it is obvious that different sets of accompanying mutations exist that are likely to modulate the effect of the druggable mutations differently in different cancers. Additionally, other influences such as for instance the local tissue chemokine composition or metabolic features may modulate the functional effects of certain mutations in a cancer-type dependent manner. It is therefore questionable not only if novel concepts such as basket trials will provide a solution but also whether NGS-driven approaches in precision medicine will live up to the high expectations unless complemented by histomorphological and more functionally oriented molecular analyses.

4 Implications of Mutational Tumor Profiling for Diagnostics and Clinical Trial Design

The observed discordance between observed mutational profiles and the established anatomic tumor classification in combination with the complexity of the mutational profiles with often low mutation frequencies lead to substantial difficulties in the design of clinical studies that evaluate targeted therapies. A prominent example is the clinical trial that showed the utility of Crizotinib in EML4-ALK positive lung cancer patients [21]. One hundred and five study centers in 27 countries were required to recruit 347 patients with EML4-ALK-positive tumors for this trial. While it might be surprising at first glance that so many centers had to collaborate, particularly given the relatively small number of patients in the trial, it becomes obvious when considering the fact that less than 5 % of lung adenocarcinomas harbor the EML4-ALK gene fusion [22].

In the future, the situation is likely to become even more difficult as combination therapies will almost certainly replace or complement current targeted approaches using single drugs against which resistance develops in almost all cases. While the rapidly increasing number of known actionable mutations and available corresponding targeted drugs makes such combination therapies technically feasible, the critical question how to select and test an appropriate drug combination remains open. Even if the majority of the 360 exon mutations in an average squamous cell lung cancer are not causally linked with cancer pathology as discussed above, genomic alterations such as EML4-ALK rearrangements or EGFR point mutations demonstrate clearly that also low-frequency mutations may represent important drug targets and that not only the ten or so most frequently mutated genes are functionally involved in oncogenetic processes and therefore the only drug targets.

Although the major cancer-related signaling pathways are well known conceptually, knowledge on the precise pathway dynamics, topology, and cross talk is incomplete. As a consequence of the lack of sufficient knowledge to allow for a rigorous preselection of drug combinations, multiple drug combinations would need to be tested preclinically, but ultimately a substantial number would also remain to be evaluated in clinical trials. Here, we review data that we have published previously on a systematic evaluation of the combinatorial of cancer precision medicine [14, 23]. As an illustration, in the case of a combination of two drugs selected out of a library against ten actionable mutations, 45 different combinations exist. While this may still be considered manageable, 190 possible 2-drug-combination therapies exist when the library has 20 components and 1,140 combinations need to be considered for a therapy of three drugs. Although a 5-drug combination selected out of a library of 50 drugs resulting in 2,118,760 possible alternatives may not seem clinically relevant in the near future, these numbers illustrate the challenges that arise already with relatively conservative assumptions. And even if it may be possible to exclude 90 % of the theoretically possible combinations through knowledge on cellular processes and preclinical experimental testing, where the combinatorial complexity described here will also lead to a significant increase in cost, a substantial number of alternative combination therapies remains to be tested in clinical trials. In this scenario, novel clinical trial concepts, such as the abovementioned basket trials will quickly reach their limits.

5 Conclusion

Next-generation sequencing technologies have become widely available in the recent years and now allow a comprehensive characterization of the mutational profiles in individual tumors. This detailed knowledge has raised hope to gain also a deeper understanding of the molecular mechanisms responsible for the development of cancer and to facilitate the development of novel targeted therapies in cancer precision medicine. The fact that despite the undoubtedly substantial gain in knowledge on the molecular properties of cancer the overall clinical utility of NGS-driven approaches has so far been limited, is due to the complexity and high variability of the mutational profiles also within defined tumor types. With respect to the genetic similarities across classical tumor types described here and the effects of actionable mutations on tumor cell function that are at least partially dependent on the organ- and tissue context and additional mutations present in the tumor, it becomes obvious that a static, exclusively genetic view of cancer without taking into account the functional implications of the mutational profiles is insufficient. This also applies to the design of current clinical trials, which is not laid out to handle the combinatorial complexity of personalized combination therapies in cancer. A possible solution to this dilemma may lie in an integration of genomic with proteomic approaches. First steps in this direction have already been made recently by an integration of proteomic data into the TCGA database, which allows to directly relate mutational and proteomic profiles [24–26]. However, we believe

that these static data need to be complemented by dynamic analyses of tumor cell function to gain a better understanding of the processes relevant for the design of precision therapies in oncology in the future.

Acknowledgement F.K. is supported as an Einstein Junior Fellow by the Einstein Foundations Berlin and the Human Frontier Science Program Organization as a Young Investigator. The work reviewed here has partly been previously published by the author in form of primary research articles and in conference proceedings in German [13, 14, 23].

References

1. Sobin LH. TNM: principles, history, and relation to other prognostic factors. Cancer. 2001;91:1589–92.
2. Klein CA. Selection and adaptation during metastatic cancer progression. Nature. 2013;501:365–72.
3. Garraway LA, Verweij J, Ballman KV. Precision oncology: an overview. J Clin Oncol. 2013;31:1803–5.
4. McDermott U, Downing JR, Stratton MR. Genomics and the continuum of cancer care. N Engl J Med. 2011;364:340–50.
5. MacConaill LE. Existing and emerging technologies for tumor genomic profiling. J Clin Oncol. 2013;31:1815–24.
6. Boveri T. Concerning the origin of malignant tumours by Theodor Boveri. Translated and annotated by Henry Harris. J Cell Sci. 2008;121 Suppl 1:1–84.
7. Cancer Genome Atlas Research Network. Integrated genomic analyses of ovarian carcinoma. Nature. 2011;474:609–15.
8. Cancer Genome Atlas Network. Comprehensive molecular portraits of human breast tumours. Nature. 2012;490:61–70.
9. Kandoth C, Schultz N, Cherniack AD, Akbani R, Liu Y, Shen H, Robertson AG, Pashtan I, Shen R, Benz CC, et al. Integrated genomic characterization of endometrial carcinoma. Nature. 2013;497:67–73.
10. Hudson TJ, Anderson W, Artez A, Barker AD, Bell C, Bernabe RR, Bhan MK, Calvo F, Eerola I, Gerhard DS, et al. International network of cancer genome projects. Nature. 2010;464:993–8.
11. Alexandrov LB, Nik-Zainal S, Wedge DC, Aparicio SA, Behjati S, Biankin AV, Bignell GR, Bolli N, Borg A, Borresen-Dale AL, et al. Signatures of mutational processes in human cancer. Nature. 2013;500:415–21.
12. Ciriello G, Miller ML, Aksoy BA, Senbabaoglu Y, Schultz N, Sander C. Emerging landscape of oncogenic signatures across human cancers. Nat Genet. 2013;45:1127–33.
13. Heim D, Budczies J, Stenzinger A, Treue D, Hufnagl P, Denkert C, Dietel M, Klauschen F. Cancer beyond organ and tissue specificity: next-generation-sequencing gene mutation data reveal complex genetic similarities across major cancers. Int J Cancer. 2014;135(10):2362–9.
14. Klauschen F. Mutational tumor profiles beyond organ and tissue specificity: implications for diagnostics and clinical study design. Pathologe. 2014;35 Suppl 2:277–80.
15. Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA, Paulovich A, Pomeroy SL, Golub TR, Lander ES, Mesirov JP. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. Proc Natl Acad Sci U S A. 2005;102:15545–50.
16. Prahallad A, Sun C, Huang S, Di Nicolantonio F, Salazar R, Zecchin D, Beijersbergen RL, Bardelli A, Bernards R. Unresponsiveness of colon cancer to BRAF(V600E) inhibition through feedback activation of EGFR. Nature. 2012;483:100–3.

17. Baselga J. Bringing precision medicine to the clinic: from genomic profiling to the power of clinical observation. Ann Oncol. 2013;24:1956–7.
18. Kim ES, Herbst RS, Wistuba II, Lee JJ, Blumenschein Jr GR, Tsao A, Stewart DJ, Hicks ME, Erasmus Jr J, Gupta S, et al. The BATTLE trial: personalizing therapy for lung cancer. Cancer Discov. 2011;1:44–53.
19. Mills GB. An emerging toolkit for targeted cancer therapies. Genome Res. 2012;22:177–82.
20. Cancer Genome Atlas Research Network. Comprehensive genomic characterization of squamous cell lung cancers. Nature. 2012;489:519–25.
21. Shaw AT, Kim DW, Nakagawa K, Seto T, Crino L, Ahn MJ, De Pas T, Besse B, Solomon BJ, Blackhall F, et al. Crizotinib versus chemotherapy in advanced ALK-positive lung cancer. N Engl J Med. 2013;368:2385–94.
22. Soda M, Choi YL, Enomoto M, Takada S, Yamashita Y, Ishikawa S, Fujiwara S, Watanabe H, Kurashina K, Hatanaka H, et al. Identification of the transforming EML4-ALK fusion gene in non-small-cell lung cancer. Nature. 2007;448:561–6.
23. Klauschen F, Andreef M, Keilholz U, Dietel M, Stenzinger A. The combinatorial complexity of cancer precision medicine. Oncoscience. 2014;1:504–9.
24. Cancer Genome Atlas Research Network. Comprehensive molecular profiling of lung adenocarcinoma. Nature. 2014;511:543–50.
25. Wilhelm M, Schlegl J, Hahne H, Moghaddas Gholami A, Lieberenz M, Savitski MM, Ziegler E, Butzmann L, Gessulat S, Marx H, et al. Mass-spectrometry-based draft of the human proteome. Nature. 2014;509:582–7.
26. Zhang B, Wang J, Wang X, Zhu J, Liu Q, Shi Z, Chambers MC, Zimmerman LJ, Shaddox KF, Kim S, et al. Proteogenomic characterization of human colon and rectal cancer. Nature. 2014;513(7518):382–7.

Standardized Decision Support in NGS Reports of Somatic Cancer Variants

Rodrigo Dienstmann

Abstract With the advent of next-generation sequencing (NGS), we have the promise of a complete genetic description of patient tumors to optimally direct therapy. Of hundreds to thousands of somatic mutations that exist in each cancer genome, a large number are unique and nonrecurrent variants. Prioritizing and annotating genetic variants identified via NGS technologies remains a major challenge. Some variants occur in tumor genes that have well-established biological and clinical relevance and are putative targets of therapy. However, most variants have limited evidence as predictive markers or are still of unknown significance. Furthermore, how to prioritize therapy when multiple potentially targetable aberrations and/or coexisting resistance mechanisms are identified in a patient's tumor still remains largely a heuristic task. In this context, there is a growing need for the biomedical research community to have access to curated and up-to-date cancer pharmacogenomic associations. In addition, the community needs to remain cognizant of the potential consequences of misuse or overinterpretation of genomic data. Herein, I describe a systematic framework for variant annotation and prioritization and propose a structured molecular pathology report using standardized terminology in order to best inform oncology clinical practice.

1 Introduction

Clinical laboratories increasingly view large cancer gene panels and NGS as a cost-effective—and tissue-saving—alternative to running a series of multiple single-gene companion tests. Large amounts of genomic data are being generated as these assays enter the clinical realm, challenging molecular pathologists and cancer genomicists in charge of interpreting and reporting the results. Manually annotating each single variant in terms of clinical significance in every possible tumor type is a

R. Dienstmann (✉)
Sage Bionetworks, Fred Hutchinson Cancer Research Center,
1100 Fairview N, Seattle, WA 98109, USA
e-mail: rodrigo.dienstmann@sagebase.org

W. Wu, H. Choudhry (eds.), *Next Generation Sequencing in Cancer Research, Volume 2*, DOI 10.1007/978-3-319-15811-2_5

daunting task. In addition, the strain on the turnaround time drives the need for prioritization strategies for the identification and reporting of clinically significant genetic variants.

Routine testing of full gene sequences as opposed to hot spots frequently identifies mutations of low frequency and unknown functional consequences, most of which are likely to be neutral or passenger alterations. On the other hand, some variants occur in cancer genes that have well-established clinical utility, driving tumorigenesis, and tumor progression. The available scientific knowledge on these mutations should be presented in the report, so that physicians and patients can make evidence-based decisions in a responsible fashion. Genetic results may provide a strong rationale for treatment with matched targeted agents in clinical trials, with the potential of directly benefitting the patient and accelerating the drug development process [1]. Consolidating so much information into a very discrete report that emphasizes the clinical significance while preserving observations that can be further looked into by the clinician is not an easy undertaking. As physicians trained in fields other than genetics are playing a more central role in the ordering and reviewing of genetic test results, the importance of translating genomic data into informative reports is further increased.

Performing NGS in the clinical laboratory is a multistep process that typically involves sample acquisition and quality control, DNA extraction, library preparation, sequencing, and genomic data generation. The process continues with three dynamic pipelines for data analysis: (1) bioinformatics tools for variant identification, (2) variant annotation and prioritization, and (3) interpretation of clinical significance and reporting to clinicians [2, 3]. In this chapter, I propose a framework for clinical interpretation of somatic cancer variants and describe how genomic data can be translated into structured evidence-based reports after a detailed variant annotation and prioritization process.

2 Prioritizing Cancer Genomic Variants

Following variant identification using bioinformatics pipelines, a computational engine is needed in order to parse the variants and suppress those that are irrelevant, highlight the ones which need manual curation, and identify pertinent “wild types” in each tumor sample. In the first step of variant prioritization, as summarized in Fig. 1, molecular pathologists have to define what is considered a “reportable” variant. Several annotation and prioritization parameters are taken into consideration so as to provide a stronger estimation of the functional significance of unknown and novel mutations. Useful tools include sequencing metric variables, external germ line single nucleotide polymorphisms (SNPs), and cancer databases for comparison of variants across populations, as well as prediction models for defining damaging/deleterious or potentially driver mutations, as discussed below.

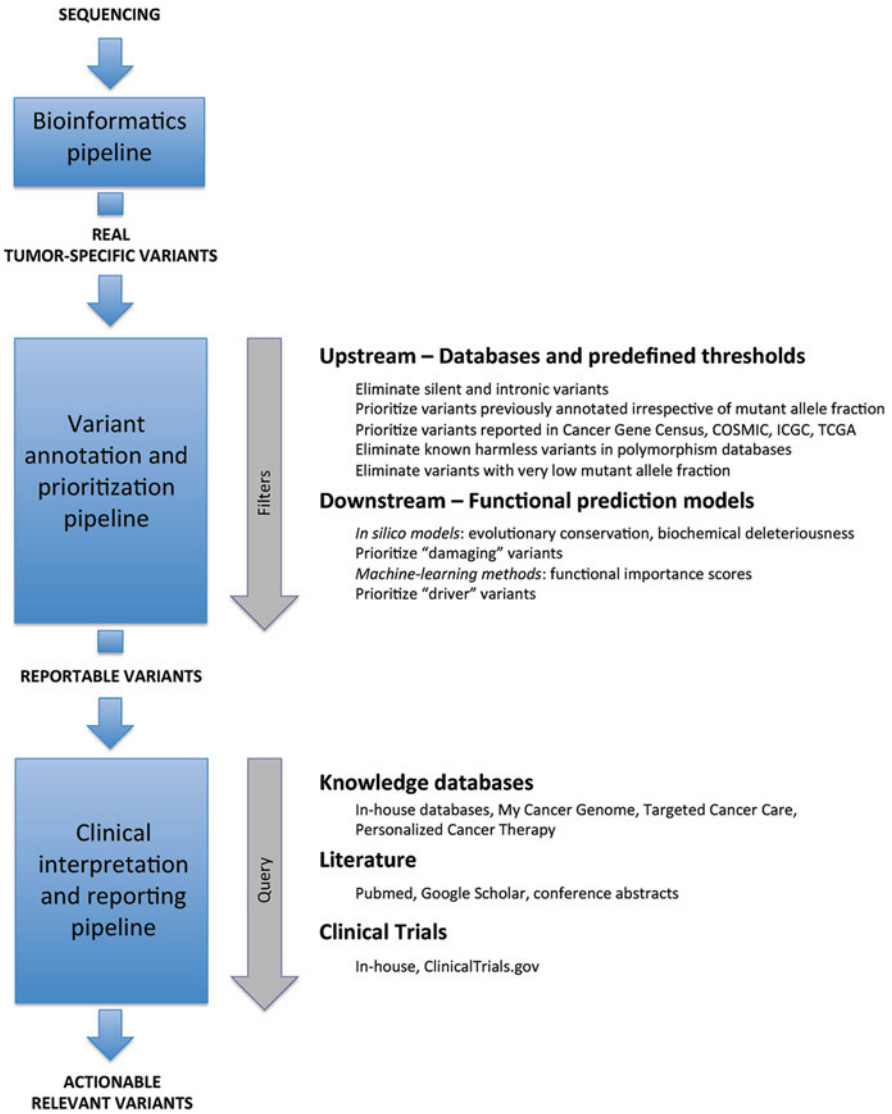


Fig. 1 Variant analysis flowchart of NGS tests performed in clinical laboratories. The bioinformatics pipeline identifies real and tumor-specific variants. During the variant annotation and prioritization pipeline, curated databases, predefined thresholds, and functional prediction models serve as filters, with reportable variants as final output. The clinical interpretation pipeline involves careful literature review and reporting of actionable variants

2.1 Upstream Filtering Tools

In the case of exome or whole genome sequencing, pairwise comparison with germ line DNA plays a pivotal role. Subtracting the genetic variation of a noncancerous "normal" genome from its cancerous counterpart allows the identification of the somatic mutations. In parallel, eliminating known harmless variants that are present in public or in-house polymorphism databases is a very helpful strategy for reducing the candidate list of deleterious mutations. The next step involves prioritizing missense, nonsense, or splice-site mutations over synonymous and intronic variants. Different bioinformatic adjustments can be used in order to improve variant detection and deal with library preparation or sequencing artifacts along with sample characteristics, including tumor purity and heterogeneity. In order to consider the variant as real and reportable, it is also advised to establish a minimum threshold of mutant allele fraction (MAF), the number of alternate reads at the genomic position divided by the total number of reads—coverage—at the same site. This threshold should take into consideration tumor cellularity and also clinical context, as rare resistant subclones in the treatment-refractory setting might be of relevance. Therefore, known gene variants previously clinically annotated are generally prioritized irrespective of MAF.

The most useful annotation tool for somatic variant interpretation involves the assessment of published cancer databases. The software used for variant prioritization should directly link genetic alterations to the Cancer Gene Census (http://cancer.sanger.ac.uk/cancergenome/projects/census/) or similar catalogues of genes for which mutations have been causally implicated in cancer [4], as well as the Catalogue of Somatic Mutations in Cancer (COSMIC) (http://cancer.sanger.ac.uk/cancergenome/projects/cosmic/), International Cancer Genome Consortium (ICGC) (https://dcc.icgc.org/), and The Cancer Genome Atlas (TCGA) (http://cancergenome.nih.gov/; http://www.cbioportal.org/), large cancer databases that present prevalence of gene variants in different tumor types. Assessing whether a newly discovered alteration may be functionally relevant rests heavily on how many times it has been reported in these international cancer genomics studies, supporting further clinical interpretation.

2.2 Downstream Filtering Tools

Prediction of the putative functional effect of a mutation is a common problem already addressed in the context of germ line SNP association studies, and several tools have been used for this purpose. These models annotate variants specifically with respect to evolutionary conservation, biochemical deleteriousness, and functional importance scores, thereby facilitating the differentiation between functional and nonfunctional variants [5–7]. At present, for alleles without prior functional analysis in genes that have been related to human cancer, such as non-hot spot/novel

variants in known oncogenes and tumor suppressor genes, prediction algorithms based on evolutionary conservation patterns are often used. Sorting Intolerant from Tolerant (SIFT) [8] and MutationAssessor [9] exploit the fact that sequences observed among living organisms are those that have not been removed by natural selection and sites with fewer observed substitutions are inferred to be under tighter constraints, having more deleterious effects when mutated. On the other hand, mutations in non-conserved residues are likely neutral. Other resources for predicting the effects of protein-coding sequence changes typically exploit the physicochemical properties of amino acids and information about the role of amino acid side chains in protein structure. These in silico protein sequence-based algorithms, such as PolyPhen2 [10], are capable of leveraging both evolutionary and biochemical information. Despite having high sensitivity for the detection of damaging variants, prediction tools that rely on conservation and structure should be used with caution. In addition to the low specificity, these methods generally have limited value in annotating gain-of-function or switch-of-function mutations [11]. Furthermore, most of these algorithms have been designed for research purposes with germ line variants, and very few databases present clinically oriented molecular annotation. As an alternative, machine learning scoring methods attempt to increase the predictive precision of somatic mutations in cancer. One example is the cancer-specific high-throughput annotation of somatic mutation (CHASM) tool, specifically designed to distinguish driver from passenger somatic missense variants [12]. It is trained on a positive class of drivers curated from the COSMIC database and a negative class of passenger variants generated in silico based on background base substitution in specific tumor types. Limitations include reduced coverage as compared to traditional algorithms—restriction to missense mutations—and the understanding that driver and passenger mutations are tumor type and context dependent, possibly changing roles during cancer evolution and therapy [7]. Whether cancer-trained methods outperform more general predictors still needs further investigation. Recent studies suggest that no method or combination of methods exceeds ~80 % accuracy [13, 14], indicating that there is still significant room for improvement in functional prediction, possibly with the development of specific algorithms for different classes of mutations.

To summarize, complex criteria involving multiple annotation sources should be used in order to select or filter out variants. Part of this process can be automated, although most of the work still needs to be done manually. As the most valuable tool consists in leveraging the cancer literature, either generated in-house or derived from publicly available databases, the genomic prioritization engine needs to be dynamic in nature, recognizing driver cancer mutations that have been previously annotated and reported. Additional tumor-specific variants with very low MAFs and those considered silent mutations are typically excluded from further clinical interpretation. Novel variants in genes that have been causally implicated in cancer are prioritized when functional models predict damaging/deleterious scores, the alteration is in the phosphorylation loop of an oncogenic kinase, or it alters the reading frame of a tumor suppressor gene.

3 Interpreting Results with Clinical Perspective

After narrowing down the list of candidate variants, the biggest challenge is to interpret the remaining genomic alterations within a biological context. Potentially "reportable" variants can be grouped in three categories: (1) those that may have a direct impact on patient care and are considered "actionable," (2) those that may have "biological relevance" but are not clearly actionable, and (3) those that are of "unknown significance." Different groups have varying definitions for clinically "actionable." This category can be restricted to variants matched to drugs that have been approved by regulatory agencies for the tumor that is being studied, but may also include those directing to off-label use of approved drugs, as well as variants that are matched to drugs being investigated in clinical trials. Academic laboratories should adopt the most inclusive definition of an actionable mutation—which accounts for variants that support treatment recommendation and enrollment in a particular clinical trial or have prognostic or diagnostic implications—even knowing that it may increase challenges in clinical decision-making, as the results sometimes lead to regulatory issues regarding the use of targeted drugs in unapproved indications.

Importantly, variants should not be reported in an uncategorized format, which can be confusing to clinicians and detrimental to patients. For actionable mutations to be fully curated, a team of experts with strong background in cancer biology and access to up-to-date knowledge resources is mandatory. Clinical interpretation of most variants identified in NGS-based cancer diagnostic tests involves the burdensome procedure of manually reviewing the published literature on four different layers: (1) gene, (2) specific variant, (3) drug or class-of-agent sensitivity/resistance patterns, and (4) tumor-type context. To facilitate this process, several groups have implemented "Sequencing Tumor Boards" or "Molecular Rounds" with up to 15 faculty members that share expertise in cancer genomics, bioinformatics, pathology, clinical genetics, bioethics, and clinical oncology as well as experimental therapeutics. Rigorous analysis of comprehensive genomic data is a time-consuming and labor-intensive task, considering that not many mutations have been validated with a high enough level of evidence to predict for response to targeted treatment. Experts should prioritize the knowledge on mutations in tumor-specific contexts, but curation of data derived from other tumor types and preclinical experiments—when clinical validation is under way—usually gives valuable information to clinicians. Unfortunately, most resources currently available cover information at limited levels: some focus on gene-tumor associations, others only on gene-drug or drug-target relationships. Moreover, databases originally developed to enable preclinical research or annotate germ line variants are of limited applicability for clinical oncology curation. Alternatively, associations on predictive, prognostic, or diagnostic variants in cancer can be retrieved in clinically oriented databases, such as My Cancer Genome (http://www.mycancergenome.org/), Targeted Cancer Care (http://www.targetedcancercare.org/), and Personalized Cancer Therapy (https://pct.mdanderson.org/). These websites are the result of large institutional efforts to provide information on cancer types, aberrant genes, and variants that are targeted by approved or experimental therapies. However, information available in these databases does not cover all genes, variants, and tumor types. In addition, it is not

accessible for download, mainly because it is presented in a descriptive format, without standardized terminology.

In order to deal with these limitations, some groups have developed internal knowledge databases with more comprehensive annotations on consensus and emerging clinical/preclinical predictive genomic markers linked to targeted therapies. When integrated to the variant prioritization computational engine and report generation system, the curated information on somatic variants that have been classified for clinical reporting is stored for future use. Maintenance of these databases involves a regular and systematic review of drug regulatory and approval status, consensus guidelines, peer-reviewed publications, and clinical trial databases. One example of detailed cancer genomics knowledge database is available for download through Synapse (https://www.synapse.org/#!Synapse:syn2370773), the collaborative cloud-based repository developed at Sage Bionetworks. As many academic groups are independently working on similar projects, an international consortium on curated cancer genomic data matching genomic aberrations to targeted therapies could have a huge clinical impact. Ideally, the information should be released as an interactive web-based tool, subjected to editing, validation, and critique from the medical community.

4 Generating NGS Reports

Previous studies evaluating single-gene reports have suggested that patient care may be compromised as a consequence of poor communication between laboratories and clinicians [15]. Developing a framework to content-rich NGS reports is complicated. The traditional "narrative" style reporting is too cumbersome for the amount of data generated by large cancer gene panels. In addition, medical oncologists prefer structured reports with results displayed in a more straightforward manner rather than detailed descriptions of each genomic alteration. Consequently, web-enabled technologies are a good alternative to text reports as they enable dynamic and interactive display of the NGS results, which could be accessed by providers and patients in different formats. Embedding links to internal and external databases allows members of the team to further explore the results and the evidence used to guide the interpretation, including more detailed information on the gene, the variant, the drug, or the clinical trial matched to a particular genomic alteration and tumor type, as well as records of PubMed identification numbers for relevant clinical literature. Unfortunately, most laboratory information systems and electronic medical records (EMR) to date do not support data formatting and metadata (data associated with the result). Therefore, reports may need to be oversimplified to a static format for inclusion in the EMR.

Wagle et al. reported the first framework to segregate genetic alterations derived from NGS tests on the basis of their predicted clinical utility [16]. The actionable category includes variants that predict tumor sensitivity or resistance to approved (tier 1) or experimental therapies (tier 2). As shown in Fig. 2a, the mutational categories are organized based on the strength of evidence supporting its predictive value. An alternative classification is presented in Fig. 2b, which represents a simplified

a

	Tier 1: FDA-approved/ standard therapies	Tier 2: Clinical trials/ experimental therapies	Prognostic/Diagnostic
A	Clinically-validated	Eligibility criteria for trial	Clinically validated
B	Limited evidence	Limited evidence	Limited evidence
C	Evidence in another tumor	Evidence in another tumor	
D	Preclinical evidence	Preclinical evidence	
E	Theoretical evidence	Theoretical evidence	

b

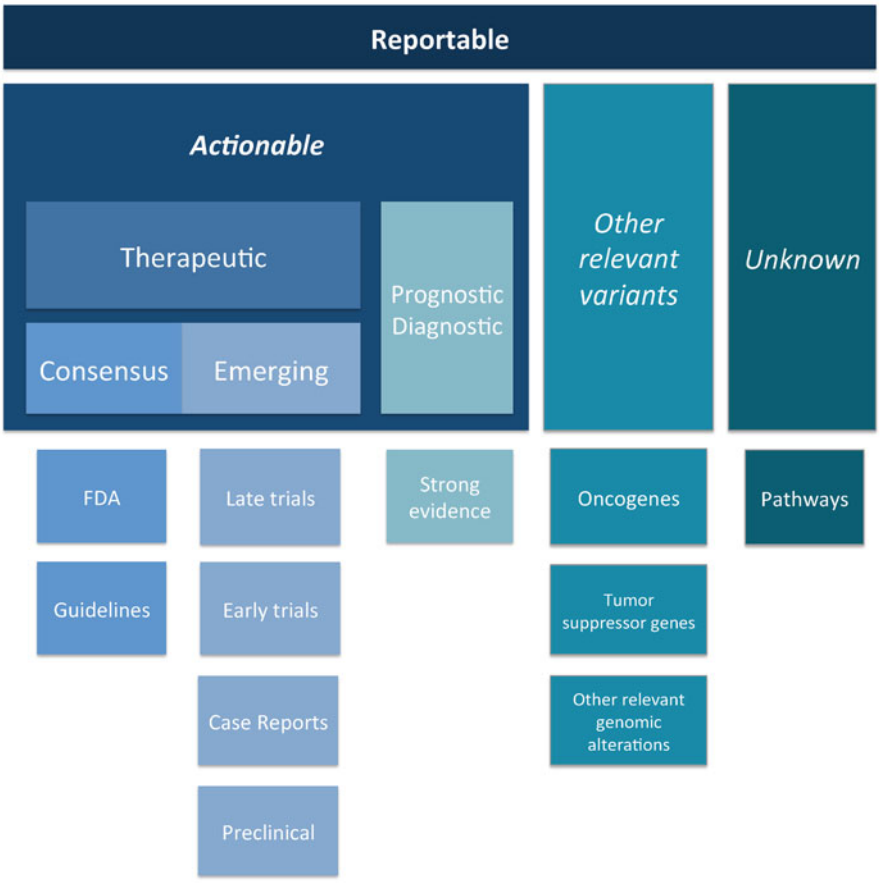
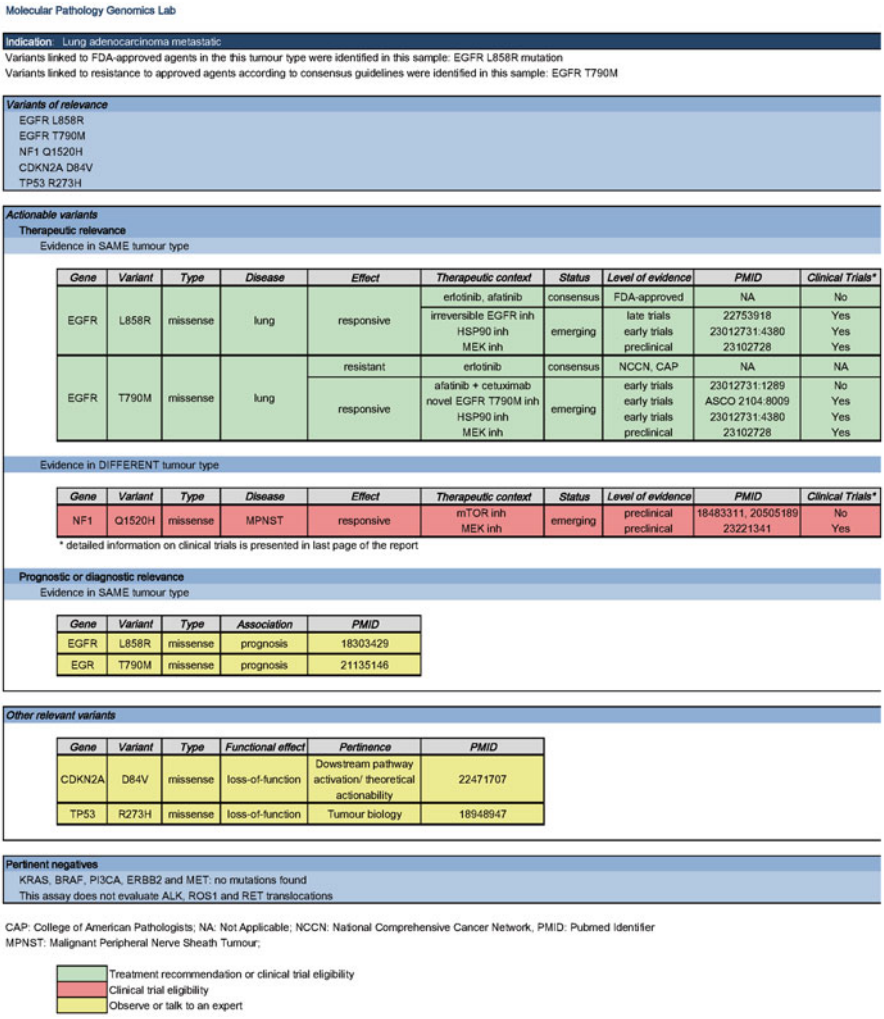


Fig. 2 Examples of somatic variant classification system for NGS reports. (**a**) Wagle et al. The actionable category includes variants that predict tumor sensitivity or resistance to approved (tier 1) or experimental therapies (tier 2) and those that have prognostic/diagnostic implications. (**b**) Dienstmann et al. Reportable variants can be grouped in three categories: (1) actionable, which support treatment recommendation (therapeutic consensus) and enrollment in clinical trials (therapeutic emerging) and/or have prognostic or diagnostic implications; (2) biologically relevant but not clearly actionable; and (3) unknown significance



Molecular Pathology Genomics Lab

Indication: Lung adenocarcinoma metastatic

Variants linked to FDA-approved agents in the this tumour type were identified in this sample: EGFR L858R mutation

Variants linked to resistance to approved agents according to consensus guidelines were identified in this sample: EGFR T790M

Variants of relevance

EGFR L858R
EGFR T790M
NF1 Q1520H
CDKN2A D84V
TP53 R273H

Actionable variants

Therapeutic relevance

Evidence in SAME tumour type

Gene	*Variant*	*Type*	*Disease*	*Effect*	*Therapeutic context*	*Status*	*Level of evidence*	*PMID*	*Clinical Trials**
EGFR	L858R	missense	lung	responsive	erlotinib, afatinib	consensus	FDA-approved	NA	No
					irreversible EGFR inh	emerging	late trials	22753918	Yes
					HSP90 inh		early trials	23012731:4380	Yes
					MEK inh		preclinical	23102728	Yes
EGFR	T790M	missense	lung	resistant	erlotinib	consensus	NCCN, CAP	NA	NA
				responsive	afatinib + cetuximab	emerging	early trials	23012731:1289	No
					novel EGFR T790M inh		early trials	ASCO 2104:8009	Yes
					HSP90 inh		early trials	23012731:4380	Yes
					MEK inh		preclinical	23102728	Yes

Evidence in DIFFERENT tumour type

Gene	*Variant*	*Type*	*Disease*	*Effect*	*Therapeutic context*	*Status*	*Level of evidence*	*PMID*	*Clinical Trials**
NF1	Q1520H	missense	MPNST	responsive	mTOR inh	emerging	preclinical	18483311, 20505189	No
					MEK inh		preclinical	23221341	Yes

* detailed information on clinical trials is presented in last page of the report

Prognostic or diagnostic relevance

Evidence in SAME tumour type

Gene	*Variant*	*Type*	*Association*	*PMID*
EGFR	L858R	missense	prognosis	18303429
EGR	T790M	missense	prognosis	21135146

Other relevant variants

Gene	*Variant*	*Type*	*Functional effect*	*Pertinence*	*PMID*
CDKN2A	D84V	missense	loss-of-function	Dowstream pathway activation/ theoretical actionability	22471707
TP53	R273H	missense	loss-of-function	Tumour biology	18948947

Pertinent negatives

KRAS, BRAF, PI3CA, ERBB2 and MET: no mutations found
This assay does not evaluate ALK, ROS1 and RET translocations

CAP: College of American Pathologists; NA: Not Applicable; NCCN: National Comprehensive Cancer Network, PMID: Pubmed Identifier
MPNST: Malignant Peripheral Nerve Sheath Tumour;

Treatment recommendation or clinical trial eligibility
Clinical trial eligibility
Observe or talk to an expert

Fig. 3 Illustrative example of sequencing results describing somatic cancer variants with structured evidence-based classification. Using the framework described in Fig. 2b, results are presented in a hierarchical and tabular format, drawing clinician's attention to associations with different levels of actionability

gene-oriented approach developed to facilitate clinical decision-making [17]. Reports based on this framework provide the information in a hierarchical/categorical format, and results can be structured in tabular view. The content is formatted in such a way as to draw the clinician's attention to associations with the highest level of evidence. As exemplified in Fig. 3, all actionable—predictive, prognostic, and diagnostic—markers are displayed first, followed by biologically relevant gene variants that warrant detailed annotation and pertinent negatives in the tumor being tested. Details are discussed in the following sections.

4.1 Predictive Associations

Consensus predictive associations include those (1) linked to drugs approved or rejected by regulatory agencies in the context of a specific gene variant and tumor type or (2) described in national guidelines as predicting response or resistance to specific therapies. Emerging predictive associations were classified in a hierarchical way based on the strength of evidence: (1) late trials, including evidence derived from trials that prospectively recruited patients based on genomic profiling as well as large trials with robust data suggesting sensitivity/resistance to targeted therapies based on retrospective analysis of biomarkers; (2) early trials, referring to phase 1 or 2 studies with genomically selected patients that show preliminary signs of efficacy (or lack of efficacy); (3) case reports of dramatic responses to targeted therapies in a specific genomic context; and (4) strong preclinical data that is being explored in clinical trials. The magnitude of the biomarker-drug effects for clinical associations is classified as "responsive," "resistant," or "not responsive" (when an expected responsive effect is not observed). In preclinical models, biomarker-drug associations are graded as "sensitive," "reduced sensitivity," or "resistant."

Some of the questions that scientists involved in clinical interpretation of genomic data have to deal with include:

- Is this an activating or inactivating mutation?
- Does this mutation engender sensitivity to targeted therapeutics—and what is the agent with highest potency?
- How to select therapy in case of multiple genomic alterations and/or coexisting resistance mechanisms?
- Is the association tumor type or context specific (treatment-naïve versus refractory setting) after exposure to which targeted agents?

Ideally, reports of NGS tests in oncology should include a list of clinical trials recruiting patients that harbor the specific genomic aberrations identified in the individual tumor sample. These are matched targeted therapies available either on-site or as part of multi-institutional collaborations. A current limitation for matching a patient's tumor genotype to clinical trials is the lack of molecular annotations in notices of national registries, such as the US National Cancer Institute clinical trial locator (www.clinicalTrials.gov). As an example, the search term "PIK3R1" does not identify any matched trial, even though many PI3K pathway inhibitors in clinical development have a clear rationale for testing in tumors that harbor *PIK3R1* inactivating mutations.

4.2 Prognostic and Diagnostic Associations

Medical oncologists are usually concerned about reporting detailed information on prognostic associations of genomic markers in cancer. First, the literature is full of inconsistent and even opposing results based on retrospective studies. Second, as

patients have access to the report, bad prognostic associations could lead to misinterpretation and anxiety, emphasizing the idea that this information should be discussed in person taking into consideration additional clinical parameters. Therefore, only prognostic markers with well-established associations in the same tumor type should be reported, preferably without description of the related outcome information. Common diagnostic associations should also be described, mainly those favoring a specific tumor subtype.

4.3 Variants with Biological Relevance

Many variants in well-known cancer genes do not fall into the prior categories but still might be causally associated with the malignant phenotype. Their relevance is justified by known biological implications (pathway activation/inactivation) or by "theoretical" actionability, when agents potentially targeting novel activating mutations in oncogenes or the downstream effects of loss-of-function mutations in tumor suppressor genes are available for clinical testing. Therefore, the expected effect of the variant on protein function (gain- or loss-of-function) is also presented in the report, as it might give insights to the ordering physician with regard to therapeutic interventions in the investigational setting. Nevertheless, until functionality is validated in preclinical studies, it is appropriate to report these novel variants as non-actionable.

4.4 Pertinent Negative Variants

Genes that have clear predictive, prognostic, or diagnostic associations in a specific tumor type and are found to be "wild type" in the NGS test should be described in the report.

4.5 Variants of Unclear Significance

The accelerated pace of advances in our understanding of cancer genomics justifies the description of all "reportable" variants in the final NGS report, even those not classified as actionable or biologically relevant when the assay is performed. These variants may become biomarkers in the near future or may be of particular interest in research settings. The most practical approach to handle variants of unknown biological/clinical significance is to present them according to the main pathway affected by the alteration. Key gene-pathway associations are increasingly being highlighted in the cancer genomics literature [18, 19]. As an example, in renal cell carcinomas, mutations in genes involved in histone modification/chromatin remodeling might dominate a report, warning the medical oncologist-translational researcher about the importance of aberrations in this pathway during cancer progression.

4.6 Germ Line Variants

The American College of Medical Genetics and Genomics (ACMG) recently published a minimum list of genes that should be reported to the patient when an incidental germ line mutation associated with heritable risk of cancer or other diseases is identified and confirmed [20]. The group prioritized disorders where preventive measures and/or treatments were available and those in which individuals with pathogenic mutations might be asymptomatic for long periods of time. Only pathogenic mutations should be reported, considering the challenges of interpreting variants of unknown significance as incidental findings. Notably, the group acknowledged the fact that insufficient data on penetrance and clinical utility support these recommendations. Considerable personnel resources, including genetic counselors with specialized training, may be needed to ensure that patients understand the potential benefits and risks of receiving somatic and germ line data and to support physicians in conveying such information.

4.7 Performance Characteristics of the Test

Specific regions interrogated by the assay and the coverage metrics by sample and target—including median depth, uniformity, and percentage of target covered at the minimum level—should be described in every NGS assay, regardless of application or platform. Minimum depth of coverage should be established during the test validation process and will depend upon the required sensitivity of the assay as well as the targeting/sequencing method. Regions of sequence not meeting the required read depth, especially genes with highest priority (see "pertinent negatives" above), should be clearly reported as indeterminate. Importantly, medical oncologists still need to be educated for the proper interpretation of MAF counts. This information is very useful in the research setting, reflecting clonal evolution and selection when NGS tests are performed in different samples and time points over the course of a disease and therapy. Of note, continued medical education is an important aspect in the process of implementing NGS reports in a clinical lab, so that physicians are trained to understand molecular profile results.

5 Conclusion

NGS tests were initially developed for research or investigational purposes but will eventually become part of cancer care. During the process of clinical implementation of these assays, many technical, legal, and ethical challenges have to be overcome. Clinical Laboratory Improvement Amendment (CLIA) or Good Clinical Laboratory Practice (GCLP) certification is required for clinical centers and

consulting biotechnology companies offering NGS-based cancer diagnostic tests. Several professional societies have generated guidelines for the implementation of NGS tests, with a focus on analytical validity or patient privacy rules. Nonetheless, recommendations for the use of computational tools and bioinformatics pipelines and reporting of somatic cancer variants are still missing. A major challenge is how to convey the amount of data obtained from NGS tests and all the information reviewed for interpretation within a reasonable time frame, so that it can be translated into a useful clinical tool. Effective communication of results with interactive reports can promote appropriate clinical decision-making and minimize the potential for patient harm. Unfortunately, at the present time, validated evidence on specific gene variants linked to predictive, prognostic, or diagnostic associations in cancer is limited. In addition, genomics knowledge is currently ahead of our ability to therapeutically target tumors, given that many mutations identified by sequencing either are linked to unapproved drugs or are not targetable by currently available molecular therapy.

Importantly, while sequencing can identify druggable targets, clinicians are often left with the task of further interpretation, treatment prioritization, and decision-making in the context of additional clinical information. When the best option is to offer the patient genomic-driven clinical trials, additional logistical challenges need to be overcome, including too strict eligibility criteria in phase 1 trials or slots not available at the time of referral and geographical limitations to access drug development units. These difficulties explain why only a small number of patients are ultimately enrolled in a specific trial based on the results of NGS assays, even when actionable genomic alterations are identified in the majority of the tumor samples tested [21]. Multi-institutional trial networks assessing novel agents that target specific mutations are needed in order to deal with these issues. Alternatively, when physicians and patients agree on off-label use of targeted therapies, another aspects that go beyond reimbursement concerns need to be taken into consideration. There is an inherent bias to publish positive results—case reports showing that sequencing results are associated with responses to off-label use of a targeted agent—and mechanisms to annotate lack of response in this setting are missing. One option is to create national formularies of targeted agents against common aberrations, so that every patient receiving a matched therapy in the off-label setting can be tracked and become a "cancer information donor." These pharmacy exchange programs could generate ever-growing data banks integrating the genomic information with therapeutic response and outcome [22]. The information derived from these registries should be added to knowledge databases such as My Cancer Genome or Personalized Cancer Therapy and become readily available to oncologists worldwide, providing annotated predictive genomic markers in cancer and potentially changing the paradigm of drug approval process.

In conclusion, structured reporting of clinically relevant variants may help addressing the current limitations of NGS to directly guide patient care. With standardized terminology and an expanding knowledge database, variant annotation, prioritization, and clinical interpretation become a fluid process with the potential to open new therapeutic options.

References

1. Dienstmann R, Rodon J, Tabernero J. Biomarker-driven patient selection for early clinical trials. Curr Opin Oncol. 2013;25:305–12.
2. Watt S, Jiao W, Brown AM, et al. Clinical genomics information management software linking cancer genome sequence and clinical decisions. Genomics. 2013;102:140–7.
3. Van Allen EM, Wagle N, Levy MA. Clinical analysis and interpretation of cancer genome data. J Clin Oncol. 2013;31:1825–33.
4. Lawrence MS, Stojanov P, Mermel CH, et al. Discovery and saturation analysis of cancer genes across 21 tumour types. Nature. 2014;505:495–501.
5. Cooper GM, Shendure J. Needles in stacks of needles: finding disease-causal variants in a wealth of genomic data. Nat Rev Genet. 2011;12:628–40.
6. Frousios K, Iliopoulos CS, Schlitt T, et al. Predicting the functional consequences of non-synonymous DNA sequence variants – evaluation of bioinformatics tools and development of a consensus strategy. Genomics. 2013;102:223–8.
7. Zhang J, Liu J, Sun J, et al. Identifying driver mutations from sequencing data of heterogeneous tumors in the era of personalized genome sequencing. Brief Bioinform. 2014;15:244–55.
8. Kumar P, Henikoff S, Ng PC. Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm. Nat Protoc. 2009;4:1073–81.
9. Reva B, Antipin Y, Sander C. Predicting the functional impact of protein mutations: application to cancer genomics. Nucleic Acids Res. 2011;39:e118.
10. Adzhubei IA, Schmidt S, Peshkin L, et al. A method and server for predicting damaging missense mutations. Nat Methods. 2010;7:248–9.
11. Flanagan SE, Patch AM, Ellard S. Using SIFT and PolyPhen to predict loss-of-function and gain-of-function mutations. Genet Test Mol Biomarkers. 2010;14:533–7.
12. Wong WC, Kim D, Carter H, et al. CHASM and SNVBox: toolkit for detecting biologically important single nucleotide mutations in cancer. Bioinformatics. 2011;27:2147–8.
13. Gnad F, Baucom A, Mukhyala K, et al. Assessment of computational methods for predicting the effects of missense mutations in human cancers. BMC Genomics. 2013;14:S7.
14. Gonzalez-Perez A, Lopez-Bigas N. Improving the assessment of the outcome of nonsynonymous SNVs with a consensus deleteriousness score, Condel. Am J Hum Genet. 2011;88:440–9.
15. Lubin IM, Caggana M, Constantin C, et al. Ordering molecular genetic tests and reporting results: practices in laboratory and clinical settings. J Mol Diagn. 2008;10:459–68.
16. Wagle N, Berger MF, Davis MJ, et al. High-throughput detection of actionable genomic alterations in clinical tumor samples by targeted, massively parallel sequencing. Cancer Discov. 2012;2:82–93.
17. Dienstmann R, Dong F, Borger D, et al. Standardized decision support in next generation sequencing reports of somatic cancer variants. Mol Oncol. 2014;8:859–73.
18. Garraway LA, Lander ES. Lessons from the cancer genome. Cell. 2013;153:17–37.
19. Vogelstein B, Papadopoulos N, Velculescu VE, et al. Cancer genome landscapes. Science. 2013;339:1546–58.
20. Green RC, Berg JS, Grody WW, et al. ACMG recommendations for reporting of incidental findings in clinical exome and genome sequencing. Genet Med. 2013;15:565–74.
21. Tran B, Brown AM, Bedard PL, et al. Feasibility of real time next generation sequencing of cancer genes linked to drug response: results from a clinical trial. Int J Cancer. 2013;132:1547–55.
22. Schilsky RL. Implementing personalized cancer care. Nat Rev Clin Oncol. 2014;11:432–8.

Clinical Considerations in the Conduct of Cancer Next-Generation Sequencing Testing and Genetic Counseling

Heather Fecteau and Tuya Pal

Abstract Over the last decade, there have been tremendous advances in genetic testing through the development of next-generation sequencing (NGS) technologies. This has led to plummeting costs of testing making it possible to test for multiple genes simultaneously at a cost comparable to testing for 1–2 genes through older Sanger sequencing technology. As a consequence, clinical practice has been greatly impacted resulting in the need to develop new models for genetic counseling and informed consent. This chapter will highlight clinical considerations when using NGS to evaluate for inherited cancer predisposition. Topics to be covered include factors to consider when conducting NGS tests, considerations of various multigene tests available, resulting paradigm shifts, and other clinical and laboratory considerations when testing is conducted. We will conclude with the evolving role of genetics health professionals given the emerging landscape and highlight the importance of education and outreach efforts.

1 Introduction

With the completion of the Human Genome Project in 2003, it was widely acknowledged that more information was needed before the genome could be translated into everyday clinical practice. During this time, DNA sequencing was performed using chain-termination method, now referred to as Sanger sequencing [1]. For over 30 years, Sanger sequencing has been the "gold standard" to accurately obtain long sequence reads (about 200 nucleotides). However, drawbacks to Sanger sequencing included restrictions in scale, turnaround time, and cost of genetic testing.

H. Fecteau (✉)
Department of Clinical Cancer Genetics, Texas Health Presbyterian Hospital, Dallas, TX, USA
e-mail: hsellers83@gmail.com

T. Pal
Department of Cancer Epidemiology and Internal Medicine, Moffitt Cancer Center, Tampa, FL, USA

W. Wu, H. Choudhry (eds.), *Next Generation Sequencing in Cancer Research, Volume 2*, DOI 10.1007/978-3-319-15811-2_6

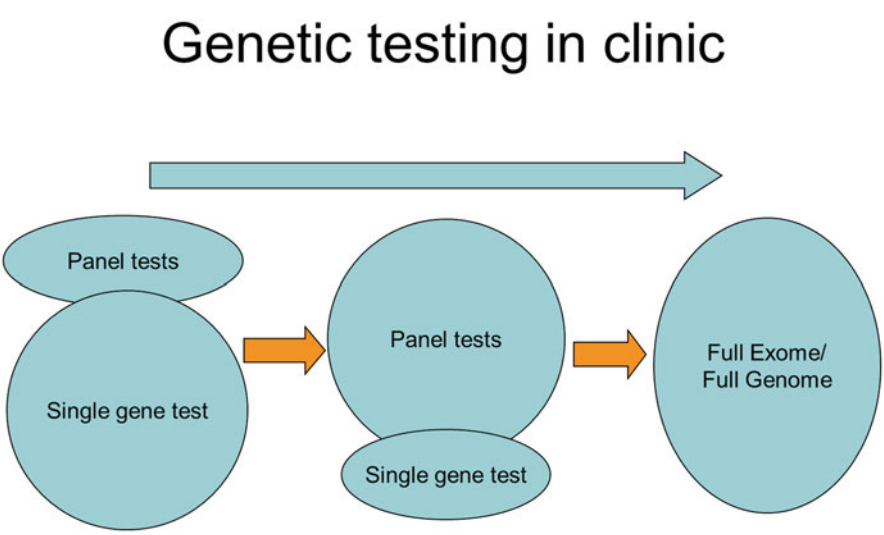


Fig. 1 The evolving paradigm of genetic testing for inherited cancer

Next-Generation Sequencing (NGS) or massively parallel sequencing technology was first described in 2000 [2]. NGS results from running multiple reactions simultaneously to generate large quantities for sequence data in parallel [3–6]. Along with this technology, sequencers were developed that could run these reactions on a larger scale. It was not until 2008 that this technology was documented as having successfully sequenced a complete human genome [7]. Since then, this technology has revolutionized clinical genetics as costs of sequencing have plummeted. Since the introduction of NGS in 2008, the National Human Genome Research Institute's (NHGRI) analysis shows the cost of sequencing one whole human genome was reduced from almost $10,000,000 to less than $10,000 [8]. Costs are expected to drop below $1,000 and take just days to complete; recognizing this does not take into account the time and cost for data interpretation of the results to the patient.

There are a plethora of NGS gene panels available for phenotypically targeting testing (e.g., deafness, cardiomyopathies, cancer) covering a handful to >100 genes and ranging in price from $1,500 to $10,000 ([9]; www.ncbi.nlm.nih.gov/gtr/; www.genetests.org). In addition, several laboratories offer clinical WGS/WES ranging from $4,500 to $10,000 [10]. Compared to the average timeframe and cost of $1,000–2,000 for single gene Sanger sequencing, NGS has shifted the genetic testing paradigm (Fig. 1). It is predicted "to become a central piece of routine healthcare management which can be practiced regularly by physicians from their offices" [11].

It is not clear how the genetic information provided by NGS should be integrated into current clinical practice. However, new models for providing genetic counseling and informed consent will clearly need to be developed and evaluated. The purpose of this chapter is to highlight issues to consider when utilizing NGS in a clinical cancer setting and to propose approaches to counseling patients about NGS genetic testing.

2 Factors to Consider with Next-Generation Sequencing Testing

Clinical applications of NGS include multigene panels, whole-genome sequencing (WGS), and whole-exome sequencing (WES). All NGS technology utilizes the human genome reference sequence as a comparison in order to identify DNA variation in a sample. It is important to keep in mind that the reference genome is incomplete and inaccurate for some regions. The accuracy of NGS is usually measured by sequencing depth; this refers to the average number of times that a specific base/nucleotide is sequenced. The greater the number of times the genome is sequenced, the greater the sequencing depth and the more accurate the individual base calls. WGS using NGS can routinely call base changes with greater than 99.9 % sensitivity and specificity at a depth of 30-fold and greater than 95 % of genome is covered at an average sequencing depth of 30-fold [12].

Currently, NGS may not consistently identify variations larger than a few base pairs in size like insertions and deletions, trinucleotide repeats, and copy number variations (CNVs) across all testing platforms. As a result, most labs supplement NGS test with other techniques, like Multiplex Ligation-dependent Probe Amplification (MLPA) or array-based tests to provide evaluation of larger, structural genomic changes. Perhaps over time as the NGS technology and bioinformatics advances, these limitations will be minimized and potentially overcome.

NGS analyzes a large number of genes, thus a large number of amino acid sequence changes ("missense" variants) called variants of uncertain significance (VUS) may be identified. For many of these genes, there are no means by which to determine whether a particular amino acid change impairs the function of the resulting protein. Although there are a number of computational tools to predict pathogenicity, the clinical validity remains uncertain without a direct functional assay and familial segregation data [13–15]. The likelihood of detecting a VUS is directly related to the number of genes tested, thus multigene testing results in a higher VUS rate given that multiple genes are tested for simultaneously. WGS can generate 3–4 million variants that differ from the human reference sequence, while WES generates 15,000–20,000 variants within the coding region [16]. Multigene panels have reported to average around 2.1 VUS per sample [17]. Given the increased likelihood for variants, the clinical challenge of accurately and efficiently interpreting the significance of the VUSs should be taken into consideration when using NGS.

It is challenging for genetic counselors and other members of the healthcare team to consistently advise patients on appropriate medical management following the detection of a VUS and this may add to patient distress [18]. VUS results add an additional layer of complexity to the conduct of cancer genetic risk assessment, as the result should be interpreted in the context of the family history and additional information available on the VUS.

Furthermore, the reporting of VUS results is not standardized between different genetic testing laboratories, thus familiarity and understanding of one laboratories classification does not translate into the same interpretation at another laboratory, which may greatly impact clinical utility. Laboratories should classify variants

according to the American College of Medical Genetics (ACMG) guidelines and document supporting evidence regarding each variant's known or possible role in disease [19]. Clinicians must have an understanding of methodology behind VUS classification, where to obtain more information about a VUS, and how to utilize the information to better guide the management of their patients.

3 Cancer Multigene Panel Testing

Next-generation sequencing can address the growing number of cancer susceptibility genes with overlapping phenotypes with potential time and cost savings with gene panel testing. Gene panels may improve the detection rate of hereditary cancer syndromes. They may also expand the range of phenotypes associated with mutations in various genes and contribute to the understanding of the natural history of hereditary cancer syndromes. The traditional approach to genetic testing has involved analyzing a single gene or a few genes related to a single syndrome based on the pattern of cancers observed in a family. However, this method may have led to underrecognition of patients with mutations given 30 and 50 % of individuals with a mutation do not have a family history significant enough to warrant genetic testing [20]. Gene panels allow for concurrent analysis of genes in which mutations confer variable levels of cancer risk and variable tumor spectrums, thus attending to syndromes with overlapping phenotypes and also addressing the limits of an uninformative family history.

Cancer panels can include genes of high, moderate, or unknown cancer risks as summarized in Table 1 [21]. High-penetrance genes are those genes that, when mutated, confer high cancer risks, with published managements guidelines for those with mutations. Moderate penetrance genes are genes that when mutated, confer moderate cancer risk, with no management guidelines for those with mutations. The last category for cancer genes included on panels is unknown penetrance; genes that when mutated are known to be prevalent within a certain cancer patient population; however, the degree of cancer risk and tumor spectrum are not well understood, and they have no management guidelines for those with mutations.

As of August 2014, there are nine laboratories offering NGS cancer susceptibility gene panels. Each laboratory has a different approach as to the number of panels it offers and/or genes included on each panel. The panels offered fall into three categories: (1) cancer-specific high-penetrance gene panel; (2) cancer-specific gene panel with high, moderate, and unknown penetrance genes; and (3) "comprehensive" cancer panels that include genes associated with multiple cancers or hereditary cancer syndromes [22]. This personalized approach using gene panels can provide a more objective risk and can parse out who is at risk for highly penetrant cancer syndrome, who is at moderate risk due to lower penetrance genes or multifactorial inheritance, and who is at average population risk [23, 24].

When addressing cancer genetic panel testing, the first challenge comes with defining the appropriate patients for this testing. There are no clear guidelines on when to order NGS cancer panels. The National Comprehensive Cancer Network

Table 1 Three categories of genes found on next-generation sequencing cancer panels

	Syndrome penetrance (cancer risk)	Understanding of phenotype	Management guidelines	Examples of genes
Mutations found in category 1	High	Good to excellent	Published guidelines likely to exist, screening or prevention for many of associated cancer risks exist, mutation likely to change management	*APC*, *BMPR1A*, *BRCA1*, *BRCA2CDH1*, *EPCAM*, *MLH1*, *MSH2*, *MSH6*, *MUTYH*, *PMS2*, *PTEN*, *SMAD4*, *STK11*, *TP53*
Mutations found in category 2	Moderate	Fair to good	Guidelines unlikely to be published, screening may exist for associated cancer risks, mutation may or may not change management	*ATM*, *CHEK2*, *PALB2*
Mutations found in category 3	Unknown	Poor	Guidelines do not exist, difficult to make recommendations and therefore unlikely to change management	*BARD1*, *BRIP1*, *MRE11*, *NBN*, *NBS1*, *RAD50*, *RAD51C*, *RAD51D*

Table 2 ACMG indications for diagnostic testing using next-generation sequencing

WGS/WES should be considered in the clinical diagnostic assessment of a phenotypically affected individual when:
• The phenotype or family history data strongly implicate a genetic etiology, but the phenotype does not correspond with a specific disorder for which a genetic test targeting a specific gene is available on a clinical basis
• A patient presents with a defined genetic disorder that demonstrates a high degree of genetic heterogeneity, making WES or WGS analysis of multiple genes simultaneously a more practical approach
• A patient presents with a likely genetic disorder but specific genetic tests available for that phenotype have failed to arrive at a diagnosis

(NCCN) addressed the use of gene panels in their 2014 Guidelines for Risk Assessment [25]. The authors of the NCCN guidelines indicated that cancer gene panels could be considered after highly penetrant syndromes have been ruled out and there is still reason to believe the family history is suggestive of a hereditary cancer syndrome. Genetic counselors and health professionals can use these early guidelines to determine when to consider counseling for NGS cancer panels. The American College of Medical Genetics (ACMG) has developed a position statement for whole-exome and whole-genome sequencing ("Points to consider in the clinical application of genomic sequencing," 2012) that can be adapted to apply to NGS cancer panels and be used by genetic counselors to guide their cancer risk assessments (Table 2). Most pediatric genetic panel testing is guided by this ACMG statement.

Table 3 Possible clinical scenarios to consider offering gene panel and/or WES/WGS genetic testing

• Individual with multiple cancer diagnosis
• Personal and/or family history of cancer meets national criteria for more than one hereditary cancer syndrome
• Second-line workup for inherited cancer risk when first-line evaluation has been noninformative
• Family history of cancer does not meet established testing guidelines due to limited or unknown family history

The American Society of Clinical Oncology (ASCO) recently updated their recommendations on genetic testing for cancer susceptibility in response to the advancements in genetic testing technology. Initially ASCO recommended that clinical genetic testing only be offered to those with a personal or family history suggestive of an inherited cancer syndrome. ASCO has since updated this recommendation indicating that individuals without a family history may be appropriate candidates for cancer susceptibility testing if analytic and clinical utility has been established, meaning the results can be adequately interpreted, and can impact medical decision making and clinical outcomes [26]. Given this recommendation, gene panel testing could be offered to a wider patient population who do not meet the standard testing criteria [27]. It should be noted that of the few published studies looking at gene panel testing in a clinical setting, they all report most of the patients testing positive for a genetic mutation either had cancer or had a significant family history [17, 28–30]. Gene panels should be considered as a testing strategy when there is a particularly complicated personal or family history, a suspicion of multiple cancer syndromes, or other clinical scenarios described in Table 3 [22].

Cancer panel testing may identify mutations in hereditary cancer genes that are both expected and unexpected by the personal and family history [22]. Research studies using panel testing have found mutations in genes that do not clearly match the family pedigree; this suggests that the current understanding of the cancer genotype–phenotype may still be incomplete since the classic style of genetic testing selected only those meeting high-risk criteria for a particular syndrome [31–33].

Thus, the interpretation of these incidental findings in family cancer risk assessment and management is evolving. While some gene alterations may have a substantial impact on cancer risk recommendations, other mutations may be more difficult to interpret clinically because of a lack of correlation with family history (e.g., a *BRCA1* mutation in a family with hereditary colon cancer) or a lack of evidence-based recommendations for management (e.g., *RAD50* mutation). Medical management guidelines do not exist for many of the genes tested and the appropriate clinical response remains unclear. In some cases, appropriate medical management will be based on a patient's personal and family history more so than genetic test results. Another option is to extrapolate risk reduction strategies from more extensively studied genes (e.g., *BRCA1/2*) that impart cancer risk [34].

Data regarding cancer risks may not be available for all genes being tested, and risk estimates may be especially difficult for patients who carry variants and/or mutations in multiple genes. A prime example is the *PALB2* gene; it has been understood that *PALB2* is associated with a moderate risk for breast cancer but the exact breast cancer risk has been not well understood. A recent study examining the breast cancer risk in families with a *PALB2* mutation found a breast cancer risk eight to nine times greater among women younger than 40 with a *PALB2* mutation compared to the general population [35].

A final factor to consider when utilizing cancer gene panels is the higher rates of VUS. VUS can be challenging clinically for several reasons, including that many patients and providers make the mistake of assuming that a VUS is responsible for disease risk in a family leading to misguided risk-reducing medical management. As discussed earlier, while it does take a great deal of time and resources, many VUS are reclassified as benign.

However, genetic panel test results may still be beneficial for excluding a diagnosis (in the case of a negative result) or allowing targeted testing for family members (in the case of a positive result). It may be difficult to get the cost of family members testing covered for mutations in moderate-penetrant or unknown-penetrant genes. It is possible that more information will be discovered about the phenotype and cancer risks related to each syndrome as more patients are tested and a larger pool of patients with hereditary cancer syndromes are identified. In much the same way that testing criteria and medical management guidelines have evolved for families at high risk for hereditary breast and ovarian cancer syndrome, it is plausible that management guidelines for cancer syndromes with incomplete penetrance will be developed in the future.

4 Exome and Genome Testing in a Cancer Setting

Further adding complexity to the genetic testing landscape is the concept of exome (i.e., the protein coding regions of the gene) and whole-genome testing. In the future, it is anticipated that multigene tests may be replaced by whole-exome or whole-genome sequencing, as sequencing costs continue to decrease.

In anticipation of these tremendous technologic advances, the American College of Medical Genetics (ACMG) recently issued guidelines pertaining to a minimal list of actionable genes (i.e., 56 genes related to roughly 25 genetic conditions) for which testing should be reported when performing exome or genome sequencing [36]. These constitute conditions that may be unrelated to the indication for ordering the sequencing, but of medical value for patient care (thus referred to as "incidental findings"). These conditions, determined by the ACMG to be well recognized and known to have a strong link of causation, were included on this list if preventative measures and treatments exist. Groups of conditions included on this list encompass cancer predisposing conditions, later-onset cardiac-related syndromes, and connective tissue syndromes.

The initial guidelines were revised in 2014 to recommend an opt out clause for incidental findings [37]. Furthermore, the ACMG guidelines recommended that seeking and reporting incidental findings not be limited by the age of the person being sequenced. It is important to consider the ACMG incidental findings guidelines in the context of ACMG guidelines pertaining to testing in children which indicate that predictive genetic testing of minors be considered only if effective medical interventions are available to treat, prevent, or retard the course of disease [38]. These guidelines are not contradictory, because incidental findings, by definition, are outside of the indication for which testing was done in contrast to specifically testing a child for an adult-onset condition.

In addition to the debate surrounding return of incidental findings from germline exome and whole-genome testing, there remain questions surrounding return of results in the setting of tumor-focused testing. Interestingly, the ACMG guidelines stated that "incidental variants should be reported for the normal sample of a tumor-normal sequenced dyad." It is important to note that this guideline could have significant implications for the field of oncology [39]. Given that the vast majority of clinical sequencing tests ordered in the oncology setting are tumor exome or genome sequencing to identify somatic mutations to guide treatment decisions, germline results are not directly related to testing indication. Consequently, these guidelines have profound implications pertaining to initial and follow-up discussions between patients and their oncologists. Specifically, a clear discussion between the oncologist and patient about the potential to include germline analysis as part of the tumor test would be required, which would include covering germline-related issues such as risks and benefits of testing, risks to family member, as well as factors related to privacy and insurability. In fact, a recent study reported on the implementation of a whole-genome sequencing protocol of tumors and paired germline DNA, which included options for receiving incidental germline findings [40]. In this study, genetic counselors documented patient family histories, secured informed consent, and actively participated in the multidisciplinary tumor board to provide clinical context of germline results and recommendations for results disclosure. This study serves to highlight the future opportunities for genetic professional involvement in these types of efforts as use of whole-genome sequencing in oncology treatment broadens.

5 Impact on Paradigm Shift from Syndrome-Based to Multigene Testing

With the availability of new testing options, there will also be changes in the delivery of genetic risk assessment services. Traditionally, cancer genetic counseling has evaluated a patient's risk based on personal and family history of cancer, age of diagnosis, and other phenotypic features. Both the ACMG and NCCN recommend that genetic counseling should be performed by a cancer genetic professional [19, 25].

Genetic counselors and professionals have used their expert knowledge to choose which genes to test and then counseled the patient about the cancer risks and management options for mutations in those specific genes. Genes that are unlikely to be mutated are not analyzed in this model. However, plummeting costs of testing are resulting in many genes being tested simultaneously (either through panels of genes focused on a particular cancer type or WGS/WES sequencing).

As a result, the need to generate an extensive differential diagnosis and eliminate possible diagnoses using a stepwise genetic testing approach is lessening and clinical practice paradigms appear to be shifting toward a model where a patient is "tested first" (without the need to generate an extensive differential diagnosis based on clinical information) following broad consent. Once results are available, additional information is collected to put the diagnosis into proper clinical context. Many patients may not be adequately prepared for the possible outcomes and/or their perceived understanding or expectations may not align with the actual results [27]. This may result in a great need and time for posttest genetic counseling than pretest counseling [41].

It is worth considering that although broad testing without the need to generate a differential diagnosis may make it easier to order comprehensive testing, it will still require proficiency in genetics due to required familiarity with the various gene panel and WGS/WES options, choosing the one best suited for each patient, result interpretation, putting the result in proper clinical context, and making appropriate management recommendations. As such, it is anticipated that provision of care based on genetic testing results will become exponentially more complex resulting in an increased need for the involvement of genetic counselors and professionals in patient care, an issue already recognized as part of several best practices guidelines from numerous professional guidelines [25, 26, 42–45].

6 Importance of Informed Consent

When genetic counseling for highly penetrant cancer syndromes was first performed, there were concerns about the lack of knowledge of the cancer risks associated with each syndrome, what early detection and/or risk-reducing options would be available for patients with mutations, and whether patients would experience significant anxiety upon learning they carried a mutation. As an increasing number of individuals with hereditary cancer syndromes were identified, the knowledge of highly penetrant cancer syndromes increased, improving the ability to create effective clinical guidelines for management.

Studies have shown that individuals receiving mutation-positive results describe an increase in anxiety, but that anxiety often returns to baseline with the passage of time [46, 47]. Organizations like the NCCN, American Society of Clinical Oncology, and the U.S. Preventive Services Task Force have acknowledged the research that shows the benefits of genetic counseling and testing for hereditary

cancer syndromes; they have written guidelines and recommendations for cancer predisposition testing, all of which include pretest counseling as part of the informed consent process [25, 26, 48].

While existing genetic counseling models encourage in-depth discussion of the syndrome to be tested, these models do not address testing multiple syndromes, simultaneously [49]. Communicating the risks for NGS testing that is usually conveyed with single gene testing would likely lead to information overload, in which there is too much information to absorb in a short time, potentially impeding patient understanding and decision-making ability [22, 41, 49].

The pretest genetic counseling model will need to involve a discussion of the range of information that could be learned from NGS genetic testing including risks, benefits, and limitations of testing and implications for both the patient and family members, such as the increased risk of discovering unanticipated results and VUSs [21, 49]. Along with the progress of genetic testing technology, genetic counseling will also have to shift and adapt to ensure patients are educated about the unique benefits and risks of NGS genetic panel testing in order to facilitate informed consent.

7 Suggested Genetic Counseling Approaches to Next-Generation Sequencing Tests

The paradigm shift in genetic testing practices will lead to changes in the approach to genetic counseling of patients, recognizing that the optimal approach is currently unknown [49]. Many of the genes on cancer panels and WGS/WES testing confer a risk for multiple different cancers. Most patients seeking genetic testing primarily based on risk for more common heritable adult malignancies (breast, colon), uncovering additional cancer risks may be unanticipated outcome of the testing and should be discussed pretest [27].

Patients should be informed of the option of single-gene, syndrome-specific testing, or WGS/WES and which may more quickly identify actionable mutations, especially when pending a treatment decision [22]. For people of childbearing age, genes that have distinct monoallelic and biallellic expression should be covered in regards to risk of having a child with a more severe autosomal recessive cancer syndrome [50].

While adapting the amount of information shared with the patient, it is important to maintain patient autonomy and the ability to make an informed decision. A suggestion to help present this information in a timely and effective manner is to group the genes into the aforementioned three categories (Table 1) [21]. This technique could help patients understand that mutations in different genes are associated with different levels of risks for cancer and not all results have clear management guidelines.

It may also be helpful to group the cancers associated with each panel test. The genetic professional could then broadly describe how the increased cancer risk for each organ may/could be managed. For example, some genes on the breast panels would put a patient at risk for breast and pancreatic cancer; the genetic counselor would explain increased breast cancer surveillance options and then explain the limited screening options for pancreatic cancer. Patients should know a deleterious mutation could mean a risk for multiple sites of cancer and understand the degree to which surveillance and management strategies exist and are efficacious for each site of cancer.

One approach adopted by the Genetic Risk Assessment Service at the Moffitt Cancer Center includes a pretest genetic counseling session during which the following is discussed: (1) a brief overview of multiple syndromes in general terms with discussion of specific conditions for which the patient may be at risk based on personal and/or family history, (2) discussion of high penetrance ("actionable") versus moderate penetrance ("not likely actionable") genes, and (3) communication of higher rates of variants of uncertain significance (VUS) [51]. A detailed discussion of specific conditions is deferred to the posttest session, during the disclosure of genetic test results.

Another suggested genetic counseling approach utilized at Dana-Farber Cancer Institute's Center for Cancer Genetics and Prevention is to present information in a framework linking function and phenotype in the pretest session. They encourage particular attention on the education of moderate-penetrance genes, the risk for variants of uncertain significance, and emphasis on genes most likely to be mutated given the family history reported [41]. They also recommend focusing on high penetrance genes associated with a severe phenotype where they defined risk-reducing strategies in order to reduce possible distress in the event an unexpected mutation is found [41] For example, if a patient were to test positive for a CDH1 mutation and the family history is negative for breast and/or gastric cancer, it could be called into question the appropriateness of a gastrectomy [22]. Posttest counseling is recommended for all patients found to carry a mutation, a VUS, or for those who test negative and have a striking cancer family history in order to ensure proper interpretation of results [41].

As noted earlier, it is important that patients understand the chance of a VUS result and the limitations of such results. Genetic counselors should consider sharing the VUS rate reported by the elected laboratory when considering NGS panels or WGS/WES testing. Patients should understand that VUSs will not be treated as deleterious nor causative of a cancer predisposition. Part of the posttest counseling session should cover expectation and plans for recontact should be discussed and patient should be encouraged to periodically check in and update contact information [22]. This is particularly important to VUS reclassification as many labs review their VUS data on a regular basis and will release updated results.

There currently remains a tremendous need to develop and refine new genetic counseling strategies to deliver genetic testing services to manage population needs particularly as use of genomic testing technologies continues to increase.

8 Documenting Genetic Testing

There are multiple NGS cancer panels with varying sets of genes, and more genes may be added to these panels as our knowledge about cancer susceptibility improves. While many genetic professionals document the type of genetic testing ordered, it will become more important to document which genes were tested for each patient and which lab was used [21]. It will also be helpful to document the testing platform, depth of coverage, and presence of a deletion/duplication assay. As part of posttest genetic counseling, genetic counselors should continue to inform patients that updated testing may be available for them in the future. The protocol for patients to be notified of such updates (e.g., who has the responsibility to follow up to discuss advances in testing options) should be clear.

9 Laboratory Considerations

Organizations that have authority to regulate genetic testing include the U.S. Food and Drug Administration (FDA) and the Centers for Medicare and Medicaid Services through the Clinical Laboratory Improvement Amendments (CLIA) [52, 53]. A genetic test may be developed as a "test kit" or a "home brew." Test kits are prepackaged with reagents and instructions and sold to laboratories, whereas home brews are assembled in house by the laboratory. Test kits are regulated by the FDA as medical devices, thus manufacturers must submit data on analytic and clinical validity and utility to the FDA for approval prior to marketing. Consequently, it is no surprise that the FDA has only approved four test kits to detect mutations in human DNA, of the hundreds of diseases for which genetic tests are currently available clinically [52]. In contrast, home brews are under CLIA oversight, which requires laboratories to perform proficiency testing themselves to demonstrate their ability to accurately perform the test and interpret the results but they do not need to demonstrate clinical validity or utility. Thus, under CLIA, the decision to offer a new genetic test is within the sole discretion of each clinical laboratory director. As a result, most genetic testing is currently overseen by CLIA rather than the FDA, which illustrates that manufacturers prefer the less regulated status and that the regulatory regime allows them to avoid stringent FDA oversight. Ultimately, there are clear opportunities to develop a regulatory system to ensure that patients and providers receive greater assurance that genetic tests are accurate and reliable and provide information that they are relevant to healthcare decision making. At the present time, mutation detection strategies and detection rates for a given gene may vary by testing laboratory due to the techniques used, the patient's mutation may be detected by one laboratory but not another. Consequently, practitioners who provide genetic testing services require familiarity with laboratory testing

approaches, as patients rely on them to research and select the laboratory best suited for their genetic needs.

Unlike many tests used in medicine, for many years there has been lack of FDA oversight for genetic testing, thus test validity and clinical utility may differ substantially between labs. Therefore, there has been heightened importance to understand variations in laboratory practices and the meaning of terms, such as analytical sensitivity, reported range, coverage, and variant filtering, when determining whether to perform a disease-targeted gene panel, exome, or genome analysis and which laboratory to utilize. Not only is the quality of the result received be impacted by these factors, but also on the ability to interpret the result itself. This is particularly true for conditions, such as those associated with moderate penetrance genes, where national best practices guidelines do not currently exist due to paucity of data.

Recent developments suggest that the FDA is planning to increase its oversight of genetic testing. In November 2013, the FDA demanded that 23andMe immediately stop selling and marketing its DNA testing service until it receives clearance from the agency. This was a direct to consumer test sold through the company's website through which saliva samples are analyzed to give clients information on risks of developing certain diseases. Subsequently, the FDA outlined plans to regulate thousands of diagnostic tests, including genetic tests. This new policy is likely to have a big impact on the increasingly common practice of using genetics to decide how to treat cancer patients [54].

In addition to the increased regulation anticipated for genetic testing through FDA oversight, there remain several factors to consider when choosing a laboratory for genetic testing. Furthermore, although these tests did not initially include the *BRCA* genes given that a single U.S. lab held and enforced their gene patent, precluding other labs from offering clinical testing. This all changed in June 2013 following the Supreme Court decision that genes cannot be patented, which has resulted in an increasing number of labs offering *BRCA* testing. Consequently, the cost of the *BRCA* test has substantially decreased. For example, prior to the loss of the patent, the list price of complete *BRCA* testing was over $4,000; in contrast, since the fall of the patent, the cost has plummeted to as low as $1,500 through a lab that offers testing for 211 genes including *BRCA*. As a result, navigating through the various testing options has become increasingly complicated for healthcare providers as they must now evaluate various factors when choosing the appropriate lab such as: (1) completeness of the testing; (2) quality of interpretation of complicated results such as variants of uncertain significance (VUS) and openness in sharing how this is done with clinicians; (3) genes included on multigene tests which fit the patient's needs best; (4) practices regarding sharing of deidentified data in public databases to enhance interpretation of genetic data worldwide rather than maintaining it internally to protect commercial interests; and (5) testing laboratory billing practices whereby health insurers are billed a much higher amount than the published list price of the test (Table 4).

Table 4 Factors to consider when choosing a laboratory for next-generation sequencing

What technology is used	• Platform of the testing
	• Depth of coverage
	• Presence of deletion/duplication assay
Which genes are included	• Number of genes (larger panels may not be of any more benefit to the patient)
	• A cancer site-specific test (e.g., breast cancer susceptibility) versus a pan-cancer test (all cancer susceptibility)
	• The proportion of genes that are considered "medically actionable," meaning mutated genes will lead to a change in medical management that is supported by guidelines
	• Option to exclude results per patient request
What is the cost of testing/ insurance coverage	• List price
	• Billing options (e.g., insurance vs. institutional billing)
	• "In network" or "out of network"
	• Medicare or medicaid billing options
	• Financial assistance or payment plans for uninsured patients
	• Presence of a patient "cap" to control patient expense
	• Requirements for letters of medical necessity
What is the turn around time (TAT)	• TAT for insurance preauthorization (if offered) and testing
	• TAT for panels vs. single genes
	• Importance of TAT may be dictated by whether or not a patient is using the information to make an immediate management decision (e.g., surgery for recent diagnosis)
Variants of unknown significance (VUS) rate	• VUS rate for the panel under consideration
	• How conservative is the laboratory in calling out a mutation versus VUS versus benign polymorphism
	• VUS reclassification process (Does the laboratory offer free VUS testing to affected family members? How are ordering providers notified when reclassifications occur?)
	• Supplementary data provided by the laboratory regarding the variant (e.g., cosegregation data, data from in silico models, population frequency, review of the literature, etc.)
	• Patient's level of anxiety about a VUS result (which may dictate the importance you place on a laboratory's VUS rate)
How reliable is the laboratory	• Past experience with the laboratory for other cancer susceptibility genetic testing
	• Laboratory's experience with NGS technology
	• Laboratory's experience with the gene(s) of interest (e.g., lab may be able to better classify missense mutations, etc.)
	• Accuracy of result interpretation
Ease of laboratory use	• Insurance preverification process
	• Reliable communication
	• Sample submission process (workload to order a test)
	• Readability of test report
	• Availability and reliability of online reporting system
	• Access to genetics professionals

10 Evolving Role of Genetics Professionals

As widely acknowledged, healthcare provider knowledge and training are insufficient to make optimal use of genetic testing services despite the general agreement that genetics competency is of high clinical relevance [55, 56]. In fact, a recent Florida-wide survey of healthcare providers who order *BRCA* testing indicated both the need for and an interest in ongoing educational opportunities and resources among community providers who order genetic testing [57].

Within the United States, despite efforts to expand community-based best practices for provision of genetic counseling and testing services, market forces are compelling an increasing number of clinicians with limited training or experience in genetic risk assessment to order and interpret genetic tests [43, 58–61]. One of the most commonly cited reasons for encouraging genetic testing without the involvement of a genetics health professional is the perceived "severe shortage" of these professionals [62]. However, while historically this perception may have been accurate, the recent survey conducted by the National Society of Genetic Counselors demonstrated that access to certified genetic counselors (CGC) is excellent and in line with physicians [63]. Moreover, in-person consultations are now supplemented with telegenetic services, particularly for patients in rural and underserved areas [64]. Furthermore, there has been tremendous growth with a 75 % increase since 2006 and 4,000 CGCs currently, with an expected annual growth rate of approximately 10 % [65].

11 Genetic Professional Issues

While the number of CGCs continues to increase, there remains a gap in reimbursement for services rendered these masters-trained healthcare providers can often not bill insurers independently for services rendered. As the reimbursement scheme in the US is primarily focused on a fee for service model, most Cancer Genetic Risk Assessment Services cost more to administer than the direct revenue they generate [66]. This recognition of the lack of reimbursement for genetic services delivered by CGCs coupled with data to suggest that provision of genetic counseling through a trained genetics professional can lead to increased cost effectiveness and enhance quality of care [67–72] is beginning to influence policy shifts at the state and payer level [50, 73].

12 Importance of Collaborative Research

Ultimately, given the limited proficiency in genetics among the U.S. healthcare workforce, there remain tremendous opportunities for genetics professionals to serve as a hub of information. This is particularly important with the tremendous advances in

genetic testing technology through which multigene testing has become a feasible and widespread option [17, 29, 74]. Innovative approaches to delivering genetics services to an increasing number of patients in community settings have been demonstrated through establishing academic–community partnerships that focus on collaboration between genetics and nongenetics providers to offer genetic testing for hereditary cancers [58, 75, 76]. These collaborative partnerships leverage the expertise of genetics professionals for challenging cases that enable patients to remain in their community and to allow for better access to resources for long-term follow-up care.

Another example of this type of partnership is the Florida-based project (called the Inherited Cancer Research (ICARE) Initiative) for which external peer-reviewed funding was secured in 2010 to develop an infrastructure to support research, education, and outreach initiatives focused on genetic counseling and testing for inherited cancer predisposition. Recognizing the limited number of genetics professionals across Florida [50], a statewide network of over 100 healthcare providers who offer genetic services was developed. These individuals are offered education and outreach about inherited cancer predisposition with the overarching goal of enhancing the provision of genetic services across the state and beyond. In addition to educational and outreach efforts, ICARE Partners refer high-risk patients to the research registry to provide the research link, which has in turn contributed to the tremendous growth of the registry since initiation of the grant in summer 2010 with almost 1,400 high-risk individuals recruited to date, including almost 900 *BRCA* carriers.

Specific educational resources available to ICARE Partners include access to:

1. *Bimonthly Case Conferences*: 1 h web-based teleconferences during which brief educational updates are provided during the first 15 min, after which 3–4 clinical cases are presented, including reason for referral, review of the pedigree including differential diagnosis, risk assessment, testing options, and management plan. Each case includes discussion items and a take-home message.
2. *Inherited Cancer Registry newsletter*: a biannual four page newsletter which briefly outlines recent clinical and research updates pertaining to risk assessment, testing options, and management of those with inherited cancer predisposition. Also included within the newsletter is a section on statewide clinical trials for those with inherited cancer, as personalized treatments based on germline mutations are often only available at a small number of study sites. The newsletter is a means by which updated information is disseminated to healthcare providers and patients who participate in the research registry (newsletters are available at the ICARE website, which can be accessed through the following link: http://inheritedcancer.net).
3. *Access to ICARE-based experts for inquiries*: A dedicated telephone line and e-mail address have been established to provide centralized access to healthcare providers requesting information about Florida genetic services. This infrastructure has facilitated access for providers across the state to seek input from genetic professionals, when faced with complicated patients. This service is provided through a board-certified GC who is specifically available to give a description of resources available through Florida genetic services' efforts and give general guidance pertaining to inherited cancer predisposition to ICARE partners.

Another issue which has become increasingly important is the interpretation of genetic tests. As more data become available, the ability to interpret test results also increases, highlighting the importance for: (1) encouraging patient participation in research registries, (2) development of international consortia to increase sample sizes through pooling data [15, 77–79], and (3) the importance of submission of data to public databases [77, 80].

13 Conclusion

While NGS-based technology is available and use of this technology is increasing, the understanding of how best to counsel patients for whom we recommend this testing is still evolving. It is essential that future research focuses on the outcomes of using this technology, with hope to limit the potential for harm and to maximize the benefit to the patient [48]. This was the approach used to develop counseling models for highly penetrant cancer syndromes such as hereditary breast and ovarian cancer syndrome and Lynch syndrome.

As clinicians are faced with the decision of a single gene/syndrome test (e.g., BRCA1/BRCA2 test) versus a cancer panel test (e.g., breast and/or ovarian cancer panel), or WGS/WES genetic testing there are multiple factors that need to be considered. For example, NGS genetic tests may lead to an improved detection rate for the causative gene mutation; however, depending on the finding, there may not be sufficient data in the medical literature to guide the clinician on how to medically manage that patient. Collaborative epidemiologic work will also be necessary to gather information about genes included in the NGS panels; this will help provide more substantial information about each gene's associated tumor spectrums and cancer risks, which will lead to the development of appropriate clinical management [81].

In addition, the improved detection rate of a cancer panel or cancer WGS/WES testing should be weighed against a higher risk to find a variant of unknown significance (VUS). Lastly, each laboratories approach to a panel and WGS/WES testing may differ, and therefore the ordering clinician will have several factors to consider when choosing between tests/laboratories (Table 1). There is not enough published literature to establish guidelines regarding which patients are best suited for a single gene/syndrome test versus a cancer panel test versus WGS/WES testing. Until such guidelines are established, all the factors should be considered when presenting a patient with genetic testing options and choosing which test to recommend.

References

1. Sanger F, Nicklen S, et al. DNA sequencing with chain-terminating inhibitors. Proc Natl Acad Sci U S A. 1977;74(12):5463–7.
2. Brenner S, Johnson M, et al. Gene expression analysis by massively parallel signature sequencing (MPSS) on microbead arrays. Nat Biotechnol. 2000;18(6):630–4.

3. Bentley DR, et al. Accurate whole human genome sequencing using reversible terminator chemistry. Nature. 2008;456(7218):53–9.
4. Eid J, et al. Real-time DNA sequencing from single polymerase molecules. Science. 2009;323(5910):133–8.
5. Shendure J, et al. Accurate multiplex polony sequencing of an evolved bacterial genome. Science. 2005;309(5741):1728–32.
6. Stoddart D, et al. Single-nucleotide discrimination in immobilized DNA oligonucleotides with a biological nanopore. Proc Natl Acad Sci U S A. 2009;106(19):7702–7.
7. Wheeler DA, et al. The complete genome of an individual by massively parallel DNA sequencing. Nature. 2008;452(7189):872–6.
8. Wetterstrand KA. DNA sequencing costs: data from the NHGRI genome sequencing program (GSP), 2013. www.genome.gov/sequencingcosts. Accessed 20 Jul 2014.
9. Rehm H. Disease-targerted sequencing: a cornerstone in the clinic. Nat Rev Genet. 2013;14(4):295–300.
10. Jamal SM, et al. Practices and policies of clinical exome sequencing providers: analysis and implications. Am J Med Genet A. 2013;161A(5):935–50.
11. Ong FS, et al. Translational utility of next-generation sequencing. Genomics. 2013;102(3): 37–139.
12. Kingsmore SF, Saunders CJ. Deep sequencing of patient genomes for disease diagnosis: when will it become routine? Sci Transl Med. 2011;3(87):p87ps23.
13. Ng PC, Henikoff S. Predicting the effects of amino acid substitutions on protein function. Annu Rev Genomics Hum Genet. 2006;7:61–80.
14. Sim NL, et al. SIFT web server: predicting effects of amino acid substitutions on proteins. Nucleic Acids Res. 2012;40(Web Server Issue):W452–7.
15. Spurdle AB, et al. ENIGMA-evidence-based network for the interpretation of germline mutant alleles: an international initiative to evaluate risk and clinical significance associated with sequence variation in BRCA1 and BRCA2 genes. Hum Mutat. 2012;33(1):2–7.
16. Bamshad MJ, et al. Exome sequencing as a tool for a Mendelian disease gene discovery. Nat Rev Genet. 2011;12(11):745–55.
17. Kurian AW, et al. Clinical evaluation of a multiple-gene sequencing panel for hereditary cancer risk assessment. J Clin Oncol. 2014;32(19):2001–9.
18. Domchek S, Weber BL. Genetic variants of uncertain significance: flies in the ointment. J Clin Oncol. 2008;26(1):16–7.
19. Rehm HL, et al. ACMG clinical laboratory standards for next-generation sequencing. Genet Med. 2013;15(9):733–47.
20. Meldrum C, et al. Next-generation sequencing for cancer diagnostics: a practical perspective. Clin Biochem Rev. 2011;32(4):177–95.
21. Fecteau H, et al. The evolution of cancer risk assessment in the era of next generation sequencing. J Genet Couns. 2014;23(4):633–9.
22. Hall MJ, et al. Gene panel testing for inherited cancer risk. J Natl Compr Canc Netw. 2014;12(9):1339–46.
23. Gail MH. Personalized estimates of breast cancer risk in clinical practice and public health. Stat Med. 2011;30(10):1090–104.
24. Riley BD, et al. Essential elements of genetic cancer risk assessment, counseling, and testing: updated recommendations of the National Society of Genetic Counselors. J Genet Couns. 2012;21(2):151–61.
25. National Comprehensive Cancer Network. Genetic/familial high-risk assessment: breast and ovarian, V.1. 2014. http://www.nccn.org/professionals/physician_gls/recently_updated.asp. Accessed 22 Jul 2014.
26. Robson ME. et al. Genetic and Genomic Testing for Cancer Susceptibility. J Clin Oncol: American Society of Clinical Oncology Policy Statement Update; 2010.
27. Hiraki S, et al. Cancer risk assessment using genetic panel testing: considerations for clinical application. J Genet Couns. 2014;23(4):604–17.
28. Selkirk CG, et al. Cancer genetic testing panels for inherited cancer susceptibility: the clinical experience of a large adult genetics practice. Fam Cancer. 2014. [Epub ahead of print].

29. Laduca H, et al. Utilization of multigene panels in hereditary cancer predisposition testing: analysis of more than 2,000 patients. Genet Med. 2014. [Epub ahead of print].
30. Mauer CB, et al. The integration of next-generation sequencing panels in the clinical cancer genetics practice: an institutional experience. Genet Med. 2014;16(5):407–12.
31. Churpek J, et al. Inherited mutations in breast cancer genes in African American breast cancer patients revealed by targeted genomic capture and next-generation sequencing. Clin Oncol. 2013;31(Suppl):abstr CRA1501.
32. Yurgelun M, et al. Germline mutations identified by a 25-gene panel in patients undergoing Lynch syndrome testing. Presented at the collaborative group of the Americas on inherited colorectal cancer (CGA-ICC) annual meeting, 7–8 Oct 2013, Anaheim, CA
33. Norquist BM, et al. Characteristics of women with ovarian carcinoma who have BRCA1 and BRCA2 mutations not identified by clinical testing. Gynecol Oncol. 2013;20(6):1483–7.
34. Domchek SM, et al. Association of risk-reducing surgery in BRCA1 or BRCA2 mutation carriers with cancer risk and mortality. JAMA. 2010;304(9):967–75.
35. Antoniou AC, et al. Breast-cancer risk in families with mutations in PALB2. N Engl J Med. 2014;371(6):497–506.
36. Green RC, et al. ACMG recommendations for reporting of incidental findings in clinical exome and genome sequencing. Genet Med. 2013.
37. ACMG updates recommendation on "Opt Out" for genome sequencing return of results. https://www.acmg.net/docs/Release_ACMGUpdatesRecommendations_final.pdf. Accessed 20 Aug 2014.
38. Ross LF, et al. Technical report: ethical and policy issues in genetic testing and screening of children. Genet Med. 2013;15(3):234–45.
39. Parsons DW, et al. Clinical tumor sequencing: an incidental casualty of the American College of Medical Genetics and Genomics recommendations for reporting of incidental findings. J Clin Oncol. 2013;32(21):2203–5.
40. Everett JN, et al. Traditional roles in a non-traditional setting: genetic counseling in precision oncology. J Genet Couns. 2014;23(4):655–60.
41. Rainville IR, Rana HQ. Next-generation sequencing for inherited breast cancer risk: counseling through the complexity. Curr Oncol Rep. 2014;16(3):371.
42. Pletcher BA, et al. Indications for genetic referral: a guide for health care professionals. Genet Med. 2007;9(6):385–9.
43. National accreditation program for breast centers genetic evaluation and management. 2013. p. 48. http://napbc-breast.org/standards/2013standardsmanual.pdf. Accessed 15 June 15 2013.
44. Trepanier A, et al. Genetic cancer risk assessment and counseling: recommendations of the national society of genetic counselors. J Genet Couns. 2004;13(2):83–114.
45. Halbert CH, et al. Long-term reactions to genetic testing for BRCA1 and BRCA2 mutations: does time heal women's concerns? J Clin Oncol. 2011;29(32):4302–6.
46. Hamilton JG, et al. Emotional distress following genetic testing for hereditary breast and ovarian cancer: a meta-analytic review. Health Psychol. 2009;28(4):510–8.
47. U.S. Preventive Services Task Force. Assessing the genetic risk for BRCA-related breast or ovarian cancer in women: recommendations from the U.S. Preventive Services Task Force. Ann Internal Med. 2014;160(4):16.
48. Domchek SM, et al. Multiplex genetic testing for cancer susceptibility: out on the high wire without a net? J Clin Oncol. 2013;31(10):1267–70.
49. Rahman N, et al. Cancer genes associated with phenotypes in monoallelic and biallelic mutation carriers: new lessons from old players. Hum Mol Genet. 2007;16(Spec No 1): I60–6.
50. Radford C, et al. Factors which impact the delivery of genetic risk assessment services focused on inherited cancer genomics: expanding the role and reach of certified genetics professionals. J Genet Couns. 2013;23(4):522–30.
51. Javitt GH, et al. Federal neglect: regulation of genetic testing. Issues Sci Technol. 2006;22(3): 59–66.
52. Horn EJ, et al. Regulating genetic tests: issues that guide policy decisions. Genet Test Mol Biomarkers. 2012;16(1):1–2.

53. FDA takes steps to help ensure the reliability of certain diagnostic tests. http://www.fda.gov/NewsEvents/Newsroom/PressAnnouncements/ucm407321.htm. Accessed 25 Aug 2014.
54. Wideroff L, et al. Hereditary breast/ovarian and colorectal cancer genetics knowledge in a national sample of US physicians. J Med Genet. 2005;42(10):749–55.
55. Vig HS, et al. Cancer genetic risk assessment and referral patterns in primary care. Genet Test Mol Biomarkers. 2009;13(6):735–41.
56. Cragun D, et al. Patient reported differences in BRCA pretest counseling based on ordering provider type. National Society of Genetic Counselors 32nd Annual Education Conference. 2013.
57. Zon RT, et al. American Society of Clinical Oncology policy statement: the role of the oncologist in cancer prevention and risk assessment. J Clin Oncol. 2009;27(6):986–93.
58. Geier LJ, et al. Clinical Cancer Genetics Remains a Specialized Area: How Do I Get There From Here? 2009.
59. Evans JP. Health care in the age of genetic medicine. JAMA. 2007;298(22):2670–2.
60. Bowen DJ, et al. Marketing genetic tests: empowerment or snake oil? Health Educ Behav. 2005;32(5):676–85.
61. Vadaparampil ST, et al. The impact of acculturation on awareness of genetic testing for increased cancer risk among Hispanics in the year 2000 National Health Interview Survey. Cancer Epidemiol Biomarkers Prev. 2006;15(4):618–23.
62. Beitsch PD, et al. Can breast surgeons provide breast cancer genetic testing? An American Society of Breast Surgeons Survey. Ann Surg Oncol. 2014. [Epub ahead of print].
63. Professional Status Survey. 2014. http://nsgc.org/p/cm/ld/fid=68. Accessed 29 Aug 2014.
64. Cohen SA, et al. Identification of genetic counseling service delivery models in practice: a report from the NSGC Service Delivery Model Task Force. J Genet Couns. 2013;22(4):411–21.
65. American Board of Genetic Counseling. http://www.abgc.net/ABGC/AmericanBoardofGeneticCounselors.asp. Accessed 4 Sept 2013.
66. McPherson E, et al. Clinical genetics provider real-time workflow study. Genet Med. 2008;10(9):699–706.
67. Miller CE, et al. Genetic counselor review of genetic test orders in a reference laboratory reduces unnecessary testing. Am J Med Genet A. 2014;164A(5):1094–101.
68. Pal T, et al. A statewide survey of practitioners to assess knowledge and clinical practices regarding hereditary breast and ovarian cancer. Genet Test Mol Biomarkers. 2013;17(5):367–75.
69. Pal T, et al. Modes of delivery of genetic testing services and the uptake of cancer risk management strategies in BRCA1 and BRCA2 carriers. Clin Genet. 2013;85(1):49–53.
70. Plon SE, et al. Genetic testing and cancer risk management recommendations by physicians for at-risk relatives. Genet Med. 2011;13(2):148–54.
71. Senter L, et al. Linking distant relatives with BRCA gene mutations: potential for cost savings. Clin Genet. 2013;85(1):54–8.
72. Cragun D, et al. Differences in BRCA counseling and testing practices based on ordering provider type. Genet Med. 2014. [Epub ahead of print].
73. Duquette D, et al. Using core public health functions to promote BRCA best practices among health plans. Public Health Genomics. 2012;15(2):92–7.
74. Cragun D, et al. Panel-based testing for inherited colorectal cancer: a descriptive study of clinical testing performed by a US laboratory. Clin Genet. 2014. [Epub ahead of print].
75. Cohen SA, et al. A collaborative approach to genetic testing: a community hospital's experience. J Genet Couns. 2009;18(6):530–3.
76. MacDonald DJ, et al. Extending comprehensive cancer center expertise in clinical cancer genetics and genomics to diverse communities: the power of partnership. J Natl Compr Canc Netw. 2010;8(5):615–24.
77. Landrum MJ, et al. ClinVar: public archive of relationships among sequence variation and human phenotype. Nucleic Acids Res. 2014;42(Database issue):D980–5.

78. Pearce CL, et al. Validating genetic risk associations for ovarian cancer through the international ovarian cancer association consortium. Br J Cancer. 2009;100(2):412–20.
79. Chenevix-Trench G, et al. An international initiative to identify genetic modifiers of cancer risk in BRCA1 and BRCA2 mutation carriers: the consortium of investigators of modifiers of BRCA1 and BRCA2 (CIMBA). Breast Cancer Res. 2007;9(2):104.
80. Hayden EC. Cancer-gene data sharing boosted. Nature. 2014;510(7504):198.
81. Offit K. Personalized medicine: new genomics, old lessons. Hum Genet. 2011;130(1):3–14.

Next-Generation Sequencing for Cancer Biomarker Discovery

Aarti N. Desai and Abhay Jere

Abstract Cancer is a genetic disorder that arises from gene mutations as well as changes in transcriptional and epigenetic profiles. These genetic changes can serve as valuable biomarkers for early detection, staging, and detailed molecular characterization of cancer for individualized therapy. Mutations in several known oncogenes (e.g., EGFR, HER2, KRAS) and tumor suppressor genes (e.g., TP53, PTEN, PI3K) are already being used as biomarkers to guide therapy in breast cancer, ovarian cancer, lung cancer, prostate cancer, etc. However, tumor heterogeneity and instability of cancer genomes poses a significant challenge to reliable and reproducible detection of biomarkers. Moreover, cancer is a multigene disorder and comprehensive knowledge of the mutational landscape is extremely important for the most effective therapeutic intervention.

Next-Generation Sequencing (NGS) is a high-throughput genome sequencing technology that enables sequencing of entire genomes or thousands of mutations simultaneously in a cost effective manner and hence can serve as a very powerful tool in biomarker detection and discovery. Many NGS-based studies published in the last few years have identified potential prognostic and predictive molecular signatures. In this chapter, we discuss the impact of NGS on cancer biomarker detection as well as discovery and the resulting paradigm shift in cancer care.

1 What Are Biomarkers?

Biomarkers are key molecular, chemical or cellular characteristics that can be objectively measured and used to describe biological processes, pathogenic state and response to therapy. Biomarkers can be either disease related or therapy related. Disease related biomarkers are diagnostic (used to establish the disease state), prognostic (provide information regarding potential clinical outcome

A.N. Desai, Ph.D. • A. Jere, Ph.D. (✉)
Persistent Labs, Persistent Systems Ltd.,
Pingala – Aryabhata 12A/12, Off Karve Road, Erandwane, Pune, Maharashtra 411004, India
e-mail: abhay_jere@persistent.co.in

W. Wu, H. Choudhry (eds.), *Next Generation Sequencing in Cancer Research, Volume 2*, DOI 10.1007/978-3-319-15811-2_7

irrespective of the treatment), or predictive (provide information regarding potential clinical outcome in response to specific treatment) [1–3].

Therapy related biomarkers provide information regarding the effectiveness of a particular drug in treating or managing a disease. Additionally, the Biomarkers and Surrogate End Point Working Group [4, 5] has classified biomarkers as Type 0 biomarkers that are markers of natural history of the disease and correlate with clinical indices (e.g., complete blood count), Type I biomarkers represent the effects of a therapeutic intervention in accordance with drug mechanism of action (e.g., decrease in urokinase-type plasminogen activator (uPA) gene expression upon dasatinib treatment in prostate cancer) and Type II biomarkers that are surrogate end points as a change in this marker indicates clinical benefit (e.g., decrease in blood/urine glucose level).

For a biomarker to become accepted for clinical application it has to have the following characteristics:

1. Readily and consistently detectable in biological fluids, tissues, or other biological specimens.
2. Rapidly detectable and stable.
3. High sensitivity and specificity.
4. Strong correlation with the phenotype or outcome of interest.
5. Detectable via a simple, noninvasive, and cost-effective test.
6. Consistent across genders.

In addition to the aforementioned properties, biomarkers to be used in cancer should be specific to the cancer subtype, provide information about the metastatic potential of the cancer, and should also be detectable in archived samples such as FFPE (Formalin-Fixed Paraffin-Embedded) blocks. Biomarkers such as Cancer Antigen 125 (CA 125) are an example of blood biomarker for ovarian cancer [6–8] whereas Prostate Specific Antigen (PSA) is a marker specific to prostate cancer [9, 10]. Additionally, imaging techniques such as CT scan, mammography, and ultrasound are also widely used in cancer detection as well as characterization [11]. As cancer arises from genetic aberrations and is a heterogeneous disease, molecular biomarkers such as genetic variations, gene expression profiles, and in some cases the genome methylation status may provide more actionable insights than traditional markers.

2 Limitation of Traditional Biomarkers

While the biomarkers mentioned in the previous section are very useful in establishing diagnosis in many cases, they suffer from low specificity and sensitivity. PSA which is an FDA approved marker for prostate cancer is also found to be elevated in other conditions such as benign prostate hyperplasia and prostatitis [12]. Another example is Nuclear Matrix Protein 22 (NMP22), a marker widely used in bladder cancer that is also elevated in pyuria, urolithiasis, or cystitis [13]. Moreover, it is

becoming increasingly obvious that even though blood biomarkers along with imaging and other techniques, may be useful in cancer diagnosis, these may not provide enough information for the best course of therapeutic intervention. With the advent of targeted therapy and more recently precision medicine, identification of reliable, precise and clinically relevant biomarkers has become extremely critical. The term "precision medicine" is used to describe therapeutic interventions derived from better understanding of the genetic as well as mechanistic underpinnings of a disease [14].

Therapy targeted to counter specific genetic aberrations has been used most successfully in cancer as it is a genetic disease resulting from mutations in oncogenes and tumor suppressor genes [15–18]. Treatment of Chronic Myelogenous Leukemia (CML) patients carrying a BCR-ABL (Breakpoint Cluster Region—Abl Tyrosine Kinase) translocation with imatinib [19, 20] and of HER2/neu (ERBB2; v-erb-b2 avian erythroblastic leukemia viral oncogene homolog 2) positive breast cancer patients with Trastuzumab [21, 22] are the most notable examples of cancer treatment based on molecular characterization of the tumors. A number of cancer drugs approved by the FDA in the last decade have been against specific genetic aberrations. For example, Gefitinib has been approved for lung adenocarcinomas harboring EGFR (Epidermal Growth Factor Receptor) mutations [23] and Vemurafinib has been approved for melanoma patients harboring the V600E BRAF (B-Raf proto-oncogene, serine/threonine kinase) mutation [24].

3 NGS in Biomarker Discovery and Testing

Next-Generation Sequencing (NGS) is a collection of latest high-throughput sequencing technologies that enable performing millions of sequencing reactions in parallel. We and others [25–27] have previously discussed the details of NGS technology and its potential applications in clinic. With its unprecedented scale and rapidly declining cost, NGS brings within grasp the possibility of sequencing the entire cancer genomes or at the very least entire cancer exomes. Given the cancer heterogeneity and the rapidly changing genetic landscape in cancer, this provides a very powerful tool to get an unbiased view of the genome and significantly improves the chances of identifying actionable genetic aberrations. In addition to detecting changes in DNA sequence, NGS can also be used in transcriptome profiling as well as methylation detection. As has been shown in breast cancer, glioblastoma, and numerous other cancers, gene expression signatures are very important markers of cancer subtypes [28], metastatic potential [29, 30], response to therapy [31], and survival [29, 32]. Currently, NGS is being used most commonly for variant detection using a "panel-based" approach for known mutations (amplicon-based target enrichment) or known oncogenes.

The application of NGS technology in clinics, loosely known as clinical NGS, has begun to provide extraordinary insights into genetic mutations in a large set of genes [33–35], novel mutations in genes previously implicated in cancer [36, 37] and genes

previously not associated with particular cancer [38, 39]. Commercially available NGS-based cancer panels are already being used in clinical practice to guide patients to most appropriate treatment [40–42].

4 Types of Biomarkers That Can Be Identified Using NGS

4.1 Genetic Variants as Biomarkers

Mutations in the form of single nucleotide variants, insertions, deletions and other structural variants are commonly found to be associated with onset, progression, and metastatic potential of cancer as well as effectiveness of therapy. Mutations in BRCA1/2 (Breast Cancer 1/2) [43, 44], KRAS (Kirsten rat sarcoma viral oncogene homolog) [45], PTEN (Phosphatase and Tensin homolog) [46], etc. are already established as prognostic as well as predictive markers. Mutations in multiple genes are typically found in cancer cells and although genetic aberrations in a single gene can be useful in determining therapeutic intervention, a more detailed picture of the mutational landscape is of immense value. A recent study by Kurian et al. [47] used a 49-gene NGS-based panel for genome characterization of 198 breast cancer patients. A majority of patients carried BRCA1/2 mutation, however, among women that tested negative for BRCA1/2 mutations, 16 potentially pathogenic mutations in other genes were identified, of which presence of 15 mutations indicated that the patients would benefit from a change in care.

A very comprehensive analysis of sequencing and expression data from 12 tumor types in The Cancer Genome Atlas (TCGA) project has shown that genes such as ERBB2, FGFR1 (Fibroblast Growth Factor Receptor 1), KRAS, PIK3CA (Phosphatidylinositol-4, 5-bisphosphate 3-kinase, catalytic subunit alpha), CDKN2A (Cyclin-Dependent Kinase inhibitor 2A), ATM (ATM serine/threonine kinase), and MDM4 are altered in multiple cancer types irrespective of the tissue of origin [48]. Moreover, the study showed that different genes were altered in patients with same type of cancer. For example, FGFR1, PIK3CA, CDKN2A, and TP53 (Tumor Protein p53) were selectively altered in patients with lung squamous cell carcinoma whereas ERBB2, PIK3CA, CCNE1 (Cyclin E1), AURKA (Aurora Kinase A), and TP53 were selectively altered in serous uterine corpus endometrioid carcinoma.

These studies emphasize the complexity of genetic alterations in cancer and provide strong evidence to support the benefits of screening for mutations in multiple known oncogenes to arrive at optimum cancer management approach. As NGS allows for simultaneous detection of mutations in multiple genes and even whole genome, NGS-based tests can be used to detect the known as well as novel cancer mutations. As mentioned above, a number of commercial tests with panels of mutations in known oncogenes are already available in market. An added advantage of using NGS for biomarker testing is that genetic material extracted from archived samples such as FFPE blocks can also be used.

4.2 Gene Expression Profiles as Biomarkers

Microarray, one of the first high-throughput genomic technologies, engendered the use of gene expression profiles comprising of several thousand genes in diagnosis and classification of cancer. Breast cancer [49–51], colon cancer [52, 53], and glioblastoma [54] are examples of cancer types where gene expression profiles are extensively used. Oncotype DX® a RT-PCR-based gene expression profiling test is available for breast cancer, colon cancer as well as prostate cancer [55]. Mammaprint® [56], BluePrint®, and TargetPrint®, offered by Agendia [57] are breast cancer tests assessing expression pattern of signature genes. However, almost all of the gene expression assays (except for those using exon arrays) report expression at "gene level." In reality, there exist many isoforms of a gene and numerous studies have shown altered expression of specific gene isoforms in cancer [58, 59]. Using NGS to perform transcriptome profiling offers a very distinct advantage, since the expression data is captured at the isoform level. Isoform level transcriptome profiling can provide unique and important insight into cancer progression and metastasis.

4.3 Epigenetic Modifications as Biomarkers

Epigenetic modifications are changes in DNA independent of variations in DNA sequence. Epigenetic modifications, such as DNA methylation, histone acetylation and methylation have profound impact on gene expression. Changes in histone modification [60, 61] and DNA methylation [62–64] pattern of genes is one of the hallmarks of cancers and can act as biomarkers for cancer detection and therapeutic intervention. Examples of epigenetic modifications used as biomarkers include hypermethylation of GSTP1 (Glutathione S-Transferase pi 1) in prostate cancer patients [65, 66] and hypermethylation of DAPK (Death Associated Protein Kinase) as well as RASSF1A (Ras association (RalGDS/AF-6) domain family member 1) genes in bladder cancer [67, 68]. More recently, Wasserkort et al. [69] reported that the cytosine residues in the v2 region of the Septin9 gene are specifically methylated in colorectal cancer tissue but not in normal colon mucosa. Using a whole-genome methylation detection technology Mah et al. [70] were able to identify more than 500 differentially methylated genes in hepatocellular carcinoma (HCC). Moreover, they were able to use the differentially methylated regions to classify the HCC patients in three subgroups with one of the group showing extremely poor survival. This study and many others have suggested that methylation status of hundreds of genes is altered in cancer and hence a high-throughput technology such as NGS is very well suited to identify epigenetic modifications as biomarkers in cancer. For an elaborate review of the role of epigenetic modifications in cancer and list of epigenetic biomarkers, please refer to Taby and Issa [71].

4.4 MicroRNA as Biomarkers

MicroRNAs are small (19–22 nucleotide long) single stranded RNA molecules that play a critical role in regulation of gene expression by targeting mRNA for degradation or by suppressing translation [72]. MicroRNA mediated regulation of gene expression is important in development, differentiation and cell growth, hence microRNAs could potentially play a key role in carcinogenesis. One of the earliest studies to demonstrate the role of microRNAs in cancer was in Chronic Lymphocytic Leukemia (CLL), where miR15 and miR16 located in a genomic region previously associated with CLL, were shown to be downregulated in 68 % of CLL cases [73]. Subsequently, many studies have shown that disruption of microRNA mediated regulation of gene expression leads to tumorigenesis and oncogenic transformation [74–77]. In a very recent study, miR206 was found to be downregulated in 93 % of breast cancer cases studied suggesting it be a good candidate as biomarker [78]. Similarly, elevated levels of miR-19A found in metastatic HER2 +ve inflammatory breast cancer patients was associated with better clinical outcome (longer progression free and overall survival) [79]. The role of microRNAs in cancer is a very active field of research and it is likely that many key molecules are yet to be discovered. Hence, NGS-based approaches for deciphering the role of microRNAs in all aspects of cancer biology will prove very valuable.

5 Examples of NGS Led Biomarker Detection in Various Cancers

In the last few years, many cancer studies have used NGS to glean very valuable information regarding different subtypes of cancer [54, 80, 81], molecular signatures, novel mutations, and mutations associated with metastasis, progression as well as response to therapy [29, 30, 82, 83]. Most recently, integrated analysis of genomic data from multiple high-throughput technologies available in TCGA was used to identify markers to classify cancers irrespective of the tissue of origin [84]. Cancers such as breast cancer, lung cancer, colon cancer, ovarian cancer, and acute myeloid leukemia have benefited the most from using NGS. Table 1 summarizes a list of novel cancer genes identified in various cancer types using the whole-genome or whole-exome sequencing methodology.

5.1 Breast Cancer

Breast cancer is the most common cancer in women in the USA, with nearly 235,030 estimated cases as well as 40,430 estimated deaths in 2014 [85]. Like all cancers, breast cancer arises from genetic mutations and almost 20 % of breast cancer patients have family history of cancer [86]. Inherited genetic mutations in the BRCA1

Table 1 Novel cancer genes identified using next-generation sequencing

Gene name	Cancer type	References
SOX9, NAV2-TCF7L1 fusion, CDH10, FAT4, DOCK2	Colorectal cancer	[81, 132]
PRPS2, PRKCZ, PRKCQ, PRKG1, PRKCE, NRC31	Triple negative breast cancer	[128]
AKT2, ARID1B, CASP8, CDKN1B, MAP3K1, MAP3K13, NCOR1, SMARCD1, TBX3	Breast cancer	[129]
LZTR1, SPTA1, ATRX, GABRA6, KEL	Glioblastoma multiforme	[130]
TSHR, ROCK1, ROCK2	Gastric adenocarcinoma	[131]
GRIN2A, TMEM132B, ZNF831, PLCB4, TAS2R60, KHDRBS2, C12orf63	Melanoma	[133]
ARID1A, PPP2R1A	Ovarian clear cell carcinoma	[134]
PBRM1, HIF1a, JARID1C, SETD2, PMS1	Renal cell carcinoma	[38, 61]
IDH1, ND4, CDC42, IMPG2, FREM2, ANKRD26, CEP170, CDH24, PCLKC, GPR1233, EBI2, KNDC1, SLC15A1, GRINL1B	Acute myeloid leukemia	[135, 136]
FUS-NCATc2 and CIC-FOXO4 fusions	Ewing sarcoma	[137]

and BRCA2 genes contribute towards 5–10 % of diagnosed breast cancer cases [86, 87] and mutations in TP53, PTEN, and STK11 (Serine/Threonine Kinase 11) contribute towards increased risk of breast cancer in the case of Li–Fraumeni syndrome [88–90], Cowden syndrome [91] and Peutz–Jeghers syndrome, respectively [92, 93]. Additionally, mutations in CHEK2 (Checkpoint Kinase 2), ATM, NBN (Nibrin), RAD50 (RAD50 Homolog), BRIP1 (BRCA1 interacting protein C-terminal helicase 1), and PALB2 (partner and localizer of BRCA2) are associated with increased risk of breast cancer [94, 95]. Table 2 provides a list of genes commonly mutated (obtained from COSMIC) in breast cancer, NGS-based tests available, and potential therapeutic intervention.

Breast cancer is the most well-characterized cancer; however, approximately 90 % of breast cancers arise from sporadic mutations in a few key cancer genes, and hence, it is important to have a biomarker assay that can in a single test provide information about potentially actionable mutations. Not surprisingly, breast cancer was one of the first cancers for which a NGS-based multigene panel for mutation detection was developed [96]. Walsh et al. [96] developed a targeted sequencing panel of 21 genes associated with breast and ovarian cancer and sequenced the DNA of 20 female cancer patients with a known mutation in at least one of the genes responsible for inherited predisposition to these diseases. They were able to successfully detect all the point mutations and small indels (1–19 bp) that were part of the test and did not detect any false positives. Additionally, they were able to detect five large deletions and one duplication event in the BRCA1 and BRCA2 genes. Thus, Walsh et al. [96] was the first group to successfully demonstrate the applicability and benefits of using NGS panels in diagnostics. Since then several publications [97–100] have demonstrated the applicability of NGS-based panels in detecting known and novel mutations as well as actionable therapeutic targets in cancer samples.

Table 2 List of genes mutated at high frequency in breast cancer and currently tested using NGS panels

Gene name	Mutation frequency[a] (%)	Available NGS panel cancer test	Potential therapeutic intervention[b]
AKAP9	2		
AKT1	3	FoundationOne™	
APC	2	Ambry Genetics CancerNext Ion AmpliSeq Cancer Hotspot v2 FoundationOne™ TruSeq Amplicon Cancer Panel	Perifosine [138, 139] MK2206 [140], Cenisertib, Iatasertib, Afuresertib, Uprosertib
ARID1A	3	FoundationOne™	
ATM	2	Ambry Genetics BreastNext Ion AmpliSeq Cancer Hotspot v2 FoundationOne™ TruSeq Amplicon Cancer Panel	Everolimus, Temsirolimus [141]
BRCA1	2	Ambry Genetics BreastNext FoundationOne™	Rucaparib [142], Niraparib [143], Veliparib, Olaprib [144]
BRCA2		Ambry Genetics BreastNext FoundationOne™	Rucaparib [142], Niraparib [143]
CHEK2		Ambry Genetics BreastNext FoundationOne™	
CDH1	12	Ambry Genetics BreastNext Ion AmpliSeq Cancer Hotspot v2 FoundationOne™ TruSeq Amplicon Cancer Panel	
GATA3	7	FoundationOne™	
KMT2D	3	FoundationOne™	
MAP2K4	2	FoundationOne™	
MED12	3	FoundationOne™	
MLL3	7		
MYH9	2		
NF1	2	Ambry Genetics BreastNext FoundationOne™	PD325901 [147]
PALB2		Ambry Genetics BreastNext Foundation One	
PIK3CA	26	Ion AmpliSeq Cancer Hotspot v2 FoundationOne™ TruSeq Amplicon Cancer panel	Buparlisib, BEZ235, BGT226, GSK2126458, GDC-0941 Bismesylate
PTEN	4	Ambry Genetics BreastNext Ion AmpliSeq Cancer Hotspot v2 FoundationOne™ TruSeq Amplicon Cancer Panel	Everolimus, Temsirolimus
RB1	3	Ambry Genetics Retinoblastoma Ion AmpliSeq Cancer Hotspot v2 FoundationOne™ TruSeq Amplicon Cancer Panel	
RUNX1	2	FoundationOne™	
TP53	23	Ambry Genetics BreastNext Ion AmpliSeq Cancer Hotspot v2 FoundationOne™ TruSeq Amplicon Cancer Panel	Ad5CMV-p53 gene [145]
UBR5	2		

[a]*Source*: COSMIC [149]

[b]Unless mentioned otherwise, the source of information for potential therapeutic intervention is [150]

Most recently, Chong et al. [40] have reported on development of BRCAPlus, a clinical diagnostic assay that detects mutations in six high risk breast cancer susceptibility genes; BRCA1, BRCA2, CDH1 (Cadherin 1), PTEN, TP53, and STK11. In this fairly large study (250 previously characterized samples and 3,000 new clinical samples), the BRCAPlus test was able to identify all the 3,025 known germ-line mutations in the 250 previously characterized samples and was also able to detect pathogenic mutations in the BRCA2 gene in two clinical samples that had previously tested negative for mutations in BRCA1 and BRCA2. This study demonstrates that NGS panels have high sensitivity and can be particularly useful in identifying low level complex mutations as well as heterozygous mutations that can be sometimes missed by Sanger sequencing.

In addition to targeted sequencing, whole-genome sequencing has also been used in identifying mutations in breast cancer, especially in the absence of mutations in BRCA1 and BRCA2. Link et al. [101] used the whole-genome sequencing approach to identify mutations in the genome of a patient with early onset breast and ovarian cancer who had also developed therapy-related acute myeloid leukemia. This patient had no family history of breast or ovarian cancer and tested negative for the conventional BRCA1 and BRCA2 mutations tested by commercial tests. Performing whole-genome sequencing on the skin (normal) and bone marrow (leukemia) genome of the patient revealed a novel 3 kb heterozygous deletion in the TP53 gene (exons 7–9) of the normal genome and a 17.6 mb region of uniparental disomy on chromosome 17 with resultant homozygous deletion of the same region (exons 7–9 of TP53 gene) of the leukemia genome. The loss of exons 7–9 resulted in loss of DNA binding domain in the TP53 protein and thus produced a functionally defective protein. The authors concluded that this deletion mutation in the TP53 gene was most likely contributing to high cancer susceptibility in this patient.

5.2 Lung Cancer

Lung cancer is the most common cancer worldwide and the most common cause of cancer-related death [102]. Lung cancer is usually divided into two broad categories: non-small-cell lung cancer (NSCLC) and small-cell lung cancer (SCLC). NSCLCs account for a majority (approximately 85 %) of lung cancers and are further divided into three subtypes: squamous-cell carcinoma (SCC), adenocarcinoma, and large-cell lung cancer. Adenocarcinomas are considered to be the most common lung carcinoma subtype, constituting approximately 40 % of all NSCLC. The primary cause of lung cancer is cigarette smoking with about 75–90 % lung cancer patients being moderate to heavy smokers, whereas 10–25 % cases of lung cancer occur in nonsmokers [103, 104]. The proposed causes of lung cancer in nonsmokers is exposure to secondhand smoke, cooking fumes, exposure to carcinogenic agents such as asbestos, arsenic, radiation, and some air pollutants and even genetic aberrations [105, 106]. Genetic aberrations are found in lung cancer patients with or without smoking history; however, nonsmokers have fewer mutations as compared

to smokers [107, 108]. Hence, even though environmental factors play a critical role in precipitating lung cancer, genetic aberrations are probably important in tumor progression and metastasis. A number of studies in the last few years have generated a long list of genes mutated at very high frequency in different types of lung cancer and many of them (e.g., mutations in EGFR and ALK fusions) serve as biomarkers for choosing appropriate therapy. Table 3 provides a list of genes commonly mutated (obtained from COSMIC) in lung cancer, NGS-based tests available and potential therapeutic intervention.

A small set of well characterized NSCLCs was recently subjected to integrated analysis of genome and transcriptome sequencing data by Govindan et al. [108] and they have reported several interesting findings particularly with regard to potential therapeutic targets in this lung cancer type. The analysis of genome sequencing data not only revealed mutations in known lung cancer associated genes such as KRAS, TP53, EGFR, BRAF, JAK2 (Janus Kinase 2), JAK3 (Janus Kinase 3), and EPHA3 (EPH receptor A3) but also identified a few significantly mutated genes not previously associated with lung cancer. Of these, DACH1 (Dachshund family transcription factor 1) is reported in other cancers such as breast cancer, gliomas and prostate cancer. Moreover, using genetic variant and gene expression data, the authors were able to identify known (e.g., mutations in KRAS, EGFR and BRAF) as well as novel [mutations in PRKCB2 (Protein Kinase C beta), MET, JAK2, HGF (hepatocyte growth factor), and ERBB2] therapeutic targets.

The TCGA group recently published a thorough molecular characterization of lung adenocarcinoma [109] by performing integrative analysis of whole-exome, gene expression, and epigenetic data. Analysis of exome sequencing data from 230 adenocarcinoma samples revealed that 62 % (143/230) samples carried known activating mutations in known driver oncogenes. Of these, 32 % samples harbored mutations in KRAS, 11 % in EGFR, 7 % in BRAF and a small fraction of samples had mutations in ERBB2, MAP2K1 (Mitogen Activated Protein Kinase Kinase 1), NRAS (Neuroblastoma RAS viral (v-ras) oncogene homolog), and HRAS (Harvey Rat Sarcoma viral oncogene homolog). In the remaining samples (38 %) which the authors have defined as oncogene negative tumors, the authors reported a significant enrichment of TP53, KEAP1 (Kelch-like ECH-Associated Protein 1), NF1 (Neurofibromin 1), and RIT1 (Ras-like without CAAX 1) mutations. On further analysis of the data, the authors recommend amplifications in MET (MET proto-oncogene, receptor tyrosine kinase) and ERBB2 as well as mutations in NF1 and RIT1 as drivers in the oncogene-negative lung adenocarcinomas. This comprehensive study underscores the potential of the NGS technology to identify molecular subtypes within a tumor population and elucidate novel markers.

Lung squamous cell carcinoma (SQCC) a form of lung cancer that does not typically harbor therapeutically relevant activating mutations in EGFR and ALK fusions were comprehensively studied by the TCGA group [83]. In this study, data from various high-throughput genomic technologies, especially NGS, was used to characterize the genomic and epigenomic landscape as well as to identify potential therapeutic targets. The group reported 22 genes [e.g., TP53, CDKN2A, PTEN,

Table 3 List of gene mutated at high frequency in lung cancer and currently tested using NGS panel

Gene name	Mutation frequency[a] (%)	Available NGS panel cancer test	Potential therapeutic intervention[b]
AKAP9	5		
ATM	5	Ambry Genetics CancerNext Ion AmpliSeq Cancer Hotspot v2 FoundationOne™ TruSeq Amplicon Cancer Panel	
ATRX	5	FoundationOne™	
CDKN2A	9	Ion AmpliSeq Cancer Hotspot v2 FoundationOne™ TruSeq Amplicon Cancer Panel	Olomoucine, Roscovitine [141], Roniciclib, Alvocidib, Dinaciclib, Seliciclib
CREBBP	5	FoundationOne™	
EGFR	28	Ion AmpliSeq Cancer Hotspot v2 FoundationOne™ TruSeq Amplicon Cancer Panel	Erlotinib, Gefitinib [140], Vandetanib, Afatinib, Icotinib, Canertinib, Epitinib
KDR	5	Ion AmpliSeq Cancer Hotspot v2 FoundationOne™ TruSeq Amplicon Cancer Panel	
KMT2D	7	FoundationOne™	
KRAS	16	Ion AmpliSeq Cancer Hotspot v2 FoundationOne™ TruSeq Amplicon Cancer Panel	
MLL3	11	FoundationOne™	
NF1	7	FoundationOne™	Everolimus, Temsirolimus [140]
NFE2L2	5	FoundationOne™	
PDE4DIP	5		
RB1	5	Ambry Genetics Retinoblastoma Ion AmpliSeq Cancer Hotspot v2 FoundationOne™ TruSeq Amplicon Cancer Panel	
SETBP1	5		
SMARCA4	5	FoundationOne™	
STK11	7	Ambry Genetics CancerNext FoundationOne™ Ion AmpliSeq Cancer Hotspot v2 TruSeq Amplicon Cancer Panel	Everolimus and Temsirolimus [148]
TP53	34	Ambry Genetics CancerNext FoundationOne™ Ion AmpliSeq Cancer Hotspot v2 TruSeq Amplicon Cancer Panel	Ad5CMV-p53 gene [146]
TRRAP	5		
ZNF521	6		

[a]*Source*: COSMIC [149]

[b]Unless mentioned otherwise, the source of information for potential therapeutic intervention is [150]

PIK3CA, KEAP1, HRAS, SMAD4 (SMAD family member 4)] most commonly mutated in SQCCs. The TP53 mutations were present in almost 93 % of samples whereas the CDKN2A gene was inactivated in 72 % of samples in the study. Moreover, 96 % of tumors were shown to have mutations in tyrosine kinases (e.g., ERBBs, FGFRs, and JAKs), serine/threonine kinases, PI3K, GPCRs (G protein coupled receptors), proteases, and tyrosine phosphatases, suggesting these to be potentially new therapeutic targets in lung SQCC.

In another fairly large study, Imielinski et al. [37] analyzed 183 lung adenocarcinoma tumor/normal pairs using the whole-exome sequencing or whole-genome sequencing approach. The study reported the presence of mutations in a number of known as well as novel lung cancer genes at different frequencies; TP53 (50 %), KRAS (27 %), EGFR (17 %), STK11 (15 %), KEAP1 (12 %), NF1 (11 %), BRAF (8 %), RBM2 (RNA binding motif protein, Y-linked, family 1, member A1; 7 %), U2AF1 (U2 small nuclear RNA auxiliary factor 1; 4 %), and SMAD4 (3 %). In addition to single base mutation, the study also identified large number of structural variations in the genomes of lung cancer patients.

It is clear from all the studies mentioned here that even though environmental factors appear to be primary drivers for lung cancer, it is a genetically heterogeneous disease. Many common as well as specific genome variations (single base and large structural variations) are found in different cohorts and different subtypes of lung cancer. Importantly, many of these identified variants are of predictive value. Continued use of comprehensive genome as well as transcriptome sequencing in large number of lung cancer samples will enable discovery of many valuable biomarkers with application in therapeutics.

5.3 Colorectal Cancer

Colorectal cancer is the fourth leading cause of cancer mortality worldwide with an estimated 694,000 deaths per year [102]. In the USA, until mid-2014, approximately 136,000 individuals were diagnosed with colon cancer and there were about 50,000 deaths, contributing to 8.5 % of all cancer deaths [110]. The two most common inherited syndromes linked with colorectal cancers are familial adenomatous polyposis (FAP) and hereditary non-polyposis colorectal cancer (HNPCC). Only 62 % individuals diagnosed with colon cancer survive for 5 years or more after diagnosis. Prognosis for patients with colorectal cancer is directly related to the timing of diagnosis. If detected early, colorectal cancer (CRC) can be managed by surgery. In the last decade there has been a steady decline in the number of CRC cases due to early detection and surgical removal of polyps.

Colorectal cancer results from accumulation of genetic mutations as well as epigenetic changes and about 5–10 % individuals that develop CRC have inherited genetic defects. In addition, a majority of sporadic CRCs harbor chromosomal instability characterized by aneuploidy, amplifications and deletions of genomic

regions, and loss of heterozygosity (LOH), microsatellite instability, and CpG Island Methylator Phenotype (CIMP) [111]. Specifically, somatic mutations in APC (Adenomatous Polyposis Coli), BRAF, KRAS, PIK3CA, TP53, and other genes have been frequently observed in CRC.

A comprehensive integrated analysis by the TCGA group [81] of 224 colorectal tumor/normal pairs using WGS and WES provides a number of insights into the biology of CRC and identifies potential therapeutic targets. The group of samples studied had a significant variation in the mutation rate between the tumor samples and the authors classified the tumors as non-hypermutated (mutation rate $\ll 1/10^6$ bases) and hypermutated (mutations rates $>100/10^6$). Interestingly, the study also reported prevalence of mutations in different sets of genes between the two groups. In the non-hypermutated set of tumors, APC, TP53, KRAS, PIK3CA, FBXW7 (F-box and WD repeat domain containing 7, E3 ubiquitin protein ligase), SMAD4, TCF7L2 (Transcription Factor 7-Like 2), and NRAS were found to be most frequently mutated. On the other hand, ACVR2A (Activin A receptor, type IIA), APC, TGFBR2 (Transforming Growth Factor, beta receptor II), MSH3 (MutS homolog 3), MSH6 (MutS homolog 6), SLC9A9 (Solute Carrier family 9, subfamily A) and TCF7L2 were frequently mutated in hypermutated tumors. Though mutations in APC, TP53, TGFBR1 (Transforming Growth Factor, beta receptor 1), TGFBR2 (Transforming Growth Factor, beta receptor 2), ACVR2A, ACVR1B (Activin A receptor, type IB), SMAD2 (SMAD family member 2), SMAD3 (SMAD family member 3), and SMAD4 were found in both groups of CRCs, there was a significant difference in the frequency of mutations between the two groups. This study suggests that different molecular signatures can be used to identify different classes of CRCs and have potentially different therapeutic targets.

Han et al. [112] have also demonstrated the feasibility of using a NGS-based panel for identifying mutations in colorectal cancer samples in a clinical setting. They created a panel of 183 genes that had predictive as well as prognostic value and were found to have high mutation frequency in the Catalogue of Somatic Mutations in Cancer (COSMIC) database. They used this panel to sequence the target region in 60 colorectal cancer patients representing different cancer stages as well as different levels of microsatellite instability. The authors reported 166 novel mutations, among which were two recurrent novel mutations, JAK1 (Janus Kinase 1) c.1595C>T (p.R532H) in two patients and EWSR1 (EWS RNA-binding protein 1) c.1769A>C (p.Q590P) in two patients. Point mutations were most frequently observed in genes well established to play a role in colorectal cancer; APC (32 mutations in 29 patients), followed by TP53 (27 in 27), KRAS (24 in 24), TTN (Titin; 36 in 21), and FBXW7 (15 in 14). This study demonstrates the utility of NGS panel in detection of known as well as potentially actionable novel molecular markers in clinical practice.

Similar to breast cancer and lung cancer, the studies conducted in colorectal cancer so far indicate that it is a genetically complex disorder with many significant genetic and epigenetic differences depending on the microsatellite stability status of the tumors. Moreover, very few patients benefit from chemotherapy and hence it is important to identify biomarkers that can be used to identify such patients using NGS technology.

6 The Cancer Genome Atlas Project

Any publication about the phenomenal contribution of novel high-throughput genomic technologies to biomarker identification and resultant impact on cancer characterization, treatment, and management would be incomplete without a brief discussion of The Cancer Genome Atlas project [113].

The Cancer Genome Atlas (TCGA) project is a pan cancer initiative undertaken in 2006 as an exploratory three year project by the National Cancer Institute (NCI) and National Human Genome Research Institute (NHGRI). Currently, the TCGA project has genomic data for more than 30 cancer types with Glioblastoma Multiforme (GBM), Breast Cancer, Ovarian Cancer, and Lung Cancer being the most comprehensive datasets. The mission, strategy, and details of genomic data available for the various cancers studied under the TCGA are described in a recent paper by the TCGA Research Network [114]. The work done in TCGA has helped identify novel biomarkers for GBM [115, 116], lung cancer [117], and ovarian cancer [80, 118]. Most of the studies published by the TCGA group have used data from genome sequencing, transcriptome sequencing, and methylation profiling to derive a very comprehensive landscape of cancer genomics. The TCGA and analysis will substantially contribute towards the development of novel biomarker assays, drugs as well as cancer management regimen.

7 Application of NGS-Based Tests in Therapeutic Intervention

In a very recent report Subbiah et al. [41] demonstrated the application of NGS technology to identify targeted therapy for a spindle cell neoplasm patient non-responsive to standard chemotherapy. A 55-year-old female patient diagnosed with malignant spindle cell neoplasm was initially treated with doxorubicin and ifosfamide followed by gemcitabine and docetaxel. However, the patient did not respond to this therapeutic regimen and was put on a combination of sorafenib (BRAF inhibitor), temsirolimus (mTOR inhibitor), and bevacizumab. This combination therapy resulted in a 25 % reduction in tumor size and decrease in chest pain as well as dyspnea. Genomic profiling of the patient using the NGS-based FoundationOne™ test revealed a KIAA1549-BRAF fusion for the first time in spindle cell neoplasm. The KIAA1549-BRAF fusion has a completely conserved kinase domain and the authors speculate that as sorafenib is a BRAF inhibitor that acts by binding to the ATP binding pocket in BRAF, the therapy was effective in this patient. Additionally, the authors suggest that as the KIAA1549-BRAF fusion has also been shown to hyperactivate the mTOR pathway in mouse models of pilocytic astrocytoma, the patient benefited from including temsirolimus, the mTOR inhibitor, in the combination therapy. This case highlights that genomic profiling using NGS can uncover novel mutations in cancers that can be used for selecting appropriate targeted therapy.

In addition to the aforementioned case study, several clinical trials are underway to test the potential of using NGS-based genomic screening to identify targeted therapies in cancer. For example, a clinical trial of CancerCode is underway to evaluate the effectiveness of targeted therapy chosen based on the genetic information in treating stage IIIB-IV NSCLC patients [119].

CancerCode is a NGS-based test that determines genetic alterations in a select group of cancer genes. Some other examples of clinical trials evaluating the impact of NGS-based testing in selecting targeted therapy are trial of FoundationOne™ test in guiding therapy in recurrent or metastatic solid tumors [120] and trial of proteomic and NGS-based genomic profiling in metastatic breast cancer [121].

8 Current Challenges

1. Cost—NGS panels sequencing small genomic regions cost a few thousand dollars unlike the traditional single gene tests which cost only a few hundred dollars. For example, the FoundationOne™ cancer panel that tests for entire coding sequence of 315 known oncogenes and few introns from 28 genes often rearranged or mutated in solid tumor cancers costs $5,800 per test 9 [122]. The NGS-based cancer tests (10–30 genes) from Ambry genetics costs about $4,000 per test [123]. On the other hand, the BRACAnalysis® Large Rearrangement Test (BART™) from Myriad Genetics that interrogates for mutations in BRCA1 and BRCA2 genes costs only $700 per test [124] and the single BRCA1 or BRCA2 test by Ambry genetics costs about $500 [123]. However, what needs to be noted is that even though the per base cost for an NGS panel is lower than a traditional test, the total out of pocket cost is higher. Moreover, currently not all NGS-based tests are reimbursed by insurance providers and that poses a significant challenge for widespread adoption.
2. Turnaround time—Typical turnaround time for NGS-based tests is about 4–6 weeks [125, 126] as opposed to 7–10 days for single gene tests [126]. The longer turnaround time is a significant challenge in cases where decision about course of treatment needs to be made at the earliest. This is usually the case as NGS tests are ordered when mutations expected in the particular cancer (e.g., BRCA1/2 mutations in breast/ovarian cancer) are not detected or when a new line of treatment is being considered.
3. Variants of unknown significance—Variants of unknown significance (VUS) are defined as DNA variants that have not been well characterized for their functional impact. On using NGS to sequence large portion of genome, a significant number of VUS are likely to be identified. Presence of VUS in genomic regions important for the disease under consideration could create substantial uncertainty in deciding future course of action and hence sometimes could be counterproductive. Tests used for biomarker identification are expected to yield precise and specific information that can be reliably used for guiding disease management and the VUS identified by NGS pose a significant challenge for its utility as a standard test.

4. Complex data analysis and interpretation—In traditional biomarker tests or single gene tests clear guidance with regard to the expected level of the biomarker or expected mutation are available. However, for NGS-based tests the likelihood of detecting a novel or unexpected mutation are high even though they often have an expected outcome. This requires that specially trained bioinformatics or medical geneticists are available for data analysis and interpretation.
5. No one test that fits all—One of the classic properties of biomarkers is that, it is detectable consistently and reliably in all tested samples. However, the biomarkers expected to be discovered and eventually used in clinic using NGS technology are likely to be such that they are found only in a specific subset of tested samples. This implies that for the same disease type (e.g., lung cancer), different markers will test positive depending on the cancer subtype. While this is not necessarily a limitation, it is certainly different from the current biomarker paradigm.

9 Future Directions

Biomarker detection using NGS technology is likely to play a very crucial role in cancer diagnosis, prognosis and disease management in the very near future. Biomarkers detected using NGS-based tests will be very different from the traditional biomarkers in that the biomarkers may not be specific to a cancer type, but rather specific to a subset of cancer patients. The data from TCGA analysis is already suggesting that many of the oncogenes that are associated with a particular cancer type are not mutated in all the patients diagnosed with that cancer. For example, association between mutations in TP53 and lung cancer is well established; however, TP53 mutations are not found in all the patients diagnosed with lung cancer. This phenomenon is not unique to lung cancer, but is prevalent in all forms of cancer. Apart from using NGS for detection of diagnostic and prognostic biomarkers, NGS-based tests can be used as a sensitive assay to predict the metastatic potential or probability of relapse. For example Fu et al. [127] recently used NGS to sequence genomes of CML patients who received allogeneic stem-cell transplant and demonstrated a higher incidence of CML relapse in patients carrying mutations in ASXL1, CBL, TET2, or NRAS. Availability of this information early in the treatment cycle will enable inclusion of preventive measures. Moreover, as metastasized tumors tend to acquire new mutations, sequencing can be performed serially on DNA obtained from blood. In this context, the recently demonstrated feasibility of detecting circulating tumor DNA and using it to perform molecular characterization of cancer is very exciting.

10 Summary

In the last few years, NGS-based studies have made immense contribution towards discovering many novel and clinically relevant biomarkers. One of the most important contribution of NGS in cancer research has been the capability to decipher individual cancer genome and truly unleash the potential of human genome in driving personalized cancer care.

References

1. Maynex R. Biomarkers: potential uses and limitations. NeuroRx. 2004;1(2):182–8.
2. Oldenhuis CN, Oosting SF, Gietema JA, de Vries EG. Prognostic versus predictive value of biomarkers in oncology. Eur J Cancer. 2008;44(7):946–53.
3. Galanis E, Wu W, Sarkaria J, Chang SM, Colman H, Sargent D, Reardon DA. Incorporation of biomarker assessment in novel clinical trial designs: personalizing brain tumor treatments. Curr Oncol Rep. 2011;13(1):42–9.
4. Atkinson AJ, Colburn WA, DeGruttola VG. Biomarkers and surrogate end points: preferred definitions and conceptual framework. Clin Pharmacol Ther. 2001;69(3):89–95.
5. Frank R, Hargreaves R. Clinical biomarkers in drug discovery and development. Nat Rev Drug Discov. 2003;2:566–80.
6. Bell R, Petticrew M, Luengo S, Sheldon TA. Screening for ovarian cancer: a systematic review. Health Technol Assess. 1998;2(2):1–84.
7. Suh KS, Park SW, Castro A, Patel H, Blake P, Liang M, Goy A. Ovarian cancer biomarkers for molecular biosensors and translational medicine. Expert Rev Mol Diagn. 2010;10(8): 1069–83.
8. Martignetti JA, Camacho-Vanegas O, Priedigkeit N, Camacho C, Pereira E, Lin L, et al. Personalized ovarian cancer disease surveillance and detection of candidate therapeutic drug target in circulating tumor DNA. Neoplasia. 2014;16(1):97–103.
9. Chan DW, Bruzek DJ, Oesterling JE, Rock RC, Walsh PC. Prostate-specific antigen as a marker for prostatic cancer: a monoclonal and a polyclonal immunoassay compared. Clin Chem. 1987;33(10):1916–20.
10. Catalona WJ, Smith DS, Ratliff TL, Dodds KM, Coplen DE, Yuan JJ, et al. Measurement of prostate-specific antigen in serum as a screening test for prostate cancer. N Engl J Med. 1991;324(17):1156–61.
11. Fass L. Imaging and cancer: a review. Mol Oncol. 2008;2:115–52.
12. Prensner JR, Rubin MA, Wei JT, Chinnaiyan AM. Beyond PSA: the next generation of prostate cancer biomarkers. Sci Transl Med. 2012;4(127):127.
13. Goodison S, Rosser CJ, Urquidi V. Bladder cancer detection and monitoring: assessment of urine- and blood-based marker tests. Mol Diagn Ther. 2013;17(2):71–84.
14. Garraway LA, Verweij J, Ballman KV. Precision oncology: an overview. J Clin Oncol. 2013;31(15):1803–5.
15. Heinrich MC, Corless CL, Demetri GD, Blanke CD, von Mehren M, Joensuu H, et al. Kinase mutations and imatinib response in patients with metastatic gastrointestinal stromal tumor. J Clin Oncol. 2003;21(23):4342–9.
16. Jia Y, Ali SM, Saad S, Chan CA, Miller VA, Halmos B. Successful treatment of a patient with Li-Fraumeni syndrome and metastatic lung adenocarcinoma harboring synchronous EGFR L858R and ERBB2 extracellular domain S310F mutations with the pan-HER inhibitor afatinib. Cancer Biol Ther. 2014;15(8):970–4.
17. O'Shaughnessy J, Osborne C, Pippen JE, Yoffe M, Patt D, Rocha C, et al. Iniparib plus chemotherapy in metastatic triple-negative breast cancer. N Engl J Med. 2011;364(3):205–14.
18. Thomas A, Rajan A, Lopez-Chavez A, Wang Y, Giaccone G. From targets to targeted therapies and molecular profiling in non-small cell lung carcinoma. Ann Oncol. 2013;24:577–85.
19. Druker BJ, Talpaz M, Resta DJ, et al. Efficacy and safety of a specific inhibitor of the BCR-ABL tyrosine kinase in chronic myeloid leukemia. N Engl J Med. 2001;344(14):1031–7.
20. Gambacorti-Passerini C, Antolini L, Mahon F-X, Guilhot F, Deininger M, Saglio G, et al. Multicenter independent assessment of outcomes in chronic myeloid leukemia patients treated with imatinib. J Natl Cancer Inst. 2011;103:553–61.
21. Vogel CL, Cobleigh MA, Tripathy D, Gutheil JC, Harris LN, Fehrenbacher L, et al. Efficacy and safety of trastuzumab as a single agent in first-line treatment of HER2-overexpressing metastatic breast cancer. J Clin Oncol. 2002;20(3):719–26.
22. Emens LA, Davidson NE. Trastuzumab in breast cancer. Oncology (Williston Park). 2004;18(9):1117–28.

23. Mok TS, Wu YL, Thongprasert S, Yang CH, Chu DT, Saijo N, et al. Gefitinib or carboplatin-paclitaxel in pulmonary adenocarcinoma. N Engl J Med. 2009;361(10):947–57.
24. Heakal Y, Kester M, Savage S. Vemurafenib (PLX4032): an orally available inhibitor of mutated BRAF for the treatment of metastatic melanoma. Ann Pharmacother. 2011;45(11):1399–405.
25. Desai AN, Jere A. Next generation sequencing for cancer genomics. Next Gen Sequence Cancer Res. 2013;1:55–74.
26. Mardis ER. Next-generation DNA sequencing methods. Annu Rev Genomics. 2008;9:387–402.
27. Nowrousian M. Next-generation sequencing techniques for eukaryotic microorganisms: sequencing-based solutions to biological problems. Eukaryot Cell. 2010;1300–10.
28. Martínez E, Yoshihara K, Kim H, Mills GM, Treviño V, Verhaak RG. Comparison of gene expression patterns across 12 tumor types identifies a cancer supercluster characterized by TP53 mutations and cell cycle defects. Oncogene. 2014. doi:10.1038/onc.2014.216.
29. Yin ZQ, Liu JJ, Xu YC, Yu J, Ding GH, Yang F, et al. A 41-gene signature derived from breast cancer stem cells as a predictor of survival. J Exp Clin Cancer Res. 2014;33:49.
30. Lee U, Frankenberger C, Yun J, Bevilacqua E, Caldas C, Chin SF, et al. A prognostic gene signature for metastasis-free survival of triple negative breast cancer patients. PLoS One. 2013;8(12):e82125.
31. Bertucci F, Finetti P, Viens P, Birnbaum D. EndoPredict predicts for the response to neoadjuvant chemotherapy in ER-positive, HER2-negative breast cancer. Cancer Lett. 2014;pii:S0304-3835(14)00513-8.
32. Peng Z, Skoog L, Hellborg H, Jonstam G, Wingmo IL, Hjälm-Eriksson M, et al. An expression signature at diagnosis to estimate prostate cancer patients' overall survival. Prostate Cancer Prostatic Dis. 2014;17(1):81–90.
33. Berger MF, Lawrence MS, Demichelis F, Drier Y, Cibulskis K, Sivachenko AY, et al. The genomic complexity of primary human prostate cancer. Nature. 2011;470(7333):214–20.
34. Fujimoto A, Totoki Y, Abe T, Boroevich KA, Hosoda F, Nguyen HH, et al. Whole-genome sequencing of liver cancers identifies etiological influences on mutation patterns and recurrent mutations in chromatin regulators. Nat Genet. 2012;44(7):760–4.
35. Wagle N, Berger MF, Davis MJ, Blumenstiel B, Defelice M, Pochanard P, et al. High-throughput detection of actionable genomic alterations in clinical tumor samples by targeted, massively parallel sequencing. Cancer Discov. 2012;2(1):82–93.
36. Brastianos PK, Horowitz PM, Santagata S, Jones RT, McKenna A, Getz G, et al. Genomic sequencing of meningiomas identifies oncogenic SMO and AKT1 mutations. Nat Genet. 2013;45(3):285–9.
37. Imielinski M, Berger AH, Hammerman PS, Hernandez B, Pugh TJ, Hodis E, et al. Mapping the hallmarks of lung adenocarcinoma with massively parallel sequencing. Cell. 2012; 150(6):1107–20.
38. Varela I, Tarpey P, Raine K, Huang D, Ong CK, Stephens P, et al. Exome sequencing identifies frequent mutation of the SWI/SNF complex gene PBRM1 in renal carcinoma. Nature. 2011;469(7331):539–42.
39. Makishima H, Jankowska AM, Tiu RV, Szpurka H, Sugimoto Y, Hu Z, et al. Novel homo- and hemizygous mutations in EZH2 in myeloid malignancies. Leukemia. 2010;24(10):1799–804.
40. Chong HK, Wang T, Lu HM, Seidler S, Lu H, Keiles S, et al. The validation and clinical implementation of BRCAplus: a comprehensive high-risk breast cancer diagnostic assay. PLoS One. 2014;9(5):e97408.
41. Subbiah V, Westin SN, Wang K, Araujo D, Wang WL, Miller VA, et al. Targeted therapy by combined inhibition of the RAF and mTOR kinases in malignant spindle cell neoplasm harboring the KIAA1549-BRAF fusion protein. J Hematol Oncol. 2014;7(1):8.
42. Frampton GM, Fichtenholtz A, Otto GA, Wang K, Downing SR, He J, et al. Development and validation of a clinical cancer genomic profiling test based on massively parallel DNA sequencing. Nat Biotechnol. 2013;31(11):1023–31.
43. King MC, Marks JH, Mandell JB, New York Breast Cancer Study Group. Breast and ovarian cancer risks due to inherited mutations in BRCA1 and BRCA2. Science. 2003;302(5645): 643–6.

44. James CR, Quinn JE, Mullan PB, Johnston PG, Harkin DP. BRCA1, a potential predictive biomarker in the treatment of breast cancer. Oncologist. 2007;12(2):142–50.
45. Bazan V, Migliavacca M, Zanna I, Tubiolo C, Grassi N, Latteri MA, et al. Specific codon 13 K-ras mutations are predictive of clinical outcome in colorectal cancer patients, whereas codon 12 K-ras mutations are associated with mucinous histotype. Ann Oncol. 2002;13(9): 1438–46.
46. Zafarana G, Ishkanian AS, Malloff CA, Locke JA, Sykes J, Thoms J, et al. Copy number alterations of c-MYC and PTEN are prognostic factors for relapse after prostate cancer radiotherapy. Cancer. 2012;118(16):4053–62.
47. Kurian AW, Hare EE, Mills MA, Kingham KE, McPherson L, Whittemore AS, et al. Clinical evaluation of a multiple-gene sequencing panel for hereditary cancer risk assessment. J Clin Oncol. 2014;32(19):2001–9.
48. Ciriello G, Miller ML, Aksoy BA, Senbabaoglu Y, Schultz N, Sander C. Emerging landscape of oncogenic signatures across human cancers. Nat Genet. 2013;45(10):1127–33.
49. van't Veer LJ, Dai H, van de Vijver MJ, He YD, Hart AA, Mao M, et al. Gene expression profiling predicts clinical outcome of breast cancer. Nature. 2002;415:530–6.
50. Molloy TJ, Roepman P, Naume B, van't Veer LJ. A prognostic gene expression profile that predicts circulating tumor cell presence in breast cancer patients. PLoS One. 2012;7(2):e32426.
51. Larsen MJ, Thomassen M, Tan Q, Lænkholm AV, Bak M, Sørensen KP, et al. RNA profiling reveals familial aggregation of molecular subtypes in non-BRCA1/2 breast cancer families. BMC Med Genomics. 2014;7:9.
52. Bertucci F, Salas S, Eysteries S, Nasser V, Finetti P, Ginestier C, et al. Gene expression profiling of colon cancer by DNA microarrays and correlation with histoclinical parameters. Oncogene. 2004;23(7):1377–91.
53. Marisa L, de Reyniès A, Duval A, Selves J, Gaub MP, Vescovo L, et al. Gene expression classification of colon cancer into molecular subtypes: characterization, validation, and prognostic value. PLoS Med. 2013;10(5):e1001453.
54. Verhaak RG, Hoadley KA, Purdom E, Wang V, Qi Y, Wilkerson MD, et al. Integrated genomic analysis identifies clinically relevant subtypes of glioblastoma characterized by abnormalities in PDGFRA, IDH1, EGFR, and NF1.Cancer. Cell. 2010;17(1):98–110.
55. http://www.oncotypedx.com/. Accessed 25 Sept 2014
56. Glas AM, Floore A, Delahaye LJ, Witteveen AT, Pover RC, Bakx N, et al. Converting a breast cancer microarray signature into a high-throughput diagnostic test. BMC Genomics. 2006;7:278.
57. http://www.agendia.com/. Accessed 25 Sept 2014
58. Hovanes K, Li TW, Munguia JE, Truong T, Milovanovic T, Lawrence Marsh J, et al. Beta-catenin-sensitive isoforms of lymphoid enhancer factor-1 are selectively expressed in colon cancer. Nat Genet. 2001;28:53–7.
59. Tomasini R, Tsuchihara K, Wilhelm M, Fujitani M, Rufini AA, Cheung CC, et al. Tap73 knockout shows genomic instability with infertility and tumor suppressor functions. Genes Dev. 2008;22:2677–91.
60. Tang J, Xiong Y, Zhou HH, Chen XP. DNA methylation and personalized medicine. J Clin Pharm Ther. 2014. doi:10.1111/jcpt.12206.
61. Dalgliesh GL, Furge K, Greenman C, Chen L, Bignell G, Butler A, et al. Systematic sequencing of renal carcinoma reveals inactivation of histone modifying genes. Nature. 2010;463(7279):360–3.
62. Tan SX, Hu RC, Liu JJ, Tan YL, Liu WE. Methylation of PRDM2, PRDM5 and PRDM16 genes in lung cancer cells. Int J Clin Exp Pathol. 2014;7(5):2305–11.
63. Sonnet M, Claus R, Becker N, Zucknick M, Petersen J, Lipka DB, et al. Early aberrant DNA methylation events in a mouse model of acute myeloid leukemia. Genome Med. 2014;6(4):34.
64. Ogino S, Nishihara R, Lochhead P, Imamura Y, Kuchiba A, Morikawa T, et al. Prospective study of family history and colorectal cancer risk by tumor LINE-1 methylation level. J Natl Cancer Inst. 2013;105(2):130–40.
65. Cairns P, Esteller M, Herman JG, Schoenberg M, Jeronimo C, Sanchez-Cespedes M, et al. Molecular detection of prostate cancer in urine by GSTP1 hypermethylation. Clin Cancer Res. 2001;7(9):2727–30.

66. Eilers T, Machtens S, Tezval H, Blaue C, Lichtinghagen R, Hagemann J, et al. Prospective diagnostic efficiency of biopsy washing DNA GSTP1 island hypermethylation for detection of adenocarcinoma of the prostate. Prostate. 2007;67(7):757–63.
67. Catto JW, Azzouzi AR, Rehman I, Feeley KM, Cross SS, Amira N, et al. Promoter hypermethylation is associated with tumor location, stage, and subsequent progression in transitional cell carcinoma. J Clin Oncol. 2005;23(13):2903–10.
68. Jarmalaite S, Jankevicius F, Kurgonaite K, Suziedelis K, Mutanen P, Husgafvel-Pursiainen K. Promoter hypermethylation in tumour suppressor genes shows association with stage, grade and invasiveness of bladder cancer. Oncology. 2008;75(3–4):145–51.
69. Wasserkort R, Kalmar A, Valcz G, Spisak S, Krispin M, Toth K, et al. Aberrant septin 9 DNA methylation in colorectal cancer is restricted to a single CpG island. BMC Cancer. 2013;13:398.
70. Mah WC, Thurnherr T, Chow PK, Chung AY, Ooi LL, Toh HC, et al. Methylation profiles reveal distinct subgroup of hepatocellular carcinoma patients with poor prognosis. PLoS One. 2014;9(8):e104158.
71. Taby R, Issa JP. Cancer epigenetics. Cancer J Clin. 2010;60(6):376–92.
72. Baek D, Villen J, Shin C, Camargo FD, Gygi SP, Bartel DP. The impact of microRNAs on protein output. Nature. 2008;455:64–71.
73. Calin GA, Dumitru CD, Shimizu M, Bichi R, Zupo S, Noch E, et al. Frequent deletions and down-regulation of micro-RNA genes miR15 and miR16 at 13q14 in chronic lymphocytic leukemia. Proc Natl Acad Sci U S A. 2002;99(24):15524–9.
74. Iorio MV, Ferracin M, Liu CG, Veronese A, Spizzo R, Sabbioni S, et al. MicroRNA gene expression deregulation in human breast cancer. Cancer Res. 2005;65(16):7065–70.
75. Mayr C, Hemann MT, Bartel DP. Disrupting the pairing between let-7 and Hmga2 enhances oncogenic transformation. Science. 2007;315(5818):1576–9.
76. Zhu ED, Li N, Li BS, Li W, Zhang WJ, Mao XH, et al. Mir-30b, down-regulated in gastric cancer, promotes apoptosis and suppresses tumor growth by targeting plasminogen activator inhibitor-1. PLoS One. 2014;9(8):e106049.
77. Ling H, Spizzo R, Atlasi Y, Nicoloso M, Shimizu M, Redis RS, et al. CCAT2, a novel noncoding RNA mapping to 8q24, underlies metastatic progression and chromosomal instability in colon cancer. Genome Res. 2013;23(9):1446–61.
78. Li Y, Hong F, Yu Z. Decreased expression of microRNA-206 in breast cancer and its association with disease characteristics and patient survival. J Int Med Res. 2013;41(3):596–602.
79. Anfossi S, Giordano A, Gao H, Cohen EN, Tin S, Wu Q, et al. High serum miR-19a levels are associated with inflammatory breast cancer and are predictive of favorable clinical outcome in patients with metastatic HER2+ inflammatory breast cancer. PLoS One. 2014; 9(1):e83113.
80. The Cancer Genome Atlas Research Network. Integrated genomic analyses of ovarian carcinoma. Nature. 2011;474(7353):609–15.
81. The Cancer Genome Atlas Network. Comprehensive molecular characterization of human colon and rectal cancer. Nature. 2012;487:330–7.
82. The Cancer Genome Atlas Network. Comprehensive molecular portraits of human breast tumors. Nature. 2012;490(7418):61–70.
83. The Cancer Genome Atlas Research Network. Comprehensive genomic characterization of squamous cell lung cancers. Nature. 2012;489(7417):519–25.
84. Hoadley KA, Yau C, Wolf DM, Cherniack AD, Tamborero D, Ng S, et al. Multiplatform analysis of 12 cancer types reveals molecular classification within and across tissues of origin. Cell. 2014;158(4):929–44.
85. http://clincancerres.aacrjournals.org/content/early/2014/09/16/1078–0432.CCR-14–2123.2.full.pdf. Accessed 25 Sept 2014
86. Carroll JC, Cremin C, Allanson J, Blaine SM, Dorman H, Gibbons CA, et al. Hereditary breast and ovarian cancers. Can Fam Phys. 2008;54(12):1691–2.
87. Apostolou P, Fostira F. Hereditary breast cancer: the era of new susceptibility genes. Biomed Res Int. 2013;2013:747318.

88. Malkin D, Li FP, Strong LC, Fraumeni Jr JF, Nelson CE, Kim DH, et al. Germ line p53 mutations in a familial syndrome of breast cancer, sarcomas, and other neoplasms. Science. 1990;250(4985):1233–8.
89. Li FP, Fraumeni JF, Mulvihill JJ, Blattner WA, Dreyfus MG, Tucker MA, et al. A cancer family syndrome in twenty four kindred. Cancer Res. 1988;48:5358–62.
90. Varley JM, McGown G, Thorncroft M, Santibanez-Koref MF, Kelsey AM, Tricker KJ, et al. Germ-line mutations of TP53 in Li-Fraumeni families: an extended study of 39 families. Cancer Res. 1997;57(15):3245–52.
91. Pradella LM, Evangelisti C, Ligorio C, Ceccarelli C, Neri I, Zuntini R, et al. A novel deleterious PTEN mutation in a patient with early-onset bilateral breast cancer. BMC Cancer. 2014;14:70.
92. Nakanishi C, Yamaguchi T, Iijima T, Saji S, Toi M, Mori T, Miyaki M. Germline mutation of the LKB1/STK11 gene with loss of the normal allele in an aggressive breast cancer of Peutz-Jeghers syndrome. Oncology. 2004;67(5–6):476–9.
93. Hearle N, Schumacher V, Menko FH, Olschwang S, Boardman LA, Gille JJ, et al. Frequency and spectrum of cancers in the Peutz-Jeghers syndrome. Clin Cancer Res. 2006;12(10):3209–15.
94. Walsh T, King MC. Ten genes for inherited breast cancer. Cancer Cell. 2007;11(2):103–5.
95. Walsh T, Casadei S, Coats KH, Swisher E, Stray SM, Higgins J, et al. Spectrum of mutations in BRCA1, BRCA2, CHEK2, and TP53 in families at high risk of breast cancer. JAMA. 2006;295(12):1379–88.
96. Walsh T, Lee MK, Casadei S, Thornton AM, Stray SM, Pennil C, et al. Detection of inherited mutations for breast and ovarian cancer using genomic capture and massively parallel sequencing. Proc Natl Acad Sci U S A. 2010;107(28):12629–33.
97. Ozcelik H, Shi X, Chang MC, Tram E, Vlasschaert M, Di Nicola N, et al. Long-range PCR and next-generation sequencing of BRCA1 and BRCA2 in breast cancer. J Mol Diagn. 2012;14(5):467–75.
98. Feliubadaló L, Lopez-Doriga A, Castellsagué E, del Valle J, Menéndez M, Tornero E, et al. Next-generation sequencing meets genetic diagnostics: development of a comprehensive workflow for the analysis of BRCA1 and BRCA2 genes. Eur J Hum Genet. 2013;21(8):864–70.
99. Castéra L, Krieger S, Rousselin A, Legros A, Baumann JJ, Bruet O, et al. Next-generation sequencing for the diagnosis of hereditary breast and ovarian cancer using genomic capture targeting multiple candidate genes. Eur J Hum Genet. 2014. doi:10.1038/ejhg.2014.16.
100. Balko JM, Giltnane JM, Wang K, Schwarz LJ, Young CD, Cook RS, et al. Molecular profiling of the residual disease of triple-negative breast cancers after neoadjuvant chemotherapy identifies actionable therapeutic targets. Cancer Discov. 2014;4(2):232–45.
101. Link DC, Schuettpelz LG, Shen D, Wang J, Walter MJ, Kulkarni S, et al. Identification of a novel TP53 cancer susceptibility mutation through whole-genome sequencing of a patient with therapy-related AML. JAMA. 2011;305(15):1568–76.
102. http://www.who.int/mediacentre/factsheets/fs297/en/. Accessed 25 Sept 2014
103. Hecht SS. Tobacco smoke carcinogens and lung cancer. J Natl Cancer Inst. 1999;91:1194–210.
104. Alberg AJ, Brock MV, Ford JG, Samet JM, Spivack SD. Epidemiology of lung cancer: diagnosis and management of lung cancer, 3rd ed: American College of Chest Physicians evidence-based clinical practice guidelines. Chest. 2013;143(5 Suppl):e1S–29.
105. Sellers TA, Weaver TW, Phillips B, Altmann M, Rich SS. Environmental factors can confound identification of a major gene effect: results from a segregation analysis of a simulated population of lung cancer families. Genet Epidemiol. 1998;15(3):251–62.
106. Pao W, Miller V, Zakowski M, Doherty J, Politi K, Sarkaria I, et al. EGF receptor gene mutations are common in lung cancers from "never smokers" and are associated with sensitivity of tumors to gefitinib and erlotinib. Proc Natl Acad Sci U S A. 2004;101(36):13306–11.
107. Ding L, Getz G, Wheeler DA, Mardis ER, McLellan MD, Cibulskis K, et al. Somatic mutations affect key pathways in lung adenocarcinoma. Nature. 2008;455(7216):1069–75.

108. Govindan R, Ding L, Griffith M, Subramanian J, Dees ND, Kanchi KL, et al. Genomic landscape of non-small cell lung cancer in smokers and never-smokers. Cell. 2012;150(6): 1121–34.
109. The Cancer Genome Atlas Research Network. Comprehensive molecular profiling of lung adenocarcinoma. Nature. 2014;511:543–50.
110. http://seer.cancer.gov/statfacts/html/colorect.html. Accessed 19 Sept 2014
111. Migliore L, Migheli F, Spisni R, Coppedè F. Genetics, cytogenetics, and epigenetics of colorectal cancer. J Biomed Biotechnol. 2011;792362.
112. Han SW, Kim HP, Shin JY, Jeong EG, Lee WC, Lee KH, et al. Targeted sequencing of cancer-related genes in colorectal cancer using next-generation sequencing. PLoS One. 2013;8(5): e64271.
113. http://cancergenome.nih.gov/. Accessed 20 Sept 2014
114. Cancer Genome Atlas Research Network, Weinstein JN, Collisson EA, Mills GB, Shaw KR, Ozenberger BA. The cancer genome atlas pan-cancer analysis project. Nat Genet. 2013;45(10):1113–20.
115. Li R, Gao K, Luo H, Wang X, Shi Y, Dong Q. Identification of intrinsic subtype-specific prognostic microRNAs in primary glioblastoma. J Exp Clin Cancer Res. 2014;33:9.
116. Kim YW, Koul D, Kim SH, Lucio-Eterovic AK, Freire PR, Yao J, et al. Identification of prognostic gene signatures of glioblastoma: a study based on TCGA data analysis. Neuro Oncol. 2013;15(7):829–39.
117. Wrangle J, Machida EO, Danilova L, Hulbert A, Franco N, Zhang W, et al. Functional identification of cancer-specific methylation of CDO1, HOXA9, and TAC1 for the diagnosis of lung cancer. Clin Cancer Res. 2014;20(7):1856–64.
118. Verhaak RG, Tamayo P, Yang JY, Hubbard D, Zhang H, Creighton CJ, et al. Prognostically relevant gene signatures of high-grade serous ovarian carcinoma. J Clin Invest. 2013;123(1):517–25.
119. https://clinicaltrials.gov/ct2/show/NCT02132884?term=sequencing&recr=Open&cond=cancer&rank=3. Accessed 25 Sept 2014
120. https://clinicaltrials.gov/ct2/show/NCT01987726?term=NCT01987726&rank=1. Accessed 25 Sept 2014
121. https://clinicaltrials.gov/ct2/show/NCT01919749?term=sequencing&recr=Open&cond=cancer&rank=2. Accessed 25 Sept 2014
122. http://www.foundationmedicine.com/pdf/resources/ONE-A-001–20120823%20Billing%20Guidelines_Physicians.pdf. Accessed 23 Sept 2014
123. http://www.ambrygen.com/brca-beyond. Accessed 23 Sept 2014
124. https://www.myriad.com/lib/brac/BART-faq.pdf. Accessed 23 Sept 2014
125. http://FoundationOne™.com/learn.php#2. Accessed 23 Sept 2014
126. http://www.ambrygen.com/turn-around-times. Accessed 23 Sept 2014
127. Fu Y, Schroeder T, Zabelina T, Badbaran A, Bacher U, Kobbe G, et al. Postallogeneic monitoring with molecular markers detected by pretransplant next-generation or Sanger sequencing predicts clinical relapse in patients with myelodysplastic/myeloproliferative neoplasms. Eur J Haematol. 2014;92(3):189–94.
128. Shah SP, Roth A, Goya R, Oloumi A, Ha G, Zhao Y, et al. The clonal and mutational evolution spectrum of primary triple-negative breast cancers. Nature. 2012;486(7403):395–9.
129. Stephens PJ, Tarpey PS, Davies H, Van Loo P, Greenman C, Wedge DC, et al. The landscape of cancer genes and mutational processes in breast cancer. Nature. 2012;486(7403):400–4.
130. Brennan CW, Verhaak RG, McKenna A, Campos B, Noushmehr H, Salama SR, et al. The somatic genomic landscape of glioblastoma. Cell. 2013;155(2):462–77.
131. Holbrook JD, Parker JS, Gallagher KT, Halsey WS, Hughes AM, Weigman VJ, et al. Deep sequencing of gastric carcinoma reveals somatic mutations relevant to personalized medicine. J Transl Med. 2011;9:119.
132. Yu J, Wu WK, Li X, He J, Li XX, Ng SS, et al. Novel recurrently mutated genes and a prognostic mutation signature in colorectal cancer. Gut. 2014;pii:gutjnl-2013–306620.

133. Wei X, Walia V, Lin JC, Teer JK, Prickett TD, Gartner J, et al. Exome sequencing identifies GRIN2A as frequently mutated in melanoma. Nat Genet. 2011;43(5):442–6.
134. Jones S, Wang TL, Shih IM, Mao TL, Nakayama K, Roden R, et al. Frequent mutations of chromatin remodeling gene ARID1A in ovarian clear cell carcinoma. Science. 2010;330(6001):228–31.
135. Ley TJ, Mardis ER, Ding L, Fulton B, McLellan MD, Chen K, et al. DNA sequencing of a cytogenetically normal acute myeloid leukemia genome. Nature. 2008;456(7218):66–72.
136. Mardis ER, Ding L, Dooling DJ, Larson DE, McLellan MD, Chen K, et al. Recurring mutations found by sequencing an acute myeloid leukemia genome. N Engl J Med. 2009;361(11):1058–66.
137. Brohl AS, Solomon DA, Chang W, Wang J, Song Y, Sindiri S, et al. The genomic landscape of the Ewing sarcoma family of tumors reveals recurrent STAG2 mutation. PLoS Genet. 2014;10(7):e1004475.
138. Pinton G, Manente AG, Angeli G, Mutti L, Moro L. Perifosine as a potential novel anticancer agent inhibits EGFR/MET-AKT axis in malignant pleural mesothelioma. PLoS One. 2012;7(5):e36856.
139. Pal SK, Reckamp K, Yu H, Figlin RA. Akt inhibitors in clinical development for the treatment of cancer. Expert Opin Investig Drugs. 2010;19(11):1355–66.
140. Ross JS, Wang K, Rand JV, Gay L, Presta MJ, Sheehan CE, et al. Next-generation sequencing of adrenocortical carcinoma reveals new routes to targeted therapies. J Clin Pathol. 2014;pii: jclinpath-2014–202514.
141. Hamilton G, Lukas Klameth U, Ulsperger E, Geissler K. Synergistic anticancer activity of topotecan-cyclin-dependent kinase inhibitor combinations against drug-resistant small cell lung cancer (SCLC) cell lines. J Cancer Ther. 2013;4:47–53.
142. Drew Y, Mulligan EA, Vong WT, Thomas HD, Kahn S, Kyle S, et al. Therapeutic potential of poly(ADP-ribose) polymerase inhibitor AG014699 in human cancers with mutated or methylated BRCA1 or BRCA2. J Natl Cancer Inst. 2011;103(4):334–46.
143. Sandhu SK, Schelman WR, Wilding G, Moreno V, Baird RD, Miranda S, et al. The poly(ADP-ribose) polymerase inhibitor niraparib (MK4827) in BRCA mutation carriers and patients with sporadic cancer: a phase 1 dose-escalation trial. Lancet Oncol. 2013;14(9):882–92.
144. To C, Kim EH, Royce DB, Williams CR, Collins RM, Risingsong R, et al. The PARP inhibitors, veliparib and olaparib, are effective chemopreventive agents for delaying mammary tumor development in BRCA1-deficient mice. Cancer Prev Res (Phila). 2014;7(7):698–707.
145. Cristofanilli M, Krishnamurthy S, Guerra L, Broglio K, Arun B, Booser DJ, et al. A nonreplicating adenoviral vector that contains the wild-type p53 transgene combined with chemotherapy for primary breast cancer: safety, efficacy, and biologic activity of a novel gene-therapy approach. Cancer. 2006;107(5):935–44.
146. Fujiwara T, Tanaka N, Kanazawa S, Ohtani S, Saijo Y, Nukiwa T, et al. Multicenter phase I study of repeated intratumoral delivery of adenoviral p53 in patients with advanced non-small-cell lung cancer. J Clin Oncol. 2006;24(11):1689–99.
147. Dodd RD, Mito JK, Eward WC, Chitalia R, Sachdeva M, Ma Y, et al. NF1 deletion generates multiple subtypes of soft-tissue sarcoma that respond to MEK inhibition. Mol Cancer Ther. 2013;12(9):1906–17.
148. Klümpen HJ, Queiroz KC, Spek CA, van Noesel CJ, Brink HC, de Leng WW, et al. mTOR inhibitor treatment of pancreatic cancer in a patient With Peutz-Jeghers syndrome. J Clin Oncol. 2011;29(6):e150–3.
149. http://cancer.sanger.ac.uk/cosmic/browse/tissue#sn=breast&ss=all&hn=all&sh=all&in=t&src=tissue&all_data=n. Accessed 29 Sept 2014
150. http://www.mycancergenome.org/content/other/molecular-medicine/anticancer-agents/. Accessed 29 Sept 2014

Validation and Implementation of Next-Generation Sequencing Technologies in a Clinical Molecular Diagnostic Laboratory

Rajesh R. Singh and Rajyalakshmi Luthra

Abstract Next-generation sequencing (NGS) technologies represent powerful tools capable of massive parallel sequencing of DNA. Their application has revolutionized characterization of genomic aberrations responsible for initiation and maintenance of cancers, resulting in a high discovery rate of therapeutic and prognostic markers. In a clinical molecular diagnostic laboratory, the high sequencing capacities of NGS technologies are well suited for routine screening of increasing number of markers using low inputs of DNA in high sample volumes. However, implementation of these technologies in the clinical environment needs thorough validation of their workflow suitability, their capability to detect a variety of genomic aberrations, and their efficiency in comparison to other orthogonal sequencing platforms being routinely used. Here, using a targeted NGS panel to screen for mutational hot spots in 46 cancer-related genes as an example, we have discussed various parameters which need to be established for validation and implementation of the NGS assays. We have highlighted various assay performance metrics which need to be established toward complete validation of the NGS platform before implementation. The criteria for filtering, annotating, and clinical reporting of variants are also discussed.

1 Introduction

Cancer is a heterogeneous disease characterized by the accumulation of stable DNA sequence abnormalities resulting in deregulation of multiple cell signaling pathways. Comprehensive profiling of these genetic changes is valuable in diagnosis, in prognosis, and in selection of suitable treatment options, a hallmark of personalized cancer therapy that aims to maximize therapeutic benefits and minimize therapy-associated risks. Thus, screening tumor DNA for mutations of diagnostic, prognostic, and therapeutic significance has now become an integral part of cancer patient management. Clinical molecular oncology laboratories strive to provide accurate and timely identification of these mutations using a variety of mutation detection

R.R. Singh (✉) • R. Luthra (✉)
Department of Hematopathology, University of Texas MD Anderson Cancer Center,
8515 Fannin Street, Houston, TX 77054, USA
e-mail: rsingh@mdanderson.org; rluthra@mdanderson.org

W. Wu, H. Choudhry (eds.), *Next Generation Sequencing in Cancer Research, Volume 2*, DOI 10.1007/978-3-319-15811-2_8

technologies such as Sanger sequencing, pyrosequencing, fragment analysis by capillary electrophoresis, primer extension coupled with mass spectrometry, and allele-specific PCR approaches. Although these technologies have been very efficiently implemented in clinical laboratories, comprehensive mutation analysis using these low-throughput technologies is difficult due to their limited ability to multiplex, especially when tissue is limited.

However, in the last decade, rapid advancements in sequencing technologies and computational methods have resulted in the emergence of the massively parallel next-generation sequencing (NGS) platforms that have drastically decreased the time and the cost associated with comprehensive genome analysis [1–3]. NGS represents a major departure from Sanger sequencing and pyrosequencing, the so-called first-generation sequencing, by permitting whole genome or exome or targeted gene panel sequencing through multiplexing flexibility. This is of high relevance to clinical diagnostic laboratories where mutational screening, until the advent of NGS, has been restricted to few markers through singleplex analysis of mutational hot spot regions or coding region of an individual gene by "first-generation" sequencing assays. The drawbacks of a single gene or exon approach are obvious and include high cost, more labor, and slower turnaround time. These limitations have been circumvented by the high multiplexing capacity of NGS platforms, making comprehensive mutational screening of tumors achievable. Furthermore, NGS technologies facilitate screening of multiple genes with very limited starting material, a significant advantage over conventional sequencing platforms that require relatively larger DNA quantities.

Many clinical laboratories are opting to use massively parallel sequencing capability of NGS which allows simultaneous sequencing of multiple genes and multiple patient samples in a single sequencing reaction as suitable alternative to traditional platforms for comprehensive mutation analysis. However, rigorous validation of the NGS technologies and multi-gene panels is warranted before they are applied for routine clinical screening of tumors. Parallel sequencing of large genomic areas and multiple genes in several multiplexed samples presents challenges not routinely faced by the laboratories using low- and medium-throughput sequencing platforms such as the selection of validation samples and the number and variety of mutations to be validated and reporting of the results [4–7]. Recently, several guidelines have been published addressing this issue [8, 9], which have helped to streamline these issues.

Here, using the example of a 46-gene hot spot NGS cancer panel (AmpliSeq Cancer Panel v1, Life Technologies, CA) that we have implemented in our molecular diagnostic laboratory utilizing Ion Torrent personal genome machine (IT-PGM) [10], we will summarize the different aspects involved in validation of an NGS-based test for clinical laboratory implementation. The IT-PGM platform performs sequencing-by-synthesis by monitoring the release of hydrogen ions or protons during the complementary strand synthesis. The sequencing is interfaced with a semiconductor chip that measures the change in pH resulting from the released hydrogen ions [11]. The recent publications also describe the applications of IT-PGM for mutation screening in hereditary diseases and characterization of microbial genomes [12–14].

2 Validation

2.1 Validation Sample

Samples for validation of a gene panel of interest should include diverse mutations, i.e., single nucleotide variants, insertions, and deletions in as many genes as possible since NGS panels are designed to screen multiple genes and multiple types of mutations simultaneously. Cell lines with known mutations can also be included. For sample diversity, one should consider including different tumor types that will be tested on the NGS platform to have a sufficient admixture of tumor-specific variants.

For the validation of the cancer panel covering 740 mutational hot spots in 46 cancer-related genes, we used 70 formalin-fixed paraffin-embedded solid tumor specimens including 22 archival specimens with known mutations and 48 specimens sequenced in parallel with alternate sequencing platforms. The 70 specimens consisted tumors of various tissue origins: melanoma ($n=36$), colorectal adenocarcinoma ($n=16$), lung ($n=5$), gastrointestinal tract ($n=5$), papillary thyroid ($n=4$), endometrial serous adenocarcinomas ($n=3$), and squamous cell carcinoma ($n=1$). These specimens were tested for mutations by MassARRAY-based multiplex mutation detection assay covering 82 hot spots in 11 genes or by individual Sanger sequencing assays developed in our laboratory.

2.2 Establishing Assay Performance

As majority of the NGS assays are meant for research use only, and the primary intention of validation of NGS assays in a clinical diagnostic laboratory is to test their suitability as routine diagnostic tests for clinical use. Hence, it is imperative to evaluate if the interested regions in the designed panel are being captured and sequenced efficiently and to establish analytical assay parameters such as accuracy, sensitivity, specificity, and reproducibility as discussed below prior to an NGS-based mutation assay deemed ready for implementation in a clinical laboratory environment. Furthermore, it is important to develop alternate assays for confirmation.

2.2.1 Target Capture and Sequencing

Different technologies of target capture are available for isolation and amplification of genomic areas of interest for sequencing on NGS platforms. The two major techniques include high-capacity multiplexed PCR and probe-hybridization capture mechanisms. The 46-gene hot spot panel employs the former multiplexed PCR approach in a single tube where a set of 190 primer pairs amplify genomic areas of interest using low levels of DNA template (10 ng). In a multiplexed PCR, the amplification performance of the primers and the nucleotide sequence in areas of interest define the efficiency of target amplification. This is eventually reflected in the sequencing depth achieved by each amplicon in the panel. For a clinical test, it is

very important to establish a minimum sequencing depth that areas of interest should achieve in order to clearly define those targeted areas that are consistently sequenced adequately and the areas that fail to meet the minimum cutoff. In our study, we defined a cutoff of minimum 300,000 AQ20 sequencing reads (one misaligned base per 100 bases) and 250× sequencing depth to be adequate. Monitoring the overall performance of the amplicons in the sequencing runs for our validation study, we could identify amplicons which consistently reached sequencing depth of ≥ 250×. The mutational status of the genomic regions covered by the failed amplicons (<250×) was recorded as indeterminate during reporting. Using this criterion, the validation studies showed that out of 190 amplicons, 11 amplicons were found to fail routinely due to inadequate sequencing depth. In the validation set of 70 samples, multiplexed sequencing using the 316 chip (4 samples) or 318 chip (8 samples) resulted in each of the samples receiving >300,000 Aq20 reads (average 520,961 reads) with median coverage of >1,000× for each nucleotide. As IT-PGM provides the flexibility of sequencing varying multiplexed samples by using chips of different capacities, we compared the sequencing performance of two chips (316-4 multiplexed samples, 318-8 multiplexed samples) and found the sequencing and variant detecting efficiency to be comparable.

2.2.2 Assay Sensitivity or Limit of Detection

This important parameter helps to establish the sensitivity or lower limit of mutation detection for the NGS platform being implemented. To establish this, we used DNA from two cancer cell lines (H2122 and DLD1) with known mutations in few genes included in the panel. H2122 harbored a homozygous *KRAS* p.G12D and a heterozygous *MET* p.M375S mutation, and DLD1 harbored heterozygous mutations in *PIK3CA* (p.D549N), *KRAS* (p.G13D), *TP53* (p.S241F), and *SMO* (p.P640A). These cell lines were fixed in formalin and embedded in paraffin to mimic the FFPE DNA of tumors. DNA extracted from each cell line was sequentially diluted into FFPE DNA from H460 cell line that did not harbor the mutations expected in either H2122 or DLD1 DNA to provide different dilution levels (100 %, 20 %, 10 %, and 5 %). Sequencing of serial dilutions of H2122 resulted in consistent detection of expected *KRAS* p.G12D and *MET* p.375S mutations at each dilution. Five different sequencing runs of sequentially diluted DNA of DLD1 (100 %, 25 %, 10 %, 5 %, and 2.5 %) showed consistent detection of expected mutations in 25 % dilution. At 10 % dilution, some mutations were not called by the variant caller software even though they were evident in the sequencing reads, which indicated the deficiency of the variant caller software.

2.2.3 Assay Reproducibility

Assessment of NGS assay reproducibility is important as it represents the performance consistency of the assay across different runs and operators. To establish the inter-run reproducibility, i.e., to check whether the same sequence is derived in a

sample when sequenced on different sequencing runs, we sequenced an FFPE patient DNA sample with mutations in *TP53*, *KRAS*, and *MET* in 25 different runs. Our studies showed that each of the three mutations was consistently detected in each of the 25 runs with very minimal variation in the allelic fractions, establishing the inter-run reproducibility.

Establishing intra-run sequencing reproducibility, a measure of consistent sequencing within a sequencing run is also essential and vital to demonstrate that comparable sequencing efficiency is obtained when samples are multiplexed on the same run. To this end, 10 aliquots of FFPE DNA from a single patient with seven mutations in five different genes [*FGFR3* p.P796S (*EGFR* p.K708T, p.R98Q, p.D837N), *NRAS* p.Q61L, *TP53* p.S241T, and *RET* p.P613L] were used as a template for library preparation with each aliquot labeled with a different barcode. The multiplexed libraries were sequenced and the consistent detection of the 7 mutations and the variation of their allelic fractions were assessed. We observed that in every barcoded sample, 5 of the 7 mutations were consistently detected. Inconsistent detection of *FGFR3* p.P796S that was present at low allelic frequency (<10 % and average of 6 %) and *RET* p.P613L that was in a region which that achieved low sequencing depth (<250×) were observed.

3 Concordance of NGS with Orthogonal Sequencing Techniques

Out of the 70 samples used for validation, 22 samples with known mutations were retrospective samples that were previously tested in the laboratory using traditional sequencing platforms. The remaining 48 clinical samples were tested in parallel with 11-gene MassARRAY (Sequenom)-based multiplex assay. Sanger sequencing was used to confirm mutations detected by IT-PGM and not covered by the Sequenom 11-gene panel. In the 22 samples sequenced in retrospective, 29 SNVs, five deletions, and one insertion were expected. IT-PGM detected each of the expected SNVs and 4 deletions except for a 12 bp deletion in *KIT* and a 6 bp insertion in *KIT*. These were evident in the sequencing reads but were not detected and called by the variant caller software version used during our initial validation. In a set of 32 melanoma samples tested in parallel, 47 missense mutations and one insertion were detected by IT-PGM. Each of these mutations was confirmed by either 11-gene panel (Sequenom MassARRAY) or Sanger sequencing except for an SNV in *RB1* that was detected by IT-PGM but was actually found to be a 16 bp deletion by Sanger sequencing. This deletion was evident in the IT-PGM sequencing reads but erroneously called as an SNV by the variant caller. In additional 16 samples of different tumor types, 18 substitution mutations were called by IT-PGM and each of them were confirmed by the 11-gene panel or Sanger sequencing. Overall, we observed a high level of concordance of IT-PGM sequencing with the orthogonal-validated sequencing techniques in our laboratory.

4 False-Positive Calls Induced by Homopolymer Stretches of DNA

As IT-PGM involves sequencing-by-synthesis by sequential flowing of single nucleotides, the presence of homopolymer stretches is a source of false-positive mutations. This occurs primarily due to lack of correlation of the extent of pH change to the nucleotides incorporated. This results mostly in calling of a false single bp deletion or false substitution mutations due to erroneous alignment. Attempt to confirm a subset of these mutations by Sanger sequencing showed that they were of spurious nature induced by homopolymer nucleotide stretches. Reanalysis of these mutations with a later release of the software showed considerable decrease in the rate of false-positive calls at the homopolymer areas.

5 Comparison of Manual and Automated Emulsion PCR Methods

Prior to sequencing, the barcoded-multiplexed DNA libraries are clonally amplified on beads (ionospheres) by emulsion PCR (e-PCR) after which the beads are deposited in wells on the surface of the semiconductor chips for sequencing. We compared the two options available for e-PCR, the first being a manual protocol and an automated option referred to as OneTouch. We compared the efficiency of these methods by the levels of polyclonal ionospheres (which have more than one amplicon amplified on them and hence do not give meaningful sequencing information) and the overall sequencing output. We found that the manual e-PCR methods performed significantly better in comparison to the automated option by providing lower polyclonal ionospheres and consequently better sequencing output. However, since then, better OneTouch instruments (OneTouch 2) have been released whose performance is comparable to the manual e-PCR and we have validated and implemented OneTouch 2 in our laboratory.

6 Interpretation and Clinical Reporting of NGS Results

This represents the most challenging aspect of the implementation of NGS technology for routine use in diagnostic labs. This includes computing power to deal with large amount of data output, establishment of cutoffs, and filtering parameters to identify right mutations in the background of a large number of mutation calls. Once the correct mutations have been identified, it is further challenging to decide which of these calls to be reported clinically and the most appropriate mode to be clear to the physicians and patients. To this end, we developed a filtering and annotating

software (OncoSeek) [10], which facilitates interfacing the patient information (tumor type, tumor percentage, patient identifiers, etc.) along with the NGS results. This software also is used to filter the variant calls, link to the known databases (COSMIC and dbSNP databases) to show the reported frequency and significance of the variants, and annotate them according to the recommended guidelines from Human Genome Variation Society (HGVS). OncoSeek also includes a capacity of visualizing each variant in the sequencing reads using the Integrated Genome Viewer (IGV) from Broad Institute, which is a very crucial step in distinguishing between the real and spurious mutations. OncoSeek also has a self-updating database, which keeps track of the patient samples analyzed and the variants detected. This is very helpful in recognizing and filtering frequent variant calls due to sequencing artifacts (like homopolymer stretches) and filter them out. As a policy, we report all bona fide somatic mutations detected in any gene covered by the panel with the mutations listed in ordered genes listed first followed by mutations in other genes. We do not include silent and intronic mutations. In the report, the mutations in ordered genes are listed first followed by mutations detected in the remaining genes of the panel.

7 Conclusions

On the whole, by following the above described approach for validation, we established that targeted sequencing of genomic regions harboring somatic cancer-related mutations in 46 genes using IT-PGM next-generation sequencer was accurate and on par with the orthogonal sequencing technologies used in our laboratory and could be implemented for routine diagnostic use for solid tumor FFPE samples (summarized in Fig. 1). It is also important to note that we designed Sanger sequencing assays to cover genomic areas interrogated by the 46-gene panel to use as confirmatory assays for NGS. Using similar guidelines, we have also validated a 50-gene hot spot panel for solid tumors (AmpliSeq Hotspot Panel V2, Life Technologies) (unpublished); a 409-gene panel (Comprehensive Cancer Panel, Life Technologies), which screens for mutations in all exons of 409 cancer-related genes using Ion Proton high-capacity sequencer [15]; and also a modified 53-gene panel for acute myeloid leukemia (AML) samples using modified TruSeq Amplicon Panel and MiSeq next-generation sequencer (Illumina Inc.) [16]. Furthermore, recently, several studies have been published from different clinical diagnostic labs regarding validation of several NGS panels, which provide insight in the approaches for validation and clinical reporting of NGS results [17–21]. This in addition to the guidelines published by the regulatory agencies has improved the clarity regarding the strategies for validation and implementation of the progressively evolving NGS technologies, which are very well suited to handle the increasing sequencing demands faced by a molecular diagnostic laboratory to screening large numbers of genes in steadily increasing sample volumes.

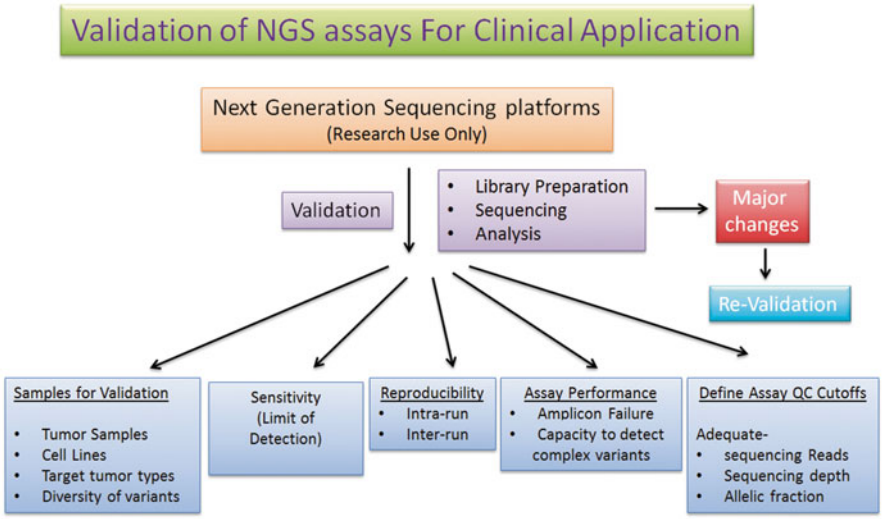


Fig. 1 The overall approach for the validation of next-generation sequencing assays in the clinical environment is summarized. Different assay parameters and quality control metrics which are required to be established for successfully implementation are depicted

References

1. Mardis ER. Next-generation DNA sequencing methods. Annu Rev Genomics Hum Genet. 2008;9:387–402.
2. Ross JS, Cronin M. Whole cancer genome sequencing by next-generation methods. Am J Clin Pathol. 2011;136(4):527–39.
3. Metzker ML. Sequencing technologies—the next generation. Nat Rev Genet. 2010;11(1):31–46.
4. Gullapalli RR, Desai KV, Santana-Santos L, Kant JA, Becich MJ. Next generation sequencing in clinical medicine: challenges and lessons for pathology and biomedical informatics. J Pathol Inform. 2012;3:40.
5. Pant S, Weiner R, Marton MJ. Navigating the rapids: the development of regulated next-generation sequencing-based clinical trial assays and companion diagnostics. Front Oncol. 2014;4:78.
6. Pilgrim SM, Pain SJ, Tischkowitz MD. Opportunities and challenges of next-generation DNA sequencing for breast units. Br J Surg. 2014;101(8):889–98.
7. Ulahannan D, Kovac MB, Mulholland PJ, Cazier JB, Tomlinson I. Technical and implementation issues in using next-generation sequencing of cancers in clinical practice. Br J Cancer. 2013;109(4):827–35.
8. Gargis AS, Kalman L, Berry MW, Bick DP, Dimmock DP, Hambuch T, Lu F, Lyon E, Voelkerding KV, Zehnbauer BA, et al. Assuring the quality of next-generation sequencing in clinical laboratory practice. Nat Biotechnol. 2012;30(11):1033–6.
9. Rehm HL, Bale SJ, Bayrak-Toydemir P, Berg JS, Brown KK, Deignan JL, Friez MJ, Funke BH, Hegde MR, Lyon E. ACMG clinical laboratory standards for next-generation sequencing. Genet Med. 2013;15(9):733–47.
10. Singh RR, Patel KP, Routbort MJ, Reddy NG, Barkoh BA, Handal B, Kanagal-Shamanna R, Greaves WO, Medeiros LJ, Aldape KD, et al. Clinical validation of a next-generation sequencing screen for mutational hotspots in 46 cancer-related genes. J Mol Diagn. 2013;15(5):607–22.

11. Rothberg JM, Hinz W, Rearick TM, Schultz J, Mileski W, Davey M, Leamon JH, Johnson K, Milgrew MJ, Edwards M, et al. An integrated semiconductor device enabling non-optical genome sequencing. Nature. 2011;475(7356):348–52.
12. Elliott AM, Radecki J, Moghis B, Li X, Kammesheidt A. Rapid detection of the ACMG/ACOG-recommended 23 CFTR disease-causing mutations using ion torrent semiconductor sequencing. J Biomol Tech. 2012;23(1):24–30.
13. Mellmann A, Harmsen D, Cummings CA, Zentz EB, Leopold SR, Rico A, Prior K, Szczepanowski R, Ji Y, Zhang W, et al. Prospective genomic characterization of the German enterohemorrhagic Escherichia coli O104:H4 outbreak by rapid next generation sequencing technology. PLoS One. 2011;6(7):e22751.
14. Vogel U, Szczepanowski R, Claus H, Junemann S, Prior K, Harmsen D. Ion torrent personal genome machine sequencing for genomic typing of Neisseria meningitidis for rapid determination of multiple layers of typing information. J Clin Microbiol. 2012;50(6):1889–94.
15. Singh RR, Patel KP, Routbort MJ, Aldape K, Lu X, Manekia J, Abraham R, Reddy NG, Barkoh BA, Veliyathu J, Medeiros LJ, et al. Clinical massively parallel next generation sequencing analysis of 409 cancer-related genes for mutations and copy number variations in solid tumors. Br J Cancer. 2014;111(10):2014–23.
16. Luthra R, Patel KP, Reddy NG, Haghshenas V, Routbort MJ, Harmon MA, Barkoh BA, Kanagal-Shamanna R, Ravandi F, Cortes JE, et al. Next-generation sequencing-based multigene mutational screening for acute myeloid leukemia using MiSeq: applicability for diagnostics and disease monitoring. Haematologica. 2014;99(3):465–73.
17. Beck J, Pittman A, Adamson G, Campbell T, Kenny J, Houlden H, Rohrer JD, de Silva R, Shoai M, Uphill J, et al. Validation of next-generation sequencing technologies in genetic diagnosis of dementia. Neurobiol Aging. 2014;35(1):261–5.
18. Zhang L, Chen L, Sah S, Latham GJ, Patel R, Song Q, Koeppen H, Tam R, Schleifman E, Mashhedi H, et al. Profiling cancer gene mutations in clinical formalin-fixed, paraffin-embedded colorectal tumor specimens using targeted next-generation sequencing. Oncologist. 2014;19(4):336–43.
19. Lin MT, Mosier SL, Thiess M, Beierl KF, Debeljak M, Tseng LH, Chen G, Yegnasubramanian S, Ho H, Cope L, et al. Clinical validation of KRAS, BRAF, and EGFR mutation detection using next-generation sequencing. Am J Clin Pathol. 2014;141(6):856–66.
20. Ong M, Carreira S, Goodall J, Mateo J, Figueiredo I, Rodrigues DN, Perkins G, Seed G, Yap TA, Attard G, et al. Validation and utilisation of high-coverage next-generation sequencing to deliver the pharmacological audit trail. Br J Cancer. 2014;111(5):828–36.
21. Dames S, Chou LS, Xiao Y, Wayman T, Stocks J, Singleton M, Eilbeck K, Mao R. The development of next-generation sequencing assays for the mitochondrial genome and 108 nuclear genes associated with mitochondrial disorders. J Mol Diagn. 2013;15(4):526–34.

Next-Generation Sequencing Technologies and Formalin-Fixed Paraffin-Embedded Tissue: Application to Clinical Cancer Research

Nadine Norton

Abstract Billions of tissue samples are now archived by formalin fixation paraffin embedding (FFPE) in tissue banks and hospitals around the world. For those biomarkers measured by immunohistochemistry and used today as a standard of care in cancer treatment, this method of preservation works well. However, the heterogeneous nature of the disease means that many patients do not respond or relapse under standard treatment.

Next-generation sequencing (NGS) technologies now provide extensive genome analyses at the level of gene expression, identification of somatic copy number aberration, somatic single nucleotide variants, fusion transcripts, and epigenetic modification. Successful application of this technology to the large volumes of archival FFPE material with long-term follow-up data will be a hugely powerful tool in identifying new biomarkers of disease outcome, disease recurrence, and treatment response.

The major hurdle for NGS application to archival material is the effect of formalin fixation on nucleic acids. The process results in chemical modification, cross-linking, and fragmentation. Chemical modification can result in false-positive mutation calls, and fragmentation can result in overrepresentation of the 3′ end of genes creating bias in gene expression. There are now a number of NGS kits and protocols which are marketed specifically for use with FFPE material. Laboratories are beginning to validate and apply these methods. In this chapter, we review the progress in the adaption of NGS technologies to FFPE tissue for clinical cancer research.

1 Introduction

The wealth of clinical data such as patient outcome and survival from large clinical trials is an invaluable tool in the successful application of genomic technologies. More than one billion samples are preserved with formalin-fixed paraffin

N. Norton, Ph.D. (✉)
Department of Cancer Biology, Mayo Clinic, 4500 San Pablo Road,
Jacksonville, FL 32224, USA
e-mail: Norton.nadine@mayo.edu

W. Wu, H. Choudhry (eds.), *Next Generation Sequencing in Cancer Research, Volume 2*, DOI 10.1007/978-3-319-15811-2_9

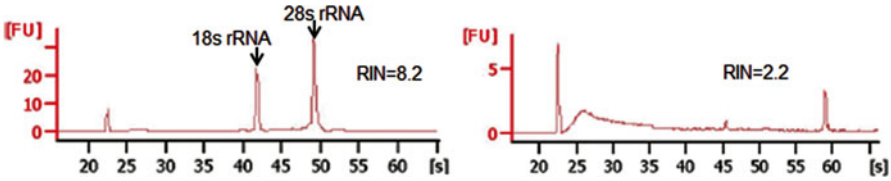


Fig. 1 RNA profiles of RNA extracted from fresh-frozen breast tumor and formalin-fixed breast tumor. Agilent profiles of RNA extracted from (*left*) fresh-frozen breast tumor and (*right*) formalin-fixed breast tumor. *rRNA* ribosomal RNA, *RIN* RNA integrity score. RIN scores are calculated from the ratio of 18s and 28s rRNA present within the sample. A RIN score >7 is determined as suitable for RNA-Seq library preparation using mRNA "pull-down" methods, while RIN scores <7 are considered too degraded

embedding (FFPE) in hospitals and tissue banks across the world [1]. Unfortunately, this method of preservation causes chemical modification and degradation to RNA (Fig. 1) and DNA [2–8], compromising the use of next-generation sequencing technologies that measure and identify differences in gene expression, single nucleotide variants (SNVs), and copy number variants/aberrations (CNVs/CNAs) and detection of gene fusion transcripts. In this chapter, we review the technical challenges and progress in the adaption of NGS technologies to FFPE tissue for clinical cancer research.

2 Technical Challenges

2.1 RNA

Our understanding of RNA complexity in the cell has changed dramatically in recent years. The composition of total human RNA has direct impact on clinical cancer research, the preferred techniques used to isolate and sequence the transcriptome, and subsequently the preferred techniques for sequencing RNA from formalin-fixed material. Early transcriptome sequencing techniques focused on mRNA, the fraction of RNA that is translated into amino acids. mRNA accounts for ~1–3 % of total RNA, while ribosomal RNA (rRNA) accounts for ~80 % and transfer RNA (tRNA) accounts for ~15 %. Early RNA sequencing techniques relied on the presence of a polyA tail on mRNA molecules, using oligo dT beads to "pull down" mRNA transcripts from total RNA. These methods are suited only to high-quality, undegraded RNA because many mRNA transcripts from formalin-fixed tissue will be missing their polyA tail due to fragmentation (Fig. 2). Further challenges in preparation of RNA from formalin-fixed tissue arise from cross-linking between nucleic acids and proteins, which limit the reverse transcription of mRNA into cDNA, a key step in the process of library preparation for transcriptome sequencing (Fig. 2). Several commercially available kits and protocol adaptations now claim

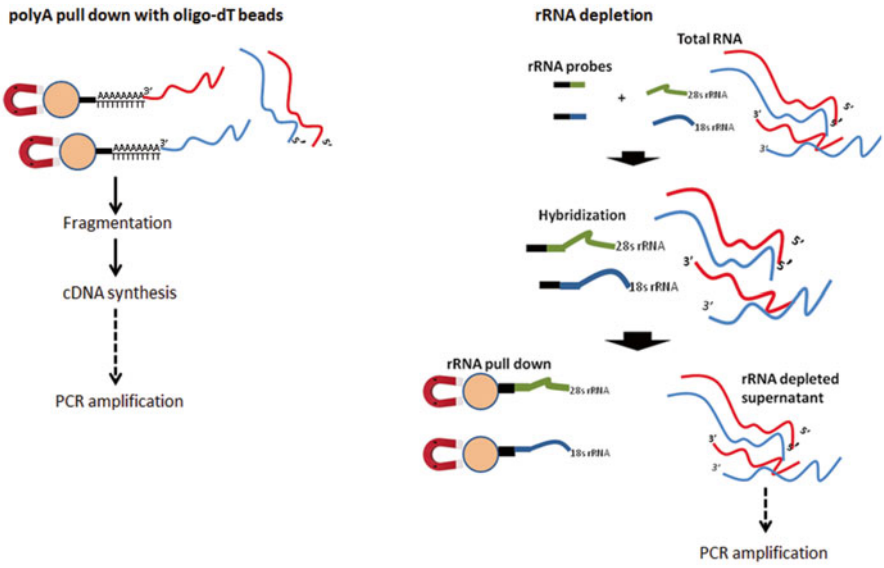


Fig. 2 RNA-Seq library preparation methods. Comparison of RNA-Seq library preparation methods. rRNA depletion methods are more appropriate for enrichment of fragmented transcripts which lack the polyA tail. *rRNA* ribosomal RNA

to remedy the technical challenges of whole-transcriptome sequencing from formalin-fixed tissue using an rRNA depletion technique rather than mRNA "pull down" [9–13] (Fig. 2).

Multiple studies employing these new methods have demonstrated that RNA expression from formalin-fixed tumor samples can reliably quantify gene expression. In addition, techniques that are not dependent on polyA "pull down" have provided relevant findings to clinical cancer research in relation to other RNA types such as microRNA (miRNA), long interspersed noncoding RNA (lincRNA), small nucleolar RNA (snoRNA), small Cajal body-specific RNA, and histone H1 cluster transcripts (which lack polyA tails), all of which can be detected and reliably quantified in formalin-fixed material.

2.2 DNA

Tumor purity, heterogeneity, and clonality make accurate variant calling challenging even with high-quality starting material. Formalin fixation confounds single nucleotide variant (SNV) calling due to chemical modification, predominantly through cytosine deamination to uracil which pairs with adenine, mimicking a C>T variant. This is clearly demonstrated by multiple studies reporting significantly more SNVs in FFPE tumor samples compared to high-quality DNA from matched

fresh-frozen tumor tissue. In addition, formalin cross-linking between DNA and protein results in fragmentation and enzymatic inhibition, which reduces library fragment size and uniformity and compromises overall success rate of NGS library preparation. Many published studies comparing FFPE and fresh-frozen DNA sequencing focus on SNV concordance of those samples that produced libraries for sequencing, neglecting to mention the proportion of FFPE samples that did not produce libraries to begin with. These issues are being addressed with new protocol adaptations: flow cytometry-based methods to isolate pure populations of tumor cell nuclei from FFPE tissue prior to sequencing, deamination removal by uracil-DNA glycosylase, increased depth of sequencing, and improved variant-calling algorithms.

3 Next-Generation RNA Sequencing from FFPE Material: Analytical Validation

The major challenge of any gene expression study is to differentiate between true biological differences and artifactual differences, but even with improved methods of whole-transcriptome sequencing, determination of what is artifact due to formalin fixation and what is artifact due to different protocols and platforms is equivocal. Norton et al. [14] systematically tested whole-transcriptome sequencing of archival material from breast tumor specifically for technical reproducibility and gene expression artifacts due to sample degradation, formalin fixation, library preparation method (polyA pull down versus rRNA depletion), and orthogonal platform difference. This set of experiments is summarized below.

3.1 RNA-Seq Protocol Comparison

RNA-Seq protocols employing ribosomal RNA depletion to enrich for mRNA are more appropriate for FFPE material than polyA pull-down methods. The goal of this experiment was to determine differences in gene expression due to protocol by comparing the same set of high-quality RNA samples with rRNA depletion (Ribo-Zero Gold/ScriptSeq V2, Epicentre) and polyA pull-down (TruSeq, Illumina) protocols. The number of reads mapping to the genome was almost identical regardless of protocol. However, the percentage of reads mapped within genes and to exon junctions was higher for the polyA pull-down protocol where the mean percentage of reads mapped to genes and exon junctions was 77.4 %, SD ± 0.74, and 15.8 %, SD ± 0.20, respectively. In the rRNA depletion protocol, the percentage of reads mapped within genes and to exon junctions was 50.6 %, SD ± 1.32, and 10.29 %, SD 0.36, suggesting the polyA pull-down protocol yielded a higher percentage of reads mapping to coding genes and the rRNA depletion protocol yielded a greater proportion of reads mapping to intronic and intergenic regions, regardless of sample quality.

3.2 Fragment Degradation

The goal of this experiment was to assess the effect of fragment degradation on an rRNA depletion protocol for whole-transcriptome sequencing without the complication of formalin fixation. The experiment used high-quality RNA from two commercially available cell lines (MDA-MB-436 and UHRR). Initial RNA quality was determined by the RNA Integrity Score (RIN) which ranges from 1 to 10 and a score of 10 means that the RNA is completely intact. Whole-transcriptome protocols based on polyA pull down require samples to have RIN scores >7. In this experiment, the undegraded cell line RNAs had RIN scores of 10 and 8.1, respectively. RNA from each cell line was manually degraded in the laboratory with heat or physical shearing to produce medium degradation (RIN scores ranging 4.7–6.8) and high degradation (RIN scores ranging 1.2–2.2). Whole-transcriptome sequencing was performed on both the high-quality (fully intact) sample and its manually degraded match. Correlation analyses were performed for each gene (23,498 RefSeq gene annotations) in each degraded sample and its undegraded match. Average Pearson correlation for MDA-MB-436 for pairs of undegraded versus degraded RNA was 0.945 under medium degradation and 0.922 when highly degraded (RIN 1.2–2.2). For UHRR, average Pearson correlation for medium and high degradation were 0.809 and 0.805, respectively, demonstrating that under this rRNA depletion protocol, gene expression can be reliably quantified in degraded RNA.

3.3 Formalin Fixation

The goal of this experiment was to assess the effects of both degradation and formalin fixation on gene expression when using the same rRNA depletion method as above. Reproducibility was assessed by correlation of gene expression in technical replicates of the same FFPE sample. Pearson correlation for technical replicates was 0.998 and in agreement with other studies summarized in Table 1. Accuracy was assessed by correlation of gene expression across nine matched pairs of RNA from fresh-frozen (RIN scores >7) and FFPE tissue (RIN scores ranging 1.2–2.3). Pearson correlation for gene expression between FFPE and matched fresh-frozen RNA ranged from 0.598 to 0.830 (mean 0.783, $p < 2 \times 10^{-16}$). These figures are in agreement with multiple studies and protocols (summarized in Table 2) and

Table 1 Reproducibility of RNA-Seq FFPE technical replicates

Pearson correlation	RNA-Seq library preparation method	References
0.947–0.985	rRNA depletion as per Morlan et al. [10]/ScriptSeq V1	Sinicropi et al. [11]
0.998	Ribo-Zero Gold/ScriptSeq V2	Norton et al. [14]
0.991	Ribo-Zero Gold/TruSeq and DSN/TruSeq	Zhao et al. [28]

Table 2 Correlation of RNA-Seq matched fresh-frozen and FFPE tissue pairs

Pearson correlation	RNA-Seq library preparation method	Type of RNA-Seq	Tissue type	Number matched pairs	Number differentially expressed genes	References
0.79	TruSeq small RNA protocol V1.5	microRNA deep sequencing	Clear cell renal cell carcinoma and benign kidney	6	Data not shown	Weng et al. [29]
0.78	Ribo-Zero Gold/ ScriptSeq V2	Whole transcriptome	Breast tumor	9	3,540 genes > log2 fold change in 1 or more pairs 76 genes > log2 fold change in 5 or more pairs	Norton et al. [14]
0.90	Ribo-Zero Gold/ ScriptSeq V2	Whole transcriptome	Bladder, prostate, and colon carcinoma; liver and colon normal tissue; reactive tonsil	38	1,494 genes significantly different expression across pairs	Hedegaard et al. [30]
0.896 0.924	Ribo-Zero Gold/ TruSeq and DSN/ TruSeq	Whole transcriptome	Breast tumor	1	Data not shown	Zhao et al. [28]

demonstrate that reliable gene expression data can be generated from whole-transcriptome studies using FFPE material. However, the study also highlighted differences in NGS mapping statistics between fresh-frozen and FFPE material. The total number of reads mapped to the genome for nine FFPE samples was 79.2 %, similar to that of fresh-frozen material, but when comparing the same protocol across matched fresh-frozen and FFPE samples, Norton et al. observed a significantly lower percentage of reads mapping to genes (fresh-frozen = 50.6 %, SD ± 3.95, and FFPE = 24.8 %, SD ± 2.8) and to exon junctions (fresh-frozen = 10.29 %, SD ± 1.09, and FFPE = 3.98 %, SD ± 0.58) in FFPE samples, which influenced further downstream NGS applications such as single nucleotide variant and fusion transcript detection.

3.4 Long Interspersed Noncoding RNA (lincRNA)

Norton et al. also aligned RNA-Seq data from to 5,749 lincRNAs from the Havana group (http://www.sanger.ac.uk/research/projects/vertebrategenome/havana/) and performed correlation analyses of the nine fresh-frozen and FFPE pairs. The percentage of reads mapping to lincRNA from FFPE libraries ranged from 4.44 to 6.32 % (mean 5.47 %), surprisingly better than that achieved in the matched fresh-frozen libraries using the same protocol (range 2.63–4.89 %, mean 3.58 %). A high degree of correlation of lincRNA expression between matched FFPE and fresh-frozen pairs was observed (Pearson correlation ranging 0.969–0.997, mean correlation $r=0.989$), suggesting studies of lincRNA expression are highly suitable for samples isolated from FFPE material.

3.5 Expressed Single Nucleotide Variants (eSNVs)

The goal of this experiment was to assess sensitivity of eSNV detection in FFPE material. On average 52.2 % of SNVs that were detected with high confidence in fresh-frozen samples were also detected in matched FFPE samples. Lower sensitivity correlated with smaller library insert size, $r^2=0.73$, a parameter that could be used to predict performance of SNV detection specifically for each FFPE sample. Performance of gene expression from FFPE material has previously been correlated with time of fixation rather than age of sample [8], although not with RNA-Seq or SNV detection, and in reality, for many available FFPE samples, fixation time will be unknown. If the sensitivity can be predicted from parameters such as library insert size, this will enable power calculations on required sample size to identify genes with cancer-associated SNVs.

3.6 Fusion Transcripts

The goal of this experiment was to assess sensitivity of fusion transcript detection in FFPE material. As demonstrated in (3.1), the rRNA depletion method used for FFPE material yielded a lower percentage of reads mapping to coding genes (even in high-quality RNA), and as demonstrated in (3.3), RNA from FFPE material showed a lower percentage of reads mapping to coding genes and exon junctions, both important parameters in detection of fusion genes. Only 24 % of high-confidence fusion transcripts detected in fresh-frozen RNA were also detected in matched FFPE RNA samples, an issue that was not overcome by an increase in depth of sequencing up to threefold (increase from ~56 to ~159 million reads).

Table 3 RNA-Seq orthogonal platform validation

Pearson correlation	RNA-Seq library preparation method	Validation platform	Note	References
0.813	rRNA depletion as per Morlan et al. 2012/ ScriptSeq V1	RT-PCR 14 genes in 136 patients	Correlation of hazard ratios across platforms	Sinicropi et al. [11]
0.838	Ribo-Zero Gold/ ScriptSeq V2	NanoString Cancer reference panel, 226 genes	Correlation of FFPE gene expression across platforms	Norton et al. [14]

3.7 Orthogonal Platform Validation (Table 3)

Transcriptome studies in discovery samples will generate a high number of disease or disease-associated genes, requiring validation and further replication. Following replication, prognostic and/predictive molecular profiles will form the basis of widely used clinical diagnostic tests. For this reason, it is important to identify discovery and replication platforms which achieve high correlation for gene expression in FFPE material. Norton et al. observed good correlation of gene expression of 236 cancer-related genes in nine FFPE RNA samples between whole-transcriptome sequencing and the NanoString nCounter platform, Pearson correlation ranging 0.468–0.923 (mean 0.839).

4 Next-Generation RNA Sequencing from FFPE Material: Discovery Studies

With the advent of RNA sequencing protocols better suited to RNA extracted from FFPE material, the number of FFPE discovery-based studies will increase and identify new biomarkers in the field of clinical cancer research. Table 4 highlights three FFPE RNA discovery studies: the first was designed to validate an rRNA depletion, the second is the RNA-Seq protocol to identify known RNA biomarkers used to determine risk of recurrence of breast cancer [11], and the third is the discovery of new RNA biomarkers in the same FFPE-derived breast cancer cohort. This study used FFPE tumor blocks from 136 patients with breast cancer. The age of the tumor blocks ranged from 5 to 12.4 years (median 8.5 years). The cohort had previously been utilized by standard microarray and reverse transcriptase (RT)-PCR-based technologies resulting in a 21 gene-based RT-PCR-based test (Oncotype DX) now used in the clinical setting to guide treatment decisions for ER+ breast cancer patients [15]. Hazard ratios obtained with the RNA-Seq protocol for the known biomarkers in this cohort were highly correlated with the hazard ratios obtained with standard RT-PCR, Pearson correlation 0.813. The advantage of the RNA-Seq

Table 4 FFPE NGS discovery studies in clinical cancer research

Method	Discovery sample description	Result	References
rRNA depletion as per Morlan et al. 2012/ ScriptSeq V1 (Epicentre), Illumina platform	RNA from Providence cohort of 136 breast cancer patients	>2,000 RNAs associated with breast cancer recurrence, many of which were intronic RNAs	Sincropi et al. [11]
rRNA depletion as per Morlan et al. 2012/ ScriptSeq V1 (Epicentre), Illumina platform Developed algorithm for detection of fusion transcripts in FFPE material	RNA from Providence cohort of 136 breast cancer patients and Rush cohort of 76 breast cancer patients	High frequency of fusion transcripts correlated with poor outcome ($P < 0.0005$)	Ma et al. [16]
Laser capture microdissection, RNA-Seq/Ovation RNA-Seq FFPE, and Encore NGS Library system 1 kits (NuGEN), Illumina platform	RNA from normal, adenocarcinoma in situ, and invasive adenocarcinoma tissue from six patients with lung cancer	5 lincRNAs and 31 mRNAs consistently associated with lung cancer progression	Morton et al. [18]
Whole-exome sequencing, SureSelect Human Whole exome kit (Agilent), SOLiD4 platform (Applied Biosystems), somatic copy number detection	Paired/normal DNA samples from 5 castration-resistant metastatic prostate cancer patients	Amplification of *YWHAZ* and *PTK2* genes in discovery cohort. Frequency of amplifications were associated with progression in a prostate cancer validation cohort	Menon et al. [24]
Targeted sequencing of 182 known cancer-related genes	74 tumors (primary and recurrent) from 43 patients with breast cancer (33 matched)	Increased frequency of CDK4/MDM2 amplifications in recurrences	Meric-Bernstam et al. [25]
Targeted sequencing of 236 cancer-related genes and 47 introns of 19 genes commonly rearranged in cancer	29 adrenocortical carcinoma patients	At least one genomic alteration found in 22/29 patients	Ross et al. [27]

protocol is the ability to assess risk of recurrence of thousands of transcripts simultaneously, including non-protein-coding intronic and intergenic sequences such as microRNAs and lincRNAs. In total, the study identified 1,307 protein-coding RNAs and 1,698 intronic RNAs that were significantly associated with risk of recurrence, at a false discovery rate of <10 %. Interestingly, for most of the intronic RNAs identified as prognostic, their related exons were not also identified as prognostic, suggesting that these intronic sequences carry biomarker information not captured in gene coding sequence.

The previous section on analytical validation describes some of the difficulties in identification of gene fusion transcripts in FFPE material. Despite this, bioinformatics methods to identify gene fusions in FFPE material have improved [16]. The 136 and 76 FFPE breast cancer patient cohorts described in [11, 17] were recently used as discovery samples for gene fusion transcripts associated with poor outcome. One hundred eighteen candidate fusion events (100 unique) were detected. Using TaqMan assays, 47/77 (61 %) fusion events were validated. Of the 100 unique fusion junctions, only one had been previously described and a high frequency of fusion transcripts correlated with poor outcome ($p<0.0005$) highlighting the potential of FFPE cohorts with mature clinical records in the discovery of novel biomarkers.

Finally, the isolation techniques such as laser capture microdissection are being used in conjunction with RNA-Seq of FFPE material allowing researchers to identify differentially expressed genes between tightly juxtaposed cell and tissue types. In a discovery study of lung cancer progression, this combination of techniques allowed differential expression analyses of mRNAs and lincRNAs from normal, adenocarcinoma in situ (AIS) (an intermediate step in progression of normal lung tissue to invasive adenocarcinoma), and invasive adenoma carcinoma FFPE samples from the same six patients [18]. The study identified five lincRNAs and 31 mRNAs that were consistently up- or downregulated from normal to AIS and more strongly to invasive carcinoma.

5 Next-Generation DNA Sequencing from FFPE Material: Analytical Validation

The major goal of next-generation DNA sequencing in clinical cancer research is the identification of somatic single nucleotide variants (SNVs), insertions/deletions, translocations, and copy number aberrations that are either prognostic of disease outcome or predictive of treatment response. To make any meaningful predictions from clinical trial datasets, variant calling must be both sensitive (low false-negative rate) and specific (low false-positive rate). Matched pairs of DNA extracted from FFPE tumor and fresh-frozen tumor tissue from the same patient are an extremely useful resource as a validation of genetic variant calling. Observations from these studies will determine subsequent experimental design and methodology in discovery studies.

5.1 Sensitivity

Concordance of SNV calls between fresh-frozen and FFPE material is not reported uniformly across studies. If we assume that all calls made in high-quality DNA extracted from fresh-frozen material are true and define sensitivity within FFPE material as the percentage of SNV calls made from fresh-frozen material that were

also detected in matched FFPE material, the sensitivity of SNV calls from FFPE tissue (within current literature) ranges from ~70 to 98 % (Table 5). Other than comparison of variant calls in DNA extracted from matched fresh-frozen and FFPE tissue, no study listed in Table 5 uses the same protocol in either library preparation or variant detection. For the purpose of SNV detection, the factors that appear to give the highest sensitivity (>95 %) are overall increased depth of coverage and calls made with at least ≥50× coverage at the variant position in both fresh-frozen and FFPE samples. These levels of coverage and sensitivity are routinely achieved with targeted sequencing and much less so with whole-exome and whole-genome sequencing.

We also note from these studies that the number of matched pairs is small (range: 1–17 pairs), making separation of other potential confounding factors, such as FFPE sample age, difficult to determine systematically. The largest study in Table 5 [19] uses a targeted NGS approach covering 88 genes. The overall sensitivity of SNV calls across all 17 matched pairs is not reported in this study (for good reason), but based on our earlier observation of depth of coverage, sensitivity should be in the highest range. What this study demonstrates is that all FFPE samples are not equal: for 3/17 FFPE samples, sensitivity was less than 60 %, despite a targeted approach. In the same study, Bourgon et al. [19] developed a sample quality control (QC) measure of functional DNA copies using a qPCR-based amplification assay at the *TRAK2* locus. Those three samples with lowest sensitivity also had the lowest number of functional DNA copies. However, even following this additional QC measure, correlation between functional copy number and sensitivity was poor, with some low functional copy number samples performing well within the range of those samples with high sensitivity.

Further observation in Table 5 is the lack of data of DNA copy number variation. Only two of eight studies report concordance of CNVs between matched fresh-frozen and FFPE pairs, likely reflecting the difficulty of these calls in FFPE material. The whole-genome study of Schweiger et al. reported perfect concordance of somatic copy number aberrations in known breast cancer copy number loci, on chromosomes 8 and 20 [20]. The second report of CNV in Table 5 used an exome sequencing approach and only matched tissue pair, reporting a high degree of noise and poor concordance [21]. A third study [22] performed exome sequencing and CNV analysis on one matched fresh-frozen-FFPE pair. This study used flow cytometry to enrich tumor nuclei specifically from FFPE tissue (prior to sequencing), thus reducing contamination of DNA from non-tumor cells. Admixture of tumor and non-tumor DNA reduces sensitivity, for both SNV and CNV. However, CNV analysis from FFPE material in this study was estimated from comparative genomic hybridization (CGH) arrays of DNA from flow-sorted nuclei and was either not attempted or not reported based on analysis of exome sequence data.

Table 5 Correlation of DNA-Seq matched fresh-frozen and FFPE tissue pairs

Scale	Library preparation method	Tissue type	Number matched pairs	SNV detection	CNV detection	References
Whole genome	Illumina 1.5 µg genomic DNA input	Breast tumor and normal breast tissue	3	81–90 % SNVs identified in FF were also identified in FFPE at ≥8× coverage 1.3–2 fold more SNVs were identified in FFPE compared to FF	All somatic copy number aberrations were concordant	Schweiger et al. [20]
Exome	FFPE libraries: flow cytometry, NuGEN Ovation WGS FFPE system, NuGEN Encore ds-DNA module, Agilent SureSelect 50 Mb Fresh-frozen libraries: Agilent SureSelect 50 Mb	Pancreatic ductal adenocarcinoma	1	22 known mutations in tumor-derived cell line from same patient used as reference. 18/22 mutations were called in both FFPE and FF samples at positions with ≥10× coverage	CNV detection not attempted with exome sequencing	Holley et al. [22]
Exome	SOLiD pre-capture 3 µg DNA input and SOLiD optimized Agilent SureSelect exome 37 Mb	Prostate cancer: tumor and normal tissue	1	85 % of SNVs identified in FFPE were present in matched FF tumor	High degree of noise and generally poor concordance	Menon et al. [23]
Exome	Illumina TruSeq 1.2 µg genomic DNA input	Colorectal carcinoma	10	FFPE samples stored <3 years showed 70–80 % SNV overlap with FF pairs at positions with ≥10× coverage. Older FFPE samples showed much greater proportion of SNVs not identified in FF. All previously known KRAS and BRAF mutations were identified in FFPE samples	CNV detection not attempted	Hedegaard et al. [30]
Exome	Agilent SureSelect 52 Mb	Radical prostatectomy	5	179 (1.2 %) discordant loci between FFPE and FF at positions with ≥20× coverage. No discordant loci at positions with ≥80× coverage	CNV detection not attempted	Kerick et al. [21]

Targeted sequence	Agilent SureSelect 3.9 Mb DNA input 0.5, 1.5, and 3 μg	Radical prostatectomy	5	SNV concordance rates between FFPE and FF were ~98 % at positions with ≥55× coverage for both known and novel SNVs	CNV detection was not performed in FF/FFPE pairs	Kerick et al. [21]
Targeted sequence	75 ng input DNA pre-amplified prior to targeted enrichment of 963 amplicons (88 genes) targeted by Fluidigm access array	4 breast, 4 lung, 4 colon, and 5 ovarian cancer	17	High concordance in samples with high number of functional copies determined by qPCR	CNV detection not attempted	Bourgon et al. [19]
Targeted sequence	Agilent SureSelect 27 genes, 1 μg input DNA	Lung adenoma carcinoma	16	99 % of SNV identified in FF were also identified in FFPE at positions with ≥50×	CNV detection not attempted	Spencer et al. [31]

5.2 Specificity

The most consistent finding reported across NGS comparisons is the significantly high number of SNVs called in DNA extracted from FFPE tissue that are not detected in matched fresh-frozen tissue, a problem that will be magnified at the level of the exome and whole genome. In the whole-genome sequencing study of Schweiger et al., ~3,000 variants were unique to each FFPE sample when compared with DNA extracted from its fresh-frozen counterpart [20]. Kerick et al. [21] reported 100 % concordance in five matched FF/FFPE pairs of tissue from radical prostatectomy, when comparing variant calls with ≥80× coverage at the level of the exome. Unfortunately, the typical average depth of coverage at the exome and whole genome (at this time) is often less than 80×, and the use of such high stringency will likely result in exclusion of true variants. Specificity was greatly improved by implementation of two strategies in the targeted DNA sequencing study of Bourgon et al. [19]. Firstly, a QC measure of each FFPE sample (prior to target enrichment) showed strong association of the number of functional copies of DNA with sensitivity. Those FFPE samples with lower numbers of functional DNA copies generated the largest number of false-positives (determined as variants not called in the matched fresh-frozen sample), and samples with higher functional copies demonstrated almost 100 % specificity when using more stringent variant read frequency cutoffs. Secondly, following sample QC, Bourgon et al. selected four FFPE samples with different numbers of functional copy number, ranging from low to high, and treated these samples with uracil-DNA glycosylase (UDG) to remove uracil-containing deaminated DNA molecules prior to target enrichment. Variant calls were compared for the same FFPE samples both treated and untreated with UDG. UDG treatment leads to 77 % and 94 % reduction in C>T and A>G variant calls, respectively, without impacting sample sensitivity.

6 Next-Generation DNA Sequencing from FFPE Material: Discovery Studies

Observations and protocol adaptations from the studies above are paving the way for the discovery of mutations and their correlation with outcome and treatment, although as we see from the example below, success will depend on multiple converging strategies including variant prioritization, replication in independent samples, and functional validation.

In the previous section we noted the difficulty in CNV detection in FFPE material, particularly from exome sequence data. Menon et al. reported a high degree of noise and poor concordance between CNV calls in a single fresh-frozen and FFPE DNA pair from prostate tissue [23]. Castration-resistant prostate cancer (CRPC) is a lethal form of prostate cancer, and obtaining this type of tissue from fresh-frozen material is a major hurdle in identifying the underlying molecular alterations in an

aggressive disease. Exome sequencing of DNA extracted from FFPE tissue of five CRPC patients and their corresponding non-tumor samples, by the same group, identified somatic copy number alterations in *YWHAZ* and *PTK2* [24]. As with the initial FFPE study, noise levels were high. The number of putative copy number aberrations was reduced by prioritizing those present in three or more samples, identifying 928 amplified genes and 647 deleted genes. This gene list was further reduced by focusing on a region already previously described to harbor the *cMYC* amplification in prostate cancer within this region, focusing on genes already known to play a role in cancer, genes with commercially available inhibitors, and genes already studied in prostate cancer. Within the filtered gene list of putative copy number aberration, *YWHAZ* and *PTK2* were amplified at higher levels than *cMYC* in the discovery sample and follow-up in a prostate cancer progression cohort consisting of cases with clinically localized prostate cancer, patients with primary and corresponding lymph node metastasis, and samples with CRPC. Both *YWHAZ* and *PTK2* showed significantly higher levels of amplification in lymph node metastases and CRPC samples compared to localized prostate cancer, providing preliminary evidence for potential therapeutic targets in CRPC.

A second example of NGS discovery in FFPE material in clinical cancer research is illustrated from comparative studies of somatic variants in primary and metastatic tissue from the same breast cancer patients. Genomic characterization of paired primary and recurrent or metastatic lesions could identify novel biomarkers or potential therapeutic targets specifically relevant to patients with recurrent or metastatic breast cancer, but metastatic tissue is often only available as archival material. Using a deep sequencing targeted NGS approach of 182 cancer-related genes (average depth of coverage of 380×), Meric-Bernstam et al. [25] showed overall high concordance between 33 matched primary and recurrent breast tumors preserved by FFPE but identified 23/159 (14.4 %) of somatic CNAs that were discordant between matched primary and recurrent tissues, with an increased frequency of *CDK4*/*MDM2* amplifications in recurrences.

A third example of NGS discovery in FFPE material is taken from the adrenocortical carcinoma (ACC) literature. AAC is a rare tumor (annual incidence 0.7–2.0 cases per million people) with poor prognosis. Disease is treated with surgical resection and systemic cytotoxic therapies, with no targeted therapies used at present. Comprehensive genomic analyses (exome sequencing and SNP arrays) of high-quality DNA from fresh-frozen tissue of 45 ACC patients previously identified mutations in driver genes both known and novel to ACC [26]. Despite this progress, the heterogeneous nature of disease will require sequencing of many more samples to determine the ACC genomic landscape. Accessing additional patient samples from archival material will aid in this task. Already this year, a further 29 patients, specifically with locally advanced or metastatic ACC refractory to their last time of cytotoxic chemotherapy, were characterized from FFPE material using a targeted deep sequencing approach (average coverage 734×) across 236 cancer-related genes and 47 introns of 19 genes commonly rearranged in cancer, identifying many additional mutations in genes [27]. 17/29 patients carried genomic alterations (SNVs, gene amplifications, or deletions) in genes with an available therapeutic or

mechanism-based clinical trial. What the study did not do (likely due to small sample size) is identify genomic alterations which are prognostic of outcome or predictive of outcome with treatment. These are analyses that may come with the accumulation of genomic data, leading to prospective clinical/translational trials.

7 Summary

7.1 RNA

Despite technical challenges, multiple studies have demonstrated that RNA expression from formalin-fixed tumor samples can (1) reliably quantify gene expression and (2) provide relevant findings to clinical cancer research in relation to both mRNA and other RNA types (microRNA (miRNA), long interspersed noncoding RNA (lincRNA), small nucleolar RNA (snoRNA), small Cajal body-specific RNA, and histone H1 cluster transcripts), but SNV calling and fusion transcript detection remains challenging.

7.2 DNA

Sensitivity of SNV calls from FFPE tissue (within current literature) ranges from ~70 to 98 % depending on depth of coverage and call criteria, but the highest estimates of sensitivity use criteria that are currently unrealistic for whole-genome and whole-exome sequencing. All FFPE samples are not equal, even when accounting for age and sample fixation time. However, SNV calls from FFPE material can be improved by (1) implementation of QC measures of the number of functional copies of DNA and (2) enzymatic removal of uracil-containing deaminated DNA molecules prior to target enrichment. Observations and protocol adaptations from the studies described in this chapter are paving the way for discovery of mutations and their correlation with outcome and treatment.

References

1. Blow N. Tissue preparation: tissue issues. Nature. 2007;448:959–63.
2. Auerbach C, Moutschen-Dahmen M, Moutschen J. Genetic and cytogenetical effects of formaldehyde and related compounds. Mutat Res. 1977;39:317–61.
3. Bresters D, Schipper ME, Reesink HW, Boeser-Nunnink BD, Cuypers HT. The duration of fixation influences the yield of HCV cDNA-PCR products from formalin-fixed, paraffin-embedded liver tissue. J Virol Methods. 1994;48:267–72.
4. Farragher SM, Tanney A, Kennedy RD, Paul HD. RNA expression analysis from formalin fixed paraffin embedded tissues. Histochem Cell Biol. 2008;130:435–45.

5. Feldman MY. Reactions of nucleic acids and nucleoproteins with formaldehyde. Prog Nucleic Acid Res Mol Biol. 1973;13:1–49.
6. Inadome Y, Noguchi M. Selection of higher molecular weight genomic DNA for molecular diagnosis from formalin-fixed material. Diagn Mol Pathol. 2003;12:231–6.
7. Karlsen F, Kalantari M, Chitemerere M, Johansson B, Hagmar B. Modifications of human and viral deoxyribonucleic acid by formaldehyde fixation. Lab Invest. 1994;71:604–11.
8. von Ahlfen S, Missel A, Bendrat K, Schlumpberger M. Determinants of RNA quality from FFPE samples. PLoS One. 2007;2:e1261.
9. Anisimova VE, Rebrikov DV, Shagin DA, Kozhemyako VB, Menzorova NI, Staroverov DB, et al. Isolation, characterization and molecular cloning of duplex-specific nuclease from the hepatopancreas of the Kamchatka crab. BMC Biochem. 2008;9:14.
10. Morlan JD, Qu K, Sinicropi DV. Selective depletion of rRNA enables whole transcriptome profiling of archival fixed tissue. PLoS One. 2012;7:e42882.
11. Sinicropi D, Qu K, Collin F, Crager M, Liu ML, Pelham RJ, et al. Whole transcriptome RNA-Seq analysis of breast cancer recurrence risk using formalin-fixed paraffin-embedded tumor tissue. PLoS One. 2012;7:e40092.
12. Vandernoot VA, Langevin SA, Solberg OD, Lane PD, Curtis DJ, Bent ZW, et al. cDNA normalization by hydroxyapatite chromatography to enrich transcriptome diversity in RNA-seq applications. Biotechniques. 2012;53:373–80.
13. Zhulidov PA, Bogdanova EA, Shcheglov AS, Vagner LL, Khaspekov GL, Kozhemyako VB, et al. Simple cDNA normalization using kamchatka crab duplex-specific nuclease. Nucleic Acids Res. 2004;32:e37.
14. Norton N, Sun Z, Asmann YW, Serie DJ, Necela BM, Bhagwate A, et al. Gene expression, single nucleotide variant and fusion transcript discovery in archival material from breast tumors. PLoS One. 2013;8:e81925.
15. Paik S, Shak S, Tang G, Kim C, Baker J, Cronin M, et al. A multigene assay to predict recurrence of tamoxifen-treated, node-negative breast cancer. N Engl J Med. 2004;351:2817–26.
16. Ma Y, Ambannavar R, Stephans J, Jeong J, Dei Rossi A, Liu ML, et al. Fusion transcript discovery in formalin-fixed paraffin-embedded human breast cancer tissues reveals a link to tumor progression. PLoS One. 2014;9:e94202.
17. Cobleigh MA, Tabesh B, Bitterman P, Baker J, Cronin M, Liu ML, et al. Tumor gene expression and prognosis in breast cancer patients with 10 or more positive lymph nodes. Clin Cancer Res. 2005;11:8623–31.
18. Morton ML, Bai X, Merry CR, Linden PA, Khalil AM, Leidner RS, et al. Identification of mRNAs and lincRNAs associated with lung cancer progression using next-generation RNA sequencing from laser micro-dissected archival FFPE tissue specimens. Lung Cancer. 2014; 85:31–9.
19. Bourgon R, Lu S, Yan Y, Lackner MR, Wang W, Weigman V, et al. High-throughput detection of clinically relevant mutations in archived tumor samples by multiplexed PCR and next-generation sequencing. Clin Cancer Res. 2014;20:2080–91.
20. Schweiger MR, Kerick M, Timmermann B, Albrecht MW, Borodina T, Parkhomchuk D, et al. Genome-wide massively parallel sequencing of formaldehyde fixed-paraffin embedded (FFPE) tumor tissues for copy-number- and mutation-analysis. PLoS One. 2009;4:e5548.
21. Kerick M, Isau M, Timmermann B, Sultmann H, Herwig R, Krobitsch S, et al. Targeted high throughput sequencing in clinical cancer settings: formaldehyde fixed-paraffin embedded (FFPE) tumor tissues, input amount and tumor heterogeneity. BMC Med Genomics. 2011;4:68.
22. Holley T, Lenkiewicz E, Evers L, Tembe W, Ruiz C, Gsponer JR, et al. Deep clonal profiling of formalin fixed paraffin embedded clinical samples. PLoS One. 2012;7:e50586.
23. Menon R, Deng M, Boehm D, Braun M, Fend F, Biskup S, et al. Exome enrichment and SOLiD sequencing of formalin fixed paraffin embedded (FFPE) prostate cancer tissue. Int J Mol Sci. 2012;13:8933–42.
24. Menon R, Deng M, Ruenauver K, Queisser A, Peifer M, Offermann A, et al. Somatic copy number alterations by whole-exome sequencing implicates YWHAZ and PTK2 in castration-resistant prostate cancer. J Pathol. 2013;231:505–16.

25. Meric-Bernstam F, Frampton GM, Ferrer-Lozano J, Yelensky R, Perez-Fidalgo JA, Wang Y, et al. Concordance of genomic alterations between primary and recurrent breast cancer. Mol Cancer Ther. 2014;13:1382–9.
26. Assie G, Letouze E, Fassnacht M, Jouinot A, Luscap W, Barreau O, et al. Integrated genomic characterization of adrenocortical carcinoma. Nat Genet. 2014;46:607–12.
27. Ross JS, Wang K, Rand JV, Gay L, Presta MJ, Sheehan CE, et al. Next-generation sequencing of adrenocortical carcinoma reveals new routes to targeted therapies. J Clin Pathol. 2014; 67:968–73.
28. Zhao W, He X, Hoadley KA, Parker JS, Hayes DN, Perou CM. Comparison of RNA-Seq by poly (A) capture, ribosomal RNA depletion, and DNA microarray for expression profiling. BMC Genomics. 2014;15:419.
29. Weng L, Wu X, Gao H, Mu B, Li X, Wang JH, et al. MicroRNA profiling of clear cell renal cell carcinoma by whole-genome small RNA deep sequencing of paired frozen and formalin-fixed, paraffin-embedded tissue specimens. J Pathol. 2010;222:41–51.
30. Hedegaard J, Thorsen K, Lund MK, Hein AM, Hamilton-Dutoit SJ, Vang S, et al. Next-generation sequencing of RNA and DNA isolated from paired fresh-frozen and formalin-fixed paraffin-embedded samples of human cancer and normal tissue. PLoS One. 2014;9:e98187.
31. Spencer DH, Sehn JK, Abel HJ, Watson MA, Pfeifer JD, Duncavage EJ. Comparison of clinical targeted next-generation sequence data from formalin-fixed and fresh-frozen tissue specimens. J Mol Diagn. 2013;15:623–33.

Applications of NGS to Screen FFPE Tumours for Detecting Fusion Transcripts

Kunbin Qu, Joffre Baker, and Yan Ma

Abstract Fusion transcripts play an important role in a variety of human cancers. But bioinformatics algorithms that use RNA-Seq data to detect fusions tend to yield high false-positive rates due to both relatively short reads from next-generation sequencing (NGS) and the repetitive elements in human genome. The primary purpose of this chapter is to discuss the design strategy of the bioinformatics methods used for detection of fusion transcripts and to compare their strengths and weaknesses, or "fit for purpose," on RNA-Seq data from non-fixed or formalin-fixed paraffin-embedded (FFPE) tumor tissues. A large number of archival FFPE tumor tissue samples are associated with mature medical records including disease outcome; these samples offer great potential for diagnostic and therapeutic target discovery. However, the chemical treatment by formalin causes RNA degradation and base deamination, which lead to low library complexity and mapping quality. It is important that bioinformatics tools are designed to address these challenges. Here we illustrate a framework to address them, using gFuse as the example. We present results by comparing the fusion transcripts discovered from analysis of RNA from fresh and FFPE MCF-7 breast cancer cells. We also describe the application of gFuse to RNA-Seq data generated from two independent breast cancer cohorts with clinical outcomes and identify candidate fusion transcripts relevant to disease progression.

1 Introduction

Chromosomal abnormalities occur frequently in human tumors. Chromosome translocations and gene fusions were initially discovered in hematological malignancies, in which they define disease subtypes [1]. More recently they have been in soft tissue sarcomas and a variety of solid tumors, including those of prostate [2], lung [3], and breast [4]. Certain recurrent gene fusions are used as cancer diagnostic markers and have been therapeutically targeted with substantial clinical success, for

K. Qu (✉) • J. Baker • Y. Ma
Genomic Health, Inc, Redwood City, CA, USA
e-mail: kqu@genomichealth.com; jbaker@genomichealth.com; yma@genomichealth.com

W. Wu, H. Choudhry (eds.), *Next Generation Sequencing in Cancer Research, Volume 2*, DOI 10.1007/978-3-319-15811-2_10

example, in leukemia and lung cancer [5, 6]. The recent advances in sequencing technology have accelerated the identification of these genetic aberrations.

The history of discovery of the first gene fusion goes back to the 1960s, when Hungerford and Nowell characterized their initial observation that two patients with chronic myelogenous leukemia (CML) had a characteristic small chromosome [7]. This was named the "Philadelphia chromosome" after the city in which it was discovered. The rearrangement is a translocation between chromosomes 9 and 22, resulting in the fusion at the breakpoint cluster region (BCR) gene on chromosome 22 with the v-abl Abelson murine leukemia viral oncogene homolog (ABL1) gene on chromosome 9. In 1990 the BCR-ABL1 fusion protein was characterized as an active tyrosine kinase [8]. Understanding the molecular mechanism of BCR-ABL1 led to the development at Novartis of imatinib (Gleevec), the first targeted cancer therapeutic agent, which recieved FDA approval in 2001 [9]. The entire development time took just 9 years, from small molecule screening to launch. Imatinib binds into the ATP-binding pocket of the fused kinase to inhibit the BCR-ABL1 kinase activity. The mortality rate of CML dropped approximately 90 % after Gleevec became available to treat the disease [10].

The success of treating CML with the specific inhibitor of the BCR-ABL1 fusion led to a strong interest in identifying more novel gene fusions in other cancer types to identify additional disease-specific targets for therapeutics. The discovery of EML4-ALK fusion in non-small cell lung cancer (NSCLC) led to the development of the therapeutic agent crizotinib (Xalkori) [11]. With a prevalence of approximately 5 % in NSCLC, the EML4-ALK fusion increases cellular growth and decreased apoptosis [3]. This fusion defines a subset of NSCLCs, which segregates from mutations in EGFR and appears more commonly in nonsmokers. Prior to the discovery of the ALK (anaplastic lymphoma kinase) fusion, Pfizer had a drug discovery program to inhibit MET for lung cancer. The chemical scaffold from the lead small molecule also showed inhibition to ALK's enzyme activity. Pfizer quickly optimized the chemical structure to be more ALK specific. Only in 3 years, FDA approved Pfizer's Xalkori to treat metastatic stage of NSCLC.

In addition to having therapeutic significance, fusion transcripts can also serve as valuable diagnostic markers. Among various genomic aberrations in cancer, recurrent gene fusions have been identified as a dominant class of mutation in hematological malignancies; they follow a distinct pattern of occurrence based on their origin, lineage, tissue specificity, structure, and function. Gene fusions in lymphomas are commonly associated with an immunoglobulin heavy chain (IGH) gene. Reciprocal translocation results in the overexpression of apparently normal transcripts at an abnormal level driven by the IGH gene regulatory elements [12]. Chromosome rearrangements in leukemia, however, result most commonly in the formation of novel gene fusion transcripts [13]. The molecular characterization of lymphomas and leukemia is now an integral part of the diagnosis. Several molecular abnormalities have been included in the latest World Health Organization classification of hematological malignancies. Similarly, molecular analysis is emerging as a tool for differential diagnosis of soft tissue sarcomas: for example, SS18-SSX fusions in synovial sarcomas, EWSR1 fusions in Ewing's sarcoma, and PAX3/7-FKHR fusions in alveolar

rhabdomyosarcomas. The ETS (erythroblastosis virus E26 transformation-specific) family rearrangement with the androgen-regulated 5′ partner genes in prostate cancer, the EML4-ALK gene fusion in lung cancer, and the RAF family gene rearrangement in a subset of different solid cancers all stratify disease. The selection of therapeutic agent targeting fusion products requires companion diagnostic assay (e.g., selection of NSCLC patients for the treatment with Xalkori requires assay of the EML4-ALK fusion sequence). To date, over 2,000 such tumor-specific gene fusions have been documented (http://cgap.nci.nih.gov/Chromosomes/Mitelman). They are real or potential prognostic biomarkers or drug targets.

2 Fusion Characterization Methods

There are many experimental and computational methods to detect fusion transcripts. Prior to next-generation sequencing (NGS), fusion identification in hematological malignancies depended on conventional cytogenetic karyotyping to detect relatively large chromosome rearrangements. Many more rearrangements have been discovered as a result of use of higher-resolution "molecular cytogenetics," which include fluorescent in situ hybridization (FISH), spectral karyotyping (SKY), multicolor FISH (M-FISH), comparative genomic hybridization (CGH), and high-density array comparative genomic hybridization (a-CGH). These technologies have been extensively applied to almost all cancer types. Non-cytogenetic methods, such as polymerase chain reaction (PCR) and Southern blotting, are also used in the research setting for screening and validation purposes [14].

Both the traditional cytogenetic methods and the non-cytogenetic approaches are based on pre-defined fusion targets [15]. Therefore, they are limited due to the need of prior knowledge and are not suitable for large-scale de novo gene fusion discovery. Also most of them can only be applied to DNA, not RNA. With the introduction of NGS technology, the high-throughput de novo gene fusion discovery is a reality. NGS can assay entire genomes and transcriptomes to exhaustively identify copy number alterations, somatic point mutations, structural rearrangements, and gene expression alterations. Large sample throughput and in-depth sequencing platforms are now widely used for cancer genome characterization. The "The Cancer Genome Atlas" (TCGA; http://cancergenome.nih.gov/) initiative has used NGS to identify DNA and RNA sequence aberrations in at least 25 different cancer types on the whole genome level. Another cost-effective NGS method is to focus on the regions of interest with laboratory enrichment procedures such as Agilent SureSelect Target Enrichment (Agilent Technologies). Some cancer mutation screening panels have predefined fusions as a component of their enrichment targets [16, 17]. Experimentally, anchored multiplex PCR technique has been used to enrich fusion-specific targets at random start points (Enzymatics, Inc.). In conjunction with NGS, it provides potential to scale up fusion transcript screening with a reasonable cost. NGS relies on bioinformatics methods to analyze the massive data output in order to enable the discovery of gene fusions and other somatic sequence variations.

Multiple events at the DNA level can generate fusion transcripts, e.g., chromosomal translocation, interstitial deletion, and chromosomal inversion. Several bioinformatics methods have been developed to detect fusion transcripts from RNA-Seq data (ChimeraScan, SnowShoes-FTD, TopHat-Fusion, FusionMap, and FusionSeq) [18–22]. The greatest computational challenge in identifying fusion transcript is the extraordinary frequency of false positives, which is caused by the direct application of short read mappers. This is due to a combination of numerous repetitive sequences in the genome and the short length of the NGS reads. Short reads make it challenging to map for repetitive regions unambiguously, because they can be aligned to multiple locations on the genome. Another challenge for fusion transcript discovery, often less appreciated, is the occurrence of false negatives. This is evidenced by discordance in results provided by different bioinformatics methods, perhaps reflecting, in part, over-stringent filtering and inaccurate mapping of sequence data.

The well-known problem of the high false-positive rate in de novo fusion discovery makes experimental validation a necessity for the assessment of computational methods. The only FDA-approved fusion companion diagnostics is based on FISH [15]. However, FISH is unpractical to scale up to the level of the NGS' throughput. Traditionally Sanger sequencing has also been used to validate the fusion junction site when sample quality is high. The major mechanism of fusion transcript generation in cancer is genome rearrangement, in which a DNA structural variation brings an mRNA donor site from one gene near an mRNA acceptor site of another gene, such that a consequent alternative splicing event can generate a fusion transcript (Fig. 1). Therefore, the existence of a DNA breakpoint consistent with the observed fusion transcript can provide cross validation of a detected fusion transcript [23–28]. In the case of datasets in which FFPE tissue RNA is available but neither DNA nor fresh tissue RNA is, multiplexed RT-PCR can be used to verify the RNA-Seq-derived fusion candidates [4]. In the example shown in Fig. 2, the RT reaction is multiplexed by using a pooled gene-specific primer set, and each tumor sample is tested with all fusion gene qPCR assays. False positives can be revealed by this multiplexed reverse transcription strategy, and the large number of true negatives serves as a clean background for ascertainment of true positives (Fig. 2). The use of an additional platform technology, such as Ion Torrent sequencing of amplicons derived from the same set of RT-PCR primers, can provide additional supporting evidence and confirm the exact sequence [4].

Here we review three methods in some detail: SnowShoes-FTD, TopHat-Fusion, and gFuse [4, 19, 20]. These methods are selected as "fit for purpose" depending on the goal of the user. ShowShoes-FTD and TopHat-Fusion are developed for high-quality samples, such as fresh tissue. gFuse is developed to analyze formalin-fixed paraffin-embedded (FFPE) tissue. We compare the performance of these three methods by examining the results obtained with fresh and FFPE RNA obtained from the breast tumor cell line MCF-7. Read pairs across a fusion junction are called bridging reads or encompassing reads; a read end that contains a fusion junction is called a split read or a spanning read (Fig. 1). Bridging reads identified from pair-end RNA-Seq data normally prompt nomination of fusion candidates, whereas split reads are used to identify exact junction sequences.

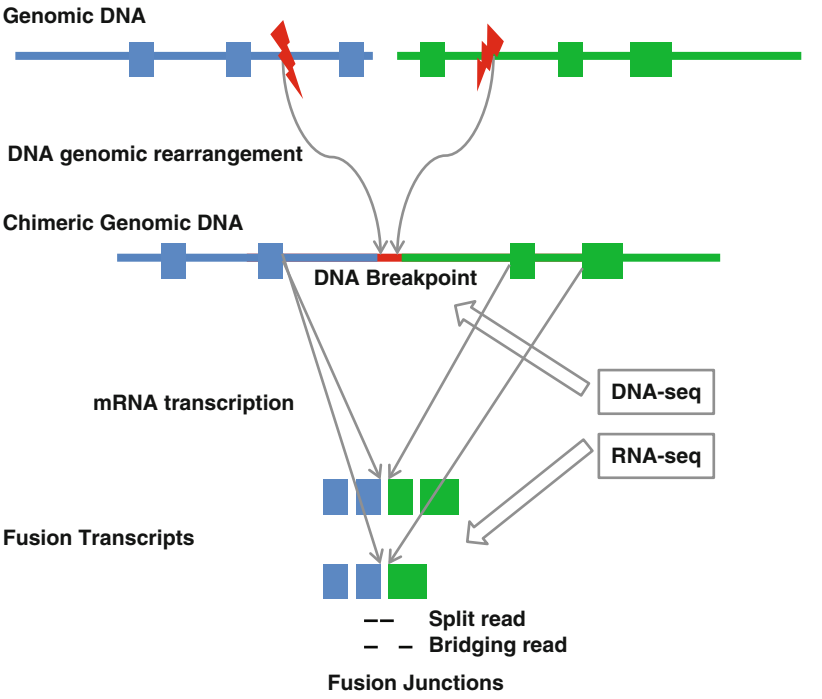


Fig. 1 DNA genomic rearrangement can produce fusion transcripts in cancer. The genomic level rearrangement during cancer development can bring two distant DNA pieces together to generate chimeric genomic DNA. If the DNA sequence around the breakpoint preserves mRNA splicing signals, fusion transcripts can be produced as a result. Two fusion isoforms are illustrated here to reflect that multiple fusion transcripts can be generated from a single DNA rearrangement. DNA-Seq can be used to interrogate DNA breakpoints, and RNA-Seq can be used to interrogate fusion junctions. In RNA-Seq, split reads contain the splicing junction, and bridging reads map to two sides of chimeric transcripts

2.1 SnowShoes-FTD (Fusion Transcripts Detection)

SnowShoes-FTD was developed to predict the fusion transcripts using RefSeq genes as the fusion references. It starts with the creation of the exon-exon boundary database using the exon and gene definition files from University of California at Santa Cruz (UCSC). The database is an exhaustive unidirectional exon junction database that provides the potential fused exon-exon sequence for any pair of RefSeq genes. The FASTA files of exon-exon boundary sequences are compatible with different pair-end sequencing lengths which include 50, 75, and 100 bases.

The major steps of the SnowShoes-FTD package consist of the following: (1) mapping the reads to the reference genome and the exon junction database by BWA or Bowtie, (2) categorizing the aligned read pairs for potential fusion candidates, (3) cleaning the false-positive candidates by a variety of filters, and (4) generating a

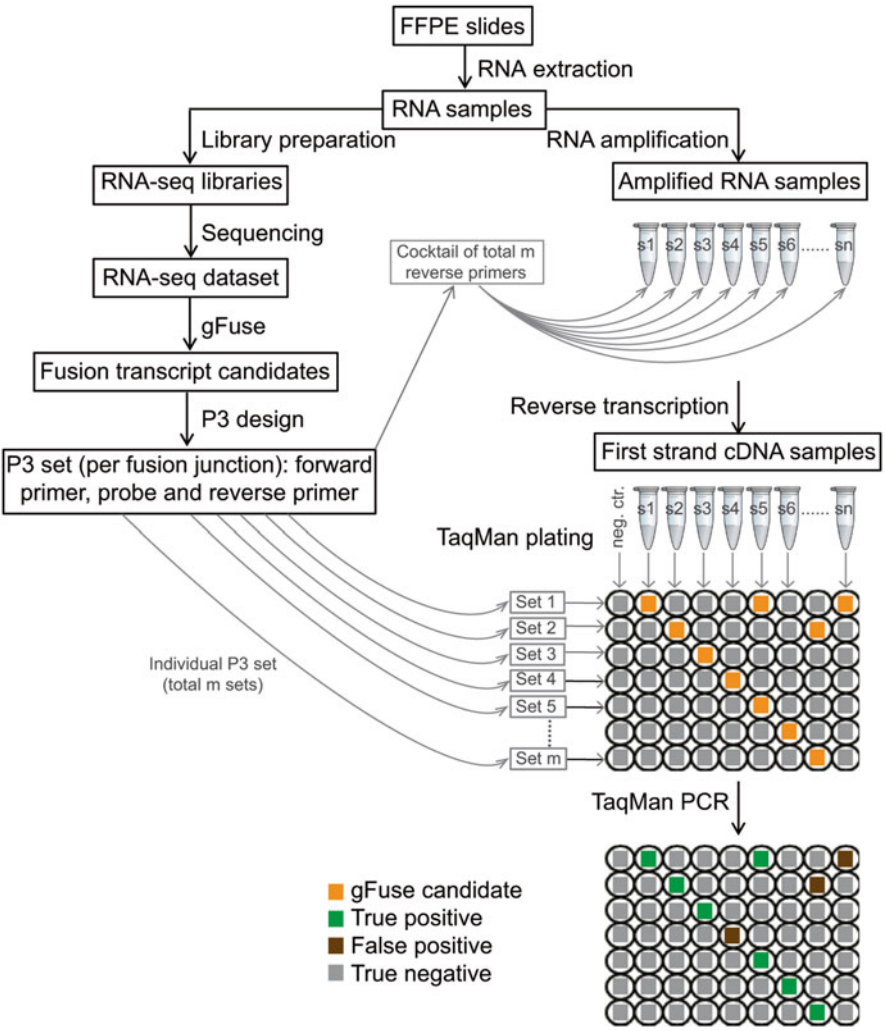


Fig. 2 The flowchart of fusion candidate validation by multiplexed TaqMan RT-PCR assays in the Providence/Rush study [4]. Primers and probes were designed using the Primer3 program restricting amplicon sizes to 65–85 bps (http://frodo.wi.mit.edu/). When Primer3 failed, primer and probe sequences were optimized manually to ensure optimal performance of the TaqMan assay design for the chimeric transcripts. Reverse transcription reaction in the absence of RNA template (i.e., water) was always used as a negative control in all assays. The samples that were previously identified as positive or negative for a particular fusion junction served as controls when needed

continuous sequence region spanning fusion junction points for PCR primer design for experimental validation. The software also suggests the fusion mechanism based on the ordering of the gene annotation and provides information about in-frame versus out-of-frame fusion products.

SnowShoes-FTD identifies the read pairs with ends mapping to two different genes. Then it uses the annotation of the those genes as the templates to align the unmapped read from the read pair whose one end can be mapped to either the genome or the exon junction, to identify the spanning reads. By requiring the existence of both the encompassing reads and spanning reads, and many other filters, it drastically reduces the false-positive rate. SnowShoes-FTD might be too conservative in calling fusion transcripts, for example, by filtering out read pairs of which neither end can be mapped.

SnowShoes-FTD has been widely used in the community. The original paper applied it to 22 breast cancer cell lines and 9 non-transformed cell lines. Fifty-five fusion candidates were identified from the cancer cell lines but none from the normal lines. All the candidates were confirmed by quantitative RT-PCR (qRT-PCR) and some were further verified by Sanger sequencing. The authors compared their results in MCF-7 cell line with other two previous publications which are commonly used for bench mark comparison [29, 30] and showed reasonable agreement. Discordances may be explained as false negatives.

2.2 TopHat-Fusion

TopHat-Fusion was developed based on the RNA-Seq mapping tool TopHat [31]. TopHat can detect and quantify genes from RNA-Seq data without database annotation. TopHat-Fusion has made several major modifications to the original TopHat algorithm, all intended to facilitate fusion transcript discovery. The focus is to identify the fusion junction through the "initially unmapped" (IUM) reads by splitting each IUM read into multiple segments, such as 25 bp a piece, then to map each segment to the different chromosomes or the same chromosome with certain distance threshold. This is the key difference from SnowShoes-FTD. The underlying alignment tool is Bowtie with parameter relaxation to allow the mapped segments to satisfy the characteristics of fusion transcripts, such as across chromosomes as well as inversions. In order to pinpoint the junction region in the fusion, TopHat-Fusion extracts 22 bp on each side of the fusion point and joins them to create 44 bp "spliced fusion contigs," which is similar to SnowShoes-FTD's junction database approach.

TopHat-Fusion implements various strategies to reduce false positives. Similar to SnowShoes-FTD, it requires both the bridging reads and split reads. We discuss two among many of those. The first is how it filters out the repeats; the program extracts the two 23 base sequences spanning each fusion point and then maps them against the entire human genome. The alignment result is kept as a list of pairs with chromosome name and genomic coordinate. For each 23-mer adjacent to a fusion point, the other 23-mer is mapped to determine if it is within 100,000 bp on the same chromosome. If so, then it is likely to be a false positive from the repeat region so that fusion candidate is removed. The second is that TopHat-Fusion tries to assess the uniformity of the reads mapped across the fusion junction. Real fusion

transcripts normally have reads mapped evenly in a wide range across the fusion junction, whereas false-positive fusions often cover much narrower range on the genome [30]. Thus, TopHat-Fusion scans a window of 600 bp around each fusion (300 bp each side) and discards fusion candidates for which the reads fail to span this gap.

TopHat-Fusion was applied to the similar RNA-Seq datasets used by the SnowShoes-FTD to benchmark its results: four breast cancer (including MCF-7) and one prostate cancer (VCaP) cell lines [29, 30]. Overall TopHat-Fusion found 76 fusion transcripts from the four breast cancer cell lines and 19 from the prostate cancer cell line. Among them 25 fusions were previously published [20]. The remaining 51 fusions identified by TopHat-Fusion were not reported previously. To focus on MCF-7 specifically, TopHat-Fusion identified more fusions overlapped with those identified previously by Maher [29] and Edgren [30] than Snowshoes-FTD did.

TopHat-Fusion in general identifies more candidates than SnowShoes-FTD and can perform de novo search without a gene annotation file. The design of TopHat-Fusion enables it to identify any chimeric sequences without examining the splicing signal, and therefore, it is suited for detection of break points in introns. On the other hand, it carries a higher risk to include false positives. TopHat-Fusion can also be applied to single-end read as well as paired-end data, whereas SnowShoes-FTD can only be used for paired-end data.

2.3 gFuse

Both methods discussed above have shown good performance with RNA data generated from fresh tissue/cells, such as cell lines and fresh frozen tissues. However, standard clinical practice generates very large numbers of FFPE tissues from biopsies and surgical resections that have associated, metadata-rich, long-term clinical records. Therefore, analysis of FFPE tissues may reveal fusion transcripts of clinical relevance. RNA from tissues fixed in formalin is altered in a number of ways, including by extensive RNA fragmentation (which continues to worsen as block archival age increases), artifactually large fractions of precursor RNAs, and chemical modifications. As a result, FFPE RNA-Seq libraries have short insert sizes, low complexity (i.e., many short sequence segments with identical nucleotide composition), a large amount of intronic sequence, and randomly erroneous reads from base conversions (C/G → T/A) [32]. The gFuse pipeline was designed, to detect fusion transcripts in RNA-Seq data from archival FFPE samples, by addressing these challenges.

The fusion transcript detection pipeline gFuse uses two strategies, a sample-based strategy and an optional cohort-based strategy (Fig. 3). The sample-based strategy interrogates each RNA-Seq sample individually and nominates candidate fusion junctions. The cohort-based strategy has two features that take advantage of multiple patients/samples within the cohort. The first feature is to combine the candidate fusion junctions in the beginning step of the cohort-based analysis, which increases

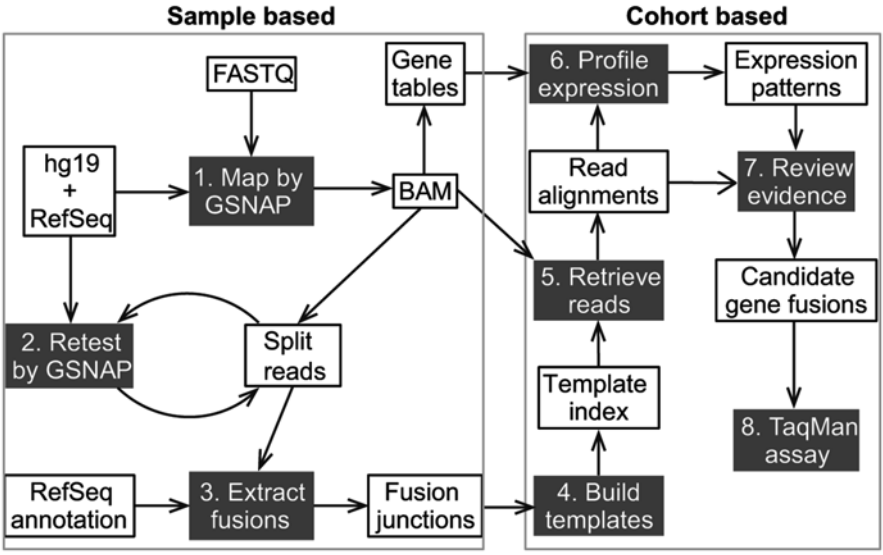


Fig. 3 The schema of gFuse, illustrated for the Providence/Rush study. The sample- and cohort-based strategies are integrated in RNA-Seq fusion transcript detection. Each step of the pipeline is numbered in *shade* and explained in Sect. 2.3 [4]

the chance of identifying recurrent fusion transcripts. The second feature is to confirm the presence of each fusion candidate in each individual sample across the entire cohort pool by examining read alignment and expression profiling evidence.

gFuse uses GSNAP [33] as the underlying mapper, which can detect a distant splice junction within a split read. Distant spliced junctions include the following categories: (1) within the same gene, but in the opposite transcription direction, (2) across different genes, and (3) across different chromosomes. The major steps in gFuse are as follows: (1) map (by GSNAP) to detect split reads; (2) retest (by GSNAP) the read candidates from the previous step with preference of local alignment, to filter out false positives; (3) extract fusion candidates based on the retained reads from the previous step to build an annotated database for the next step; (4) build junction alignment templates based on the annotated candidates for the next step; (5) retrieve reads including split reads and bridging reads when in paired-end mode and realign them to the junction templates from step 4 to enhance sensitivity and specificity; and (6) profile expression to exploit the observation that if there is a fusion it is highly likely to exhibit a marked expression discontinuity between the preserved side and discarded side of a given fusion junction. In this step, candidates from multiple cohorts (if available) are pooled together to increase sensitivity. Since more than 50 % of RNA-Seq reads from older FFPE tissue samples map to introns [34], intron reads are also included to calculate the expression. The performance of gFuse is discussed in more detail in the following section.

3 The Effect of FFPE on Fusion Discovery in a Cell Line Model

The breast cancer cell line MCF-7 has been well studied with abundant knowledge about the genomic rearrangements and fusion transcripts [20, 30, 35–37]. This cell line has also been used as a benchmark to assess multiple fusion detection methods [19, 20]. We prepared FFPE MCF-7 and archived them at room temperature for 9 years. This FFPE MCF-7 sample provides a model system to assess the impact of formalin treatment on fusion detection (Fig. 4a). With the availability of paired fresh and FFPE MCF-7 samples, we compare the performance of the three fusion detection programs discussed above (Fig. 4b, c), to illustrate the extent to which FFPE fixation hampers fusion detection and to demonstrate how this can be mitigated.

3.1 Fresh MCF-7 Cell Line

The RNA-Seq library was prepared using TruSeq RNA Sample Preparation Kit (Illumina Inc.) and sequenced in a HiSeq 2000 (Illumina Inc.) sequencer to a depth of 93 million by 2×50 bp read pairs. Three fusion detection algorithms including SnowShoes-FTD, TopHat-Fusion, and gFuse (sample-based approach only) were applied to this RNA-Seq dataset (Fig. 4b). In order to compare the results across these three methods, fusion candidates annotated in the intron region from TopHat-Fusion were removed from the following analysis.

Recently 25 fusion transcripts have been reported and validated by quantitative RT-PCR in MCF-7 cell line, which is the biggest set so far [38]. In order to expand this "truth set," we combined the 25 known fusion transcripts with the DNA-Seq breakpoint cross-validated fusion transcripts (Fig. 4a). We took a fusion transcript-guided approach to manually assemble DNA breakpoints which could provide a mechanistic explanation for the fusion transcript.

MCF-7 was sequenced by both transcriptome and whole genome sequencing (WGS) methods in our laboratory. Starting from the 57 fusion transcripts identified by gFuse from RNA-Seq data (Table 1), we identified all WGS paired reads which bridge both donor and acceptor genes. Each of these paired reads was evaluated manually to identify the portion of DNA sequence that contains two genes and harbors the DNA breakpoint. Multiple reads covering this portion of DNA sequence were recovered and assembled manually. The assembled DNA sequence was verified by aligning to the genome with BLAT (https://genome.ucsc.edu/index.html) to confirm the location of DNA breakpoints. Another feature we used to check the assembly result was to recognize the abrupt read coverage change at putative DNA breakpoints, which is very similar to the expression profiling in gFuse and reflects the discontinuous of the genome sequence at disturbed sites. After careful examination, we called 30 DNA breakpoints corresponding to 36 fusion transcripts, 18 overlapping with the published 25 fusion transcripts. A full automatic systematic search of MCF-7 WGS dataset for DNA breakpoints didn't confirm additional fusion transcripts called by gFuse, or Snowshoes-FTD, or TopHat-Fusion beyond these 36 verified fusion transcripts, suggesting that this manual assembly approach

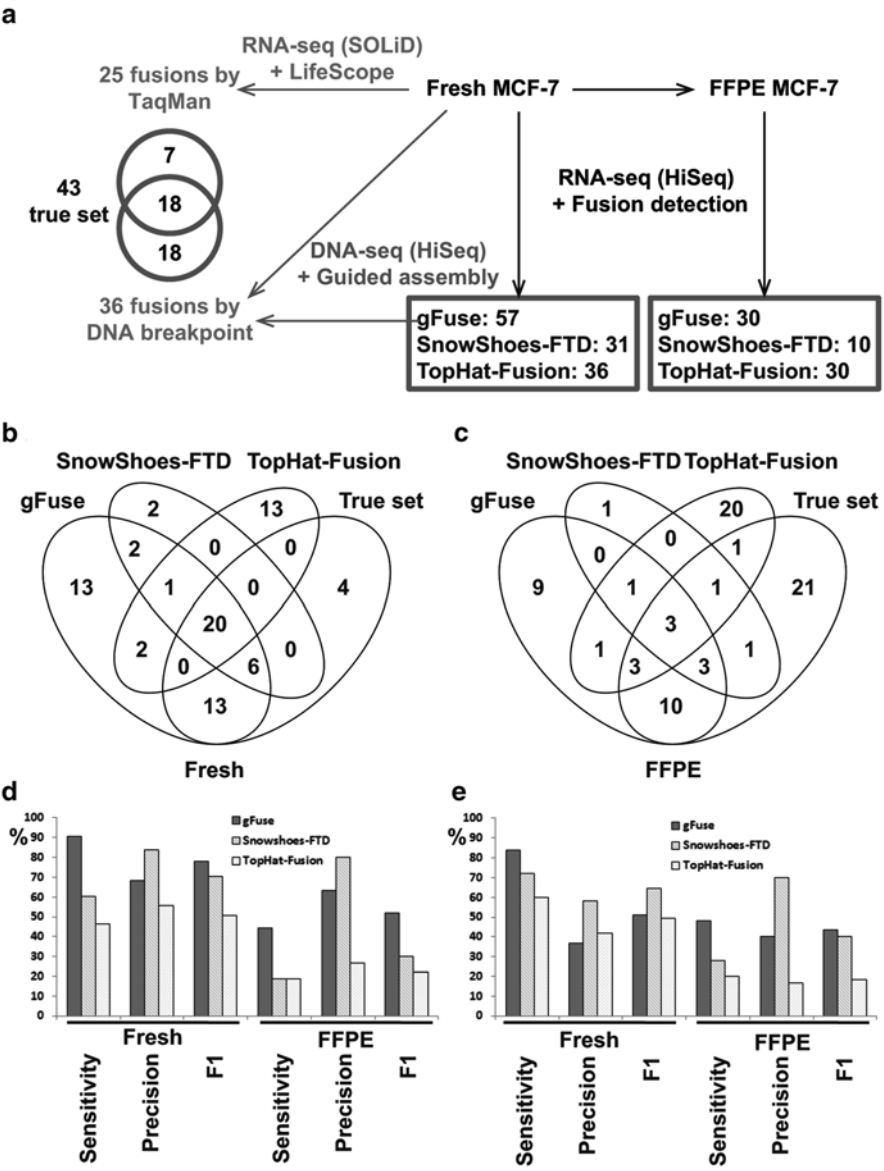


Fig. 4 Cross-platform comparisons of gene fusion transcript detection in fresh and FFPE MCF-7 demonstrate the challenge of fusion detection in FFPE samples. (**a**) The flowchart of seven compared datasets. A true set consisting of 43 fusion transcripts are generated by combining 25 fusions validated by TaqMan [38] and 36 fusions validated by DNA breakpoints. (**b**) A Venn diagram of the results from three bioinformatics methods plus the true set in fresh MCF-7. (**c**) A Venn diagram of the results from three bioinformatics methods plus the true set in FFPE MCF-7. (**d**) The performance of three bioinformatics methods is assessed by sensitivity, precision, and F1 score in fresh and MCF-7 samples when the 43 true set is used as the comparator. (**e**) The performance of three bioinformatics methods is assessed by sensitivity, precision, and F1 score in fresh and FFPE MCF-7 samples when 25 TaqMan is used as the comparator

Table 1 Fusion transcripts identified by gFuse from fresh MCF-7 cell line

Fusion gene	Fusion junction	True set (TaqMan[a] and DNA breakpoint)	gFuse	SnowShoes-FTD		TopHat-Fusion	
			FFPE	Fresh	FFPE	Fresh	FFPE
ABCA5>PPP4R1L	−chr17:67309233→−chr20:56847931	TaqMan;−chr17:67305620_4DUP_−chr20:56848361	Yes	Yes		Yes	
ADAMTS19>SLC27A6	+chr5:128796175→+chr5:128364038	TaqMan;−chr5:128363760_4DUP_−chr5:128836260		Yes		Yes	
ADAMTS19>SLC27A6	+chr5:128797450→+chr5:128364038	TaqMan;−chr5:128363760_4DUP_−chr5:128836260	Yes	Yes		Yes	
AHCYL1>RAD51C	+chr1:110527794→ +chr17:56811479	TaqMan;+chr1:110547809_1DUP_+chr17:56810937	Yes				
ANKS1A>UHRF1BP1	+chr6:34857376→+chr6:34801994	−chr6:34797722_1DUP_−chr6:34883016	Yes				
APPBP2>PPM1D	−chr17:58603155→+chr17:58700882						
ARFGEF2>SULF2	+chr20:47538547→−chr20:46365686	TaqMan;+chr20:46378557_−chr20:47540643	Yes	Yes	Yes	Yes	Yes
ATL2>LINC00330	−chr2:38604285→−chr13:45379166						
ATP1A2>CSDE1	+chr1:160105392→−chr1:115263338						
ATXN7L3>FAM171A2	−chr17:42275388→−chr17:42433956	TaqMan;−chr17:42275371_7DUP_−chr17:42435579	Yes				
BAGE>UBA52	−chr21:11059726→+chr19:18684103						
BCAS3>ATXN7	+chr17:58824640→+chr3:63981181					Yes	
BCAS4>BCAS3	+chr20:49411710→+chr17:59445688	−chr17:59287702_4DUP_+chr20:58845400; +chr20:49430579_3DUP_−chr20:58845848[b]	Yes	Yes		Yes	Yes
BCAS4>BCAS3	+chr20:49411710→+chr17:59469338	TaqMan;−chr17:59287702_4DUP_+chr20:58845400; +chr20:49430579_3DUP_−chr20:58845848[b]	Yes	Yes			
BCAS4>ZMYND8	+chr20:49411710→−chr20:45850125						
C16orf45>ABCC1	+chr16:15528616→+chr16:16170183	TaqMan;+chr16:15593076_2DUP_+chr16:16166573		Yes		Yes	
CAPZB>EEF2	−chr1:19811930→−chr19:3984348						
CDH15>CDH1	+chr16:89260014→+chr16:68863557		Yes				
DEPDC1B>ELOVL7	−chr5:59934577→−chr5:60053472	TaqMan;−chr5:59929635_3DUP_−chr5:60057483		Yes		Yes	
ESR1>CCDC170	+chr6:152023140→+chr6:151857453	−chr6:151853473_21INS_−chr6:152086162					
ESR1>CCDC170	+chr6:152023140→+chr6:151894309	TaqMan;−chr6:151884552_5DUP_−chr6:152045414		Yes		Yes	
GATAD2B>NUP210L	−chr1:153895209→−chr1:154000073		Yes	Yes			

GATAD2B>NUP210L	−chr1:153895209→−chr1:154002530	TaqMan		Yes	Yes	Yes	Yes
GCN1L1>MSI1	−chr12:120628101→−chr12:120785317	TaqMan;−chr12:120626069_2INS_−chr12:120789654	Yes	Yes		Yes	
GCN1L1>MSI1	−chr12:120628101→−chr12:120789203	−chr12:120626069_2INS_−chr12:120789654		Yes			
HIPK1>DENND2C	+chr1:114484081→−chr1:115161103					Yes	Yes
MLL3>TPTE	−chr7:151902191→−chr21:10987877						
MYO6>SENP6	+chr6:76459140→+chr6:76388299	TaqMan;−chr6:76387997_3DUP_−chr6:76479962	Yes	Yes	Yes	Yes	Yes
MYO9B>FCHO1	+chr19:17213367→+chr19:17881234	+chr19:17226392_2DUP_+chr19:17880014	Yes	Yes	Yes		
MYO9B>FCHO1	+chr19:17213367→+chr19:17881601	+chr19:17226392_2DUP_+chr19:17880014					
PAPOLA>AK7	+chr14:96968937→+chr14:96904172	+chr14:96981506_33INS_+chr14:96895053	Yes	Yes		Yes	Yes
PARD6B>BCAS3	+chr20:49348389→+chr17:59445688		Yes	Yes			
PDE8A>SCAND2	+chr15:85525579→ +chr15:85180578	−chr15:85178652_3DUP_−chr15:85527687	Yes				
PLCG1>TOP1	+chr20:39766498→+chr20:39729849	−chr20:39729440_20INS_−chr20:39782017					
POP1>MATN2	+chr8:99129618→+chr8:99042690	TaqMan;−chr8:99042398_25INS_−chr8:99133087		Yes			
RAD51C>ATXN7	+chr17:56801461→+chr3:63965591	−chr3:63962499_4DUP_−chr17:56808562					
RBM6>BCAR1	+chr3:50036946→−chr16:75276988	+chr3:50069871_2DUP_−chr16:75282029		Yes			
RPS6KB1>DIAPH3	+chr17:58007535→−chr13:60240980	TaqMan;+chr13:60320934_9INS_−chr17:58008048	Yes	Yes			
RPS6KB1>VMP1	+chr17:57987972→+chr17:57915656	TaqMan;−chr17:57914774_36INS_−chr17:57988497	Yes	Yes		Yes	
RPS6KB1>VMP1	+chr17:57990165→+chr17:57915656						
RPS6KB1>VMP1	+chr17:57990165→+chr17:57917129	−chr17:57915876_−chr17:57999131					
RPS6KB1>VMP1	+chr17:57992064→+chr17:57917129	−chr17:57915876_−chr17:57999131	Yes	Yes		Yes	
SLC35B3>ANP32B	−chr6:8435672→+chr9:100774703						
SMARCA4>CARM1	+chr19:11097269→+chr19:11015627	TaqMan;−chr19:11001041_1DUP_−chr19:11097675		Yes		Yes	
SULF2>PRICKLE2	−chr20:46414579→−chr3:64085601	+chr3:64102178_+chr20:46411670					
SULF2>PRICKLE2	−chr20:46414792→−chr3:64085601	+chr3:64102178_+chr20:46411670		Yes		Yes	
SYTL2>PICALM	−chr11:85468668→−chr11:85685855	TaqMan					
TAF4>BRIP1	−chr20:60639507→−chr17:59924581	+chr17:59928989_2DUP_+ chr20:(60632861-60633817)[c]	Yes	Yes		Yes	Yes

(continued)

Table 1 (continued)

Fusion gene	Fusion junction	True set (TaqMan[a] and DNA breakpoint)	gFuse	SnowShoes-FTD		TopHat-Fusion	
			FFPE	Fresh	FFPE	Fresh	FFPE
TAF4>BRIP1	−chr20:60639507→−chr17:59926617	TaqMan;+chr17:59928989_2DUP_+chr20:(60632861-60633817)[c]	Yes	Yes	Yes	Yes	
TANC2>CA4	+chr17:61151375→+chr17:58232675		Yes	Yes	Yes	Yes	Yes
TBL1XR1>RGS17	−chr3:176914909→−chr6:153365178	TaqMan	Yes	Yes	Yes	Yes	Yes
TMOD4>VPS72	−chr1:151145975→−chr1:151149507						
TWSG1>ANKRD12	+chr18:9337350→+chr18:9275522	−chr18:9267246_2DUP_−chr18:9359451					
TXLNG>SYAP1	+chrX:16804712→+chrX:16753350	TaqMan;−chrX:16748446_3DUP_−chrX:16813167	Yes	Yes	Yes	Yes	
TYW1>TYW1B	+chr7:66548526→−chr7:72209578						
UBE2T>AXDND1	−chr1:202311023→+chr1:179437578	−chr1:179416879_1INS_+chr1:202308345					
ZNF664>LINC00330	+chr12:124472685→−chr13:45379166						

The complete list of 57 fusions identified by gFuse from fresh MCF-7 cell line is compared with results from SnowShoes-FTD, TopHat-Fusion in fresh, and FFPE samples. The definition of symbols used to define fusion junctions is as follows: "→" indicates the splicing direction from donor to acceptor, "+" indicates the transcription direction on the top of chromosome strand, and "−" indicates the transcription direction on the bottom of the chromosome strand. The donor genomic position is the last base of the preserved side of the donor, and the acceptor genomic position is the first base of the preserved side of the acceptor. Each DNA breakpoint is separated into two (*left*, *right*) or three (*left*, *middle*, *right*) segments based on the chromosomal locations as well as chromosome numbers ordered from chromosome 1–22, X and Y, with the left segment preceding the right segment. In the *left segment*, "+" indicates that the segment runs along the top strand and stops at the indicated genomic location, and "−" indicates that the segment runs along the bottom strand and stops at the indicated genomic location. In the *right segment*, "+" indicates that the segment starts at the indicated genomic location and runs along the top strand, and "−" indicates that the segment starts at the indicated genomic location and runs along the bottom strand. The *middle short segment* labeled with "DUP" indicates this segment can be mapped to either left or right segment (the genomic mapping of left or right segment is greedy and each contains the middle segment), and the *one labeled with* "INS" indicates this segment cannot be mapped to either left or right segment

[a]Fusion transcripts were identified by LifeScope on SOLiD platform and validated by TaqMan from Sakarya et al. PLOS Computational Biology 2012 [38]

[b]Two DNA breakpoints connected by a 449 bp DNA segment correspond to these fusion transcripts

[c]Due to multiple local repeats, the precise location of one DNA breakpoint cannot be determined

was not biased towards gFuse results. We combined this set with the original 25 and defined the combined 43 fusion transcripts as the MCF-7 truth set (Fig. 4a).

Three statistical measures including sensitivity, precision, and F1 score are compared across the three methods (Fig. 4d, e). The F1 score is the geometric mean of sensitivity and precision, this providing an assessment of the accuracy of a test. TopHat-Fusion has the lowest scores in all assessed measures in both fresh and FFPE samples and therefore is the poorest performer among three methods (Fig. 4d). Snowshoes-FTD has slightly higher precision than gFuse in fresh MCF-7 sample, but its sensitivity lags significantly behind gFuse. gFuse has 1.5 times (90.7 % vs. 60.5 %) greater sensitivity than Snowshoes-FTD in fresh MCF-7. Overall, the accuracy of the methods measured by F1 scores is higher in gFuse than in Snowshoes-FTD (Fig. 4d). Thus, the sample-based part of gFuse demonstrates an attractive balance between sensitivity and precision in discovering fusion transcripts in this sample.

3.2 FFPE MCF7

We carried out RNA-Seq using RNA from the archival FFPE MCF-7 cells, sequencing to a depth of 250 million by 2×50 bp read pairs, and compared the results to those obtained using RNA from the unfixed MCF-7 cells (Fig. 5a).

Overall, fewer fusion transcripts were detected with FFPE RNA, regardless of bioinformatics pipeline: 40.4 %, 32.2 %, and 19.4 % of fusions identified in unfixed cells were rediscovered in FFPE cells, by gFuse, ShowShoes-FTD, and TopHat-Fusion, respectively (Fig. 5a). Within the 43 truth set, the discovery rate in FFPE cells was 44.2 %, 18.6 %, and 18.6 % by gFuse, SnowShoes-FTD, and TopHat-Fusion, respectively. Thus, among the three methods compared here, gFuse exhibited superior performance for fusion transcript discovery from the archival FFPE cells. The primary reason of the general decrease in fusion discovery in FFPE cells, relative to unfixed cells, might be the reduced library complexity caused by extensive RNA fragmentation. Although several features of gFuse compensate for certain features of RNA-Seq data generated from FFPE RNA, such as higher error rate and higher intron percentage, bioinformatics cannot offer a solution to the fragmentation issue.

When we use the 25 validated fusion transcript set as the comparator, the performance difference between gFuse and SnowShoes-FTD is less obvious (Fig. 4e). SnowShoes-FTD nominates fewer candidates which leads to low sensitivity and high specificity for both fresh and FFPE MCF-7 cell lines. SnowShoes-FTD has higher F1 in fresh and slightly lower in FFPE, comparing with gFuse, whereas TopHat-Fusion is still in the third place (Fig. 4e). With the performance trade-off for each method, fit for purpose becomes important when choosing which algorithm to use.

It is noteworthy that a small number of new fusion transcripts were detected by gFuse in the FFPE cells. Although this might be due to the deeper read coverage (2.7 times) in FFPE cells, the more likely explanation is that the low quality of the FFPE reads caused inaccurate mapping and that these new calls are false positives. The Circos plot representing gFuse-predicted MCF-7 fusion junctions is consistent with BAC library data (Fig. 5b) [35].

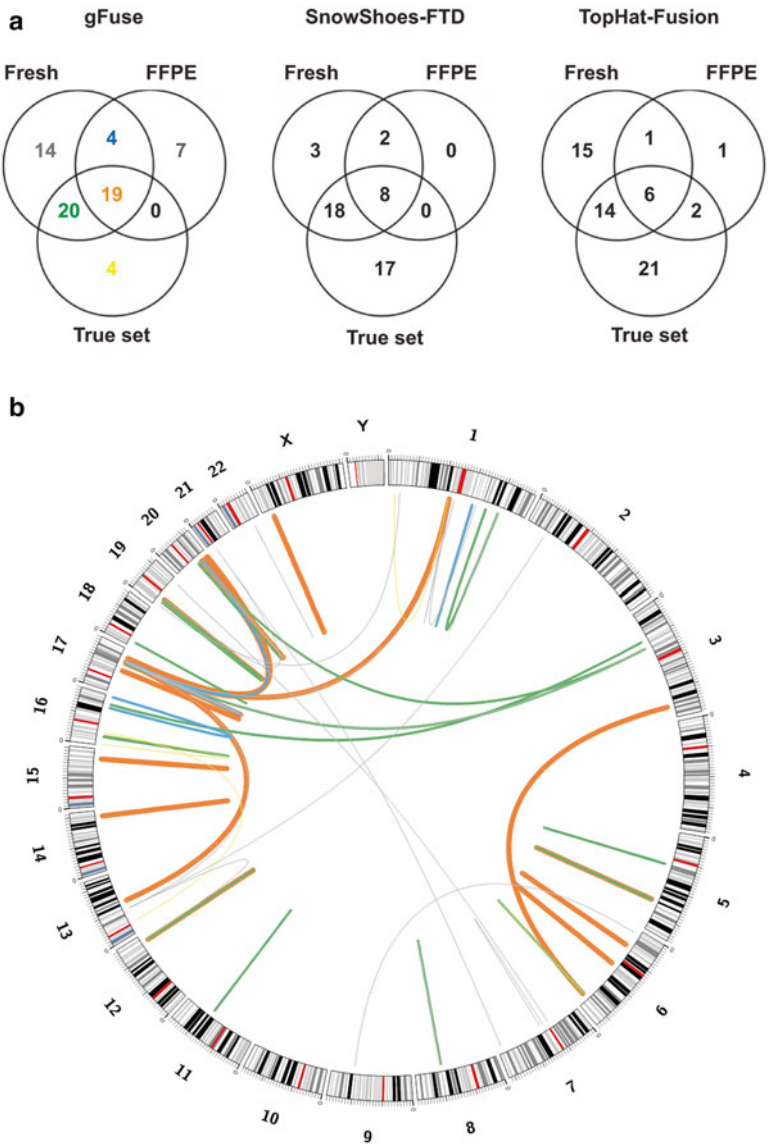


Fig. 5 (**a**) Venn diagrams of fresh and FFPE MCF-7 plus the 43 true set by each of the three bioinformatics methods, respectively. (**b**) The gFuse-identified fusion candidates represented by the Circos plot are from both fresh and FFPE samples, and the colors correspond to gFuse in (**a**)

4 Application to FFPE Cancer Tissue

We applied gFuse to detect fusion transcripts in two breast cancer clinical cohorts with records of clinical outcomes, namely, the Providence cohort of 136 patients (provided by Providence St. Joseph Medical Center, Burbank, CA) and the Rush

cohort of 76 patients (provided by Rush University Medical Center, Chicago, IL) with average FFPE block archive ages of 8.5 years and 13.4 years, respectively [34, 39]. These two cohorts have been previously used by Genomic Health in the development of the 21-gene qRT-PCR breast cancer recurrence risk assay.

Overall, 118 fusion events, representing 100 unique fusion junctions, were identified across the two cohorts. A total of 60 fusion junction candidates were selected for confirmation assay by qRT-PCR. A few of the candidate fusions selected for TaqMan assay were observed in multiple samples. By using 60 designs, 77 candidate fusion events were tested by qRT-PCR in amplified RNA from those patients harboring the corresponding candidate fusions. A total of 47 of the 77 fusion events (61 %) were validated by TaqMan across the two cohorts. To further confirm fusion junctions identified by the TaqMan assays, 19 fusion events identified by TaqMan were selected for Personal Genome Machine (PGM, Ion Torrent) sequencing. Fusion junctions were amplified using TaqMan primers, and PCR products containing fusion amplicons were sequenced on the PGM. In all 19 PCR reactions, the PCR amplicons matched the predicted fusion junction sequences.

Using the cohort-based approach in gFuse, three recurrent fusion events including TFG → GPR128, ESR1 → AKAP12, and RABEP1 → DNAH9 were identified and verified in 6, 3, and 2 patients, respectively, in the two cohorts of 212 total patients. Among the three ESR1 → AKAP12 fusion events in three different patients, there are two unique fusion junctions with the same acceptor junction site but differing at the donor junction sites by one exon (Fig. 6). Since both of these ESR1 → AKAP12 fusion junctions are in frame and the differing ESR1 exon doesn't contain any known functional domains, these two fusion transcripts may function similarly. Both fusion protein isoforms replace the ESR1 ligand binding site with functional domains from AKAP12. The lost ligand binding site of ESR1 is documented to interact with AKAP13, another AKAP family member. AKAP12 is a scaffold protein present in the plasma membrane, cytosol, or endoplasmic reticulum for protein kinases A and C which regulates actin-cytoskeleton reorganization [40]. It has been reported that AKAP12 also is a tumor suppressor with recurrent loss in colorectal cancer and re-expression of AKAP12 inhibits cancer progression and decreases metastasis potential [41]. Since the functional domains of AKAP12 are preserved in these fusion protein isoforms, we postulate that the fused AKAP12 protein might undergo functional alteration, with the fused ER protein perhaps impacting its cellular localization. In addition, both fusion protein isoforms may induce constitutive ligand-independent signaling. In consequence, the patients with ESR1 → AKAP12 fusion may show different responses to breast cancer hormone therapy.

The patients from the two clinical cohorts were stratified based on the number of fusion events identified. The patients with more than two fusions demonstrated significant increased recurrence risk compared to patients with fewer detected fusion genes [4]. To determine whether gene expression profiles can distinguish between the 82 tumors without detected fusion transcripts and the 8 tumors with multiple gene fusion transcripts in the Providence cohort, the differentially expressed genes were analyzed by edgeR [42] based on gene tables tallied from GSNAP mapping results. Due to the strong influence of estrogen receptor (ER) status on gene expression patterns [43], the additive model of edgeR was used to obtain differentially expressed genes between multiple fusion cases versus cases with no fusions, adjusting for differences in ER

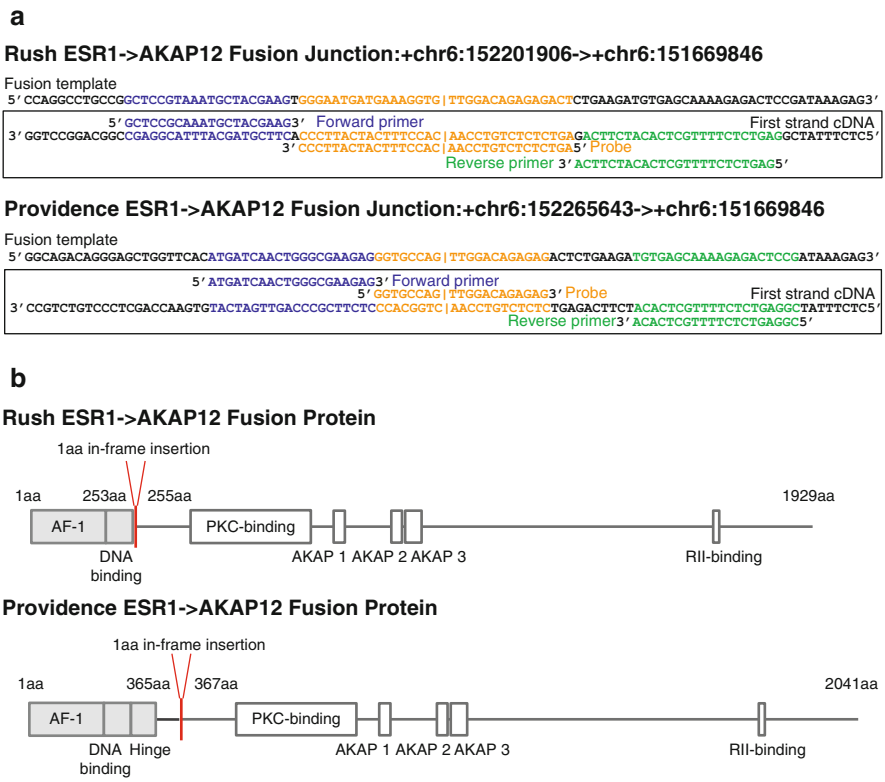


Fig. 6 Recurrent gene fusion ESR1→AKAP12 has two predicted fusion protein isoforms [4]. (**a**) TaqMan designs for two predicted fusion junctions. (**b**) Protein domains of two protein isoforms ESR1→AKAP12 are illustrated based on UniProt database (www.uniprot.org). The protein domains of ESR1 are from protein P03372 (UniProt ID). The protein domains of AKAP12 are from protein Q02952 (UniProt ID). The *red vertical line* indicates the fusion position on the corresponding protein. The one amino acid insertion generated from the fusion event is labeled on each fusion protein. The amino acid length and amino acid positions of each fusion position are labeled on the *top* of each protein

status. A set of 134 genes that were differentially expressed between tumors with no observed fusions and tumors with multiple observed fusions was uploaded to the Reactome FI (functional interaction) database via the Cytoscape Plugin [44]. This reveals a protein interaction network having 84 genes distributed in five functional modules (Fig. 7a). These 84 genes are all upregulated in the multiple fusion group

Fig. 7 (continued) Nodes are manually arranged to display the sub-modules properly. Edges display FI direction attribute values as the following: "→" for activating/catalyzing, "-|" for inhibition, "-" for FIs extracted from complexes or inputs, and "---" for predicted FIs. (**b**) Fusion signature indexes are plotted for each of the fusion number categories in Providence and Rush datasets. The fusion signature index is the average expression level of 84 fusion gene signatures. The base counts of each signature gene are normalized by library size then scaled across the patient cohort before averaged in the signature index. The *p*-values are derived from Wilcoxon tests

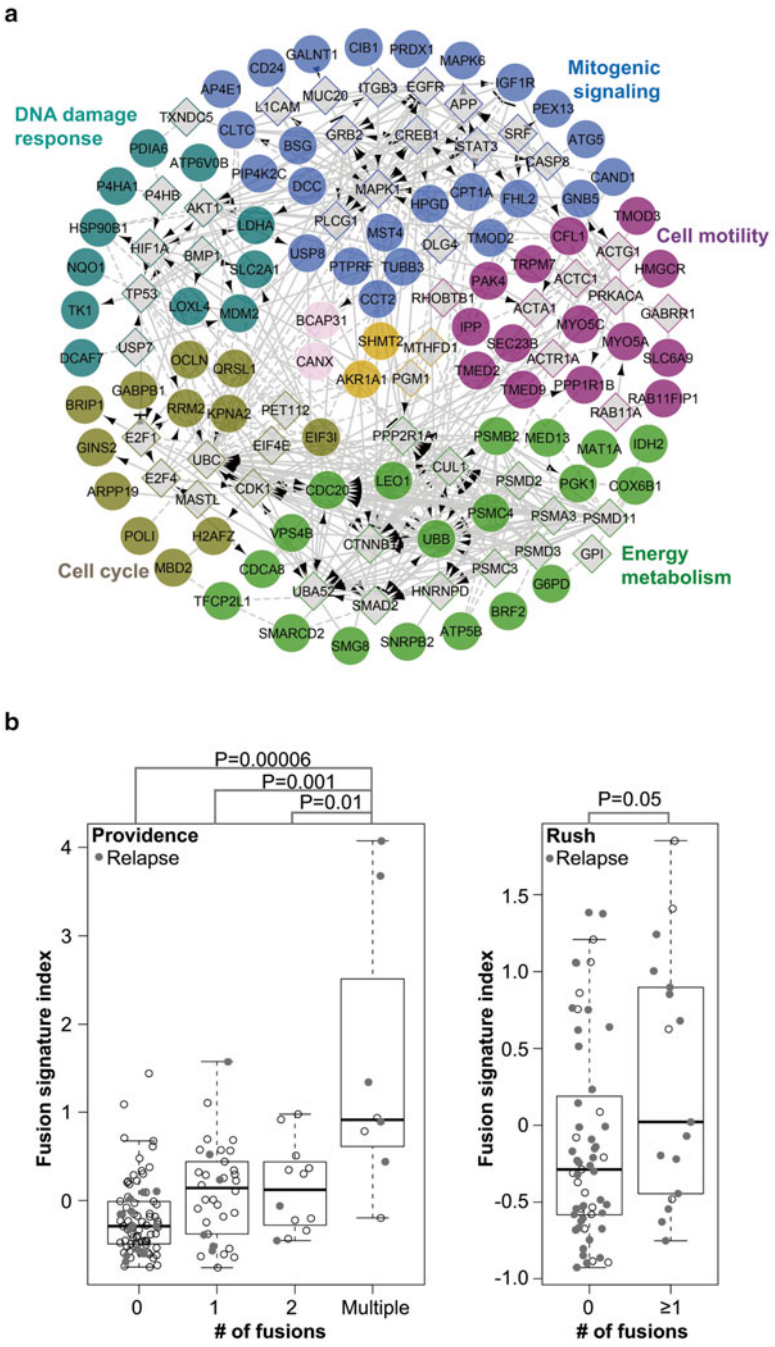


Fig. 7 Fusion gene index level is associated with fusion frequency across patient cohorts. (**a**) The differentially expressed genes between multiple fusion samples versus no fusion samples in the Providence cohort are mapped to the Reactome FI database and clustered into five core sub-modules represented by different colors via linker proteins (in *gray-shaded rectangles*) by the Reactome FI Cytoscape Plugin. The connected 84 genes are referred as the fusion gene signature.

compared to no fusion group. Strikingly their functions are all carcinoma related, in pathways representing some of the prominent pathological hallmarks of cancer [45]. Therefore, we term this network of 84 genes the fusion gene signature. The fusion signature index is the average normalized expression of these 84 genes. The fusion signature index in Providence patient tumors was on average significantly greater in patients with multiple fusions than in patients with 1 or 2 detected fusions (Fig. 7b). Further, in the Rush cohort the expression of this signature is on average significantly greater in tumors with identified fusions (Fig. 7b).

This study demonstrates the technical feasibility and potential biomedical value of being able to detect fusion transcripts in archival FFPE tumor specimens having attached clinical records. Although the average frequency of detected fusion transcripts is relatively low per patient, plausibly attributable to the low quality of FFPE RNA-Seq libraries, the frequency of fusion events found in our cohort nevertheless appears to have prognostic significance. This is further supported by an identified breast cancer fusion gene signature enriched with genes that have functions associated with tumor progression. Many of the identified fusion partner genes belong to the kinase, phosphatase, and ubiquitin ligase families, which are attractive pharmaceutical targets in oncology. The association of fusion frequency with disease prognosis likely reflects the link between chromosome rearrangements and genome instability.

5 Conclusion

Cancer arises from diverse genetic alterations, including gene fusions. Identifying recurrent gene fusions carries great potential to discover targets for both the diagnosis and treatment of various cancers. RNA-Seq provides a wealth of data for fusion transcripts discovery. However, because of the high frequency of repetitive sequences in human genome and the short length of NGS reads, false positives plague data mining for fusion transcripts. Many computational algorithms such as SnowShoes-FTD and TopHat-Fusion have been specifically designed to address these issues with great success in fresh tissues. RNA from archival FFPE tissues presents extra challenges as library complexity tends to be low, insert sizes short, and intron percentages high. With these in mind, gFuse was designed to work well with FFPE RNA. It is important to understand the methods' performance trade-off so the best tool can be selected for a specific application.

Cancer gene fusions could be either disease drivers or passengers. If the fused genes are drivers, they tend to be recurrent and may represent therapeutic targets for disease intervention, such as BCR-ABL and EML4-ALK fusions. However, many fusions detected from multiple cancer types are rare or private, i.e., only appear within individual patients [24–28]. Such gene fusions are likely passengers, but they could still serve as biomarkers for diagnosis and disease monitoring. Our results show that the prognosis of breast cancer is associated with the number of fusion events, independent of the identities of the fused sequences [4]. Furthermore, tumor volumes

are associated with the fusions detected in the plasma from lung cancer patients [17]. One big advantage of using gene fusions over single nucleotide variations (SNV) as biomarkers is that the sequence signature of the former is much less susceptible to sequencing errors. As the sequencing technology evolves, we anticipate that the therapeutic and diagnostic importance of gene fusions will increase proportionally.

References

1. Pui C-H, Robison LL, Look AT. Acute lymphoblastic leukaemia. Lancet. 2008;371:1030–43. doi:10.1016/S0140-6736(08)60457-2.
2. Tomlins SA, Rhodes DR, Perner S, et al. Recurrent fusion of TMPRSS2 and ETS transcription factor genes in prostate cancer. Science. 2005;310:644–8. doi:10.1126/science.1117679.
3. Soda M, Choi YL, Enomoto M, et al. Identification of the transforming EML4-ALK fusion gene in non-small-cell lung cancer. Nature. 2007;448:561–6. doi:10.1038/nature05945.
4. Ma Y, Ambannavar R, Stephans J, et al. Fusion transcript discovery in formalin-fixed paraffin-embedded human breast cancer tissues reveals a link to tumor progression. PLoS One. 2014;9:e94202. doi:10.1371/journal.pone.0094202.
5. Habeck M. FDA licences imatinib mesylate for CML. Lancet Oncol. 2002;3:6. doi:10.1016/S1470-2045(01)00608-8.
6. Costa DB. More than just an oncogene translocation and a kinase inhibitor: Kevin's story. J Clin Oncol. 2012;30:110–2. doi:10.1200/JCO.2011.39.4486.
7. Nowell PC, Hungerford DA. A minute chromosome in human chronic granulocytic leukemia. Science. 1960;142:1497.
8. Lugo TG, Pendergast AM, Muller AJ, Witte ON. Tyrosine kinase activity and transformation potency of bcr-abl oncogene products. Science. 1990;247:1079–82.
9. Druker BJ, Talpaz M, Resta DJ, et al. Efficacy and safety of a specific inhibitor of the BCR-ABL tyrosine kinase in chronic myeloid leukemia. N Engl J Med. 2001;344:1031–7. doi:10.1056/NEJM200104053441401.
10. Cortes J, Quintás-Cardama A, Kantarjian HM. Monitoring molecular response in chronic myeloid leukemia. Cancer. 2011;117:1113–22. doi:10.1002/cncr.25527.
11. Kwak EL, Bang Y-J, Camidge DR, et al. Anaplastic lymphoma kinase inhibition in non-small-cell lung cancer. N Engl J Med. 2010;363:1693–703. doi:10.1056/NEJMoa1006448.
12. Hatzivassiliou G, Miller I, Takizawa J, et al. IRTA1 and IRTA2, novel immunoglobulin superfamily receptors expressed in B cells and involved in chromosome 1q21 abnormalities in B cell malignancy. Immunity. 2001;14:277–89. doi:10.1016/S1074-7613(01)00109-1.
13. Alcalay M, Meani N, Gelmetti V, et al. Acute myeloid leukemia fusion proteins deregulate genes involved in stem cell maintenance and DNA repair. J Clin Invest. 2003;112:1751–61. doi:10.1172/JCI17595.
14. Chinnaiyan AM, Palanisamy N. Chromosomal aberrations in solid tumors. Prog Mol Biol Transl Sci. 2010;95:55–94. doi:10.1016/B978-0-12-385071-3.00004-6.
15. Yi ES, Chung J-H, Kulig K, Kerr KM. Detection of anaplastic lymphoma kinase (ALK) gene rearrangement in non-small cell lung cancer and related issues in ALK inhibitor therapy: a literature review. Mol Diagn Ther. 2012;16:143–50. doi:10.2165/11632830-000000000-00000.
16. Frampton GM, Fichtenholtz A, Otto GA, et al. Development and validation of a clinical cancer genomic profiling test based on massively parallel DNA sequencing. Nat Biotechnol. 2013;31:1023–31. doi:10.1038/nbt.2696.
17. Newman AM, Bratman SV, To J, et al. An ultrasensitive method for quantitating circulating tumor DNA with broad patient coverage. Nat Med. 2014. doi:10.1038/nm.3519.
18. Iyer MK, Chinnaiyan AM, Maher CA. ChimeraScan: a tool for identifying chimeric transcription in sequencing data. Bioinformatics. 2011;27:2903–4. doi:10.1093/bioinformatics/btr467.

19. Asmann YW, Hossain A, Necela BM, et al. A novel bioinformatics pipeline for identification and characterization of fusion transcripts in breast cancer and normal cell lines. Nucleic Acids Res. 2011;39:e100. doi:10.1093/nar/gkr362.
20. Kim D, Salzberg SL. TopHat-Fusion: an algorithm for discovery of novel fusion transcripts. Genome Biol. 2011;12:R72. doi:10.1186/gb-2011-12-8-r72.
21. Ge H, Liu K, Juan T, et al. FusionMap: detecting fusion genes from next-generation sequencing data at base-pair resolution. Bioinformatics. 2011;27(14):1922–8.
22. Sboner A, Habegger L, Pflueger D, et al. FusionSeq: a modular framework for finding gene fusions by analyzing paired-end RNA-sequencing data. Genome Biol. 2010;11:R104. doi:10.1186/gb-2010-11-10-r104.
23. Chen K, Navin NE, Wang Y, et al. BreakTrans: uncovering the genomic architecture of gene fusions. Genome Biol. 2013;14:R87. doi:10.1186/gb-2013-14-8-r87.
24. Network TCGAR. Comprehensive genomic characterization of squamous cell lung cancers. Nature. 2012;489:519–25. doi:10.1038/nature11404.
25. Network TCGAR. Integrated genomic characterization of endometrial carcinoma. Nature. 2013;497:67–73. doi:10.1038/nature12113.
26. The Cancer Genome Atlas Research Network. Comprehensive molecular profiling of lung adenocarcinoma. Nature. 2014;511:543–50. doi:10.1038/nature13385.
27. The Cancer Genome Atlas Research Network. Comprehensive molecular characterization of gastric adenocarcinoma. Nature. 2014;513:202–9. doi:10.1038/nature13480.
28. The Cancer Genome Atlas Research Network. Comprehensive molecular characterization of urothelial bladder carcinoma. Nature. 2014;507:315–22. doi:10.1038/nature12965.
29. Maher CA, Palanisamy N, Brenner JC, et al. Chimeric transcript discovery by paired-end transcriptome sequencing. Proc Natl Acad Sci. 2009;106:12353–8. doi:10.1073/pnas.0904720106.
30. Edgren H, Murumagi A, Kangaspeska S, et al. Identification of fusion genes in breast cancer by paired-end RNA-sequencing. Genome Biol. 2011;12:R6. doi:10.1186/gb-2011-12-1-r6.
31. Trapnell C, Pachter L, Salzberg SL. TopHat: discovering splice junctions with RNA-Seq. Bioinformatics. 2009;25:1105–11. doi:10.1093/bioinformatics/btp120.
32. Yost SE, Smith EN, Schwab RB, et al. Identification of high-confidence somatic mutations in whole genome sequence of formalin-fixed breast cancer specimens. Nucleic Acids Res. 2012;40:e107. doi:10.1093/nar/gks299.
33. Wu TD, Nacu S. Fast and SNP-tolerant detection of complex variants and splicing in short reads. Bioinformatics. 2010;26:873–81. doi:10.1093/bioinformatics/btq057.
34. Sinicropi D, Qu K, Collin F, et al. Whole transcriptome RNA-Seq analysis of breast cancer recurrence risk using formalin-fixed paraffin-embedded tumor tissue. PLoS One. 2012;7:e40092. doi:10.1371/journal.pone.0040092.
35. Hampton OA, Hollander PD, Miller CA, et al. A sequence-level map of chromosomal breakpoints in the MCF-7 breast cancer cell line yields insights into the evolution of a cancer genome. Genome Res. 2009;19:167–77. doi:10.1101/gr.080259.108.
36. Den Hollander P, Osadchey I, Hampton O, et al. Evolution of genomic diversity in the breast cancer cell line MCF-7. Cancer Res. 2009;69:3171–1. doi:10.1158/0008-5472.SABCS-09-3171.
37. Hampton OA, Miller CA, Koriabine M, et al. Long-range massively parallel mate pair sequencing detects distinct mutations and similar patterns of structural mutability in two breast cancer cell lines. Cancer Genet. 2011;204:447–57. doi:10.1016/j.cancergen.2011.07.009.
38. Sakarya O, Breu H, Radovich M, et al. RNA-Seq mapping and detection of gene fusions with a suffix array algorithm. PLoS Comput Biol. 2012;8:e1002464. doi:10.1371/journal.pcbi.1002464.
39. Cobleigh MA, Tabesh B, Bitterman P, et al. Tumor gene expression and prognosis in breast cancer patients with 10 or more positive lymph nodes. Clin Cancer Res. 2005;11:8623–31. doi:10.1158/1078-0432.CCR-05-0735.
40. Akakura S, Gelman IH. Pivotal role of AKAP12 in the regulation of cellular adhesion dynamics: control of cytoskeletal architecture, cell migration, and mitogenic signaling. J Signal Transduct. 2012;2012:e529179. doi:10.1155/2012/529179.
41. Liu W, Guan M, Hu T, et al. Re-expression of AKAP12 inhibits progression and metastasis potential of colorectal carcinoma in vivo and in vitro. PLoS One. 2011;6:e24015. doi:10.1371/journal.pone.0024015.

42. Robinson MD, McCarthy DJ, Smyth GK. edgeR: a bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics. 2010;26:139–40. doi:10.1093/bioinformatics/btp616.
43. DeSombre ER, Jensen EV. Estrophilin assays in breast cancer: quantitative features and application to the mastectomy specimen. Cancer. 1980;46:2783–8.
44. Wu G, Feng X, Stein L. A human functional protein interaction network and its application to cancer data analysis. Genome Biol. 2010;11:R53. doi:10.1186/gb-2010-11-5-r53.
45. Hanahan D, Weinberg RA. Hallmarks of cancer: the next generation. Cell. 2011;144:646–74. doi:10.1016/j.cell.2011.02.013.

Clinical Applications of Next-Generation Sequencing of Formalin-Fixed Paraffin-Embedded Tumors

Cheryl L. Thompson and Vinay Varadan

Abstract Significant progress has been made in next-generation sequencing (NGS) of tumors from formalin-fixed paraffin-embedded (FFPE) tumors. In this chapter, we review some of the recent developments in RNA and DNA sequencing from FFPE. We highlight some of the current challenges and considerations that must be in place before using NGS in a clinical setting. We begin with a discussion of the technical challenges in dealing with FFPE tumors, including nucleic acid degradation and the potential utility of adding in dissection methods to the tumors. We then discuss bioinformatical and statistical considerations that must be employed when analyzing data obtained from FFPE tissues. We end with a brief discussion of ethical implications and other issues that will need to be addressed before translating discoveries into the clinic.

1 Introduction

The use of formalin-fixed and paraffin-embedded (FFPE) tissue samples for next-generation sequencing (NGS), including both DNA sequencing (DNA-Seq) and RNA sequencing (RNA-Seq), has recently received increasing attention as a result of improved techniques for extracting DNA and RNA from these samples, which are often widely available to researchers and much less costly to obtain and store. Further, many of these samples not only carry extensive clinical data, including longer-term follow-up data, but also have well-preserved tissue architecture that is

C.L. Thompson, Ph.D. (✉)
Department of Family Medicine and Community Health, Case Comprehensive Cancer Center, Case Western Reserve University, Cleveland, OH, USA

Department of Epidemiology and Biostatistics, Case Comprehensive Cancer Center, Case Western Reserve University, Cleveland, OH, USA
e-mail: Cheryl.L.Thompson@case.edu

V. Varadan, Ph.D.
Department of General Medical Sciences (Oncology), Case Comprehensive Cancer Center, Case Western Reserve University, Cleveland, OH, USA
e-mail: vxv89@case.edu

W. Wu, H. Choudhry (eds.), *Next Generation Sequencing in Cancer Research, Volume 2*, DOI 10.1007/978-3-319-15811-2_11

amenable to dissection using techniques such as laser capture microdissection (LCM) [1, 2], allowing scientists to extract regions of interest and minimize noise due to unwanted tissue.

Further, sequencing allows for more comprehensive analysis of tumors. Traditionally, for example, chemotherapy targets have been tested only using a "candidate" approach. For example, trastuzumab has been shown to be an effective agent against breast tumors which have amplifications in the HER2 gene, and thus, its expression is typically measured in breast tumors. However, using sequencing allows not only for characterization of aberrations in HER2, but for every gene in the genome, and, unlike microarrays, can also identify novel mutations that are not captured in the microarray but may have significant clinical implications. This comprehensive characterization also leads to the ability to identify aberrant pathways, which, as a whole, may be important even if the individual mutations are not significant themselves [3].

2 Use of Dissection Methods for Isolating Tumor and Normal Tissue from FFPE Samples

It is often helpful to take advantage of the tissue architecture available in FFPE tumors to isolate areas of interest. For research purposes, it may be of interest to isolate only areas of invasive tissue or areas that have yet to become invasive, for example. Further, by isolating specific areas of the FFPE tissue, we are also able to remove areas that do not contain tumor tissue and thus reduce noise due to contamination from normal tissue. For example, in our recent study of gene expression aberrations associated with cancer initiation and invasion, we used laser capture microdissection (LCM) to successfully isolate areas of normal, preinvasive adenocarcinoma in situ, and invasive components of non-small cell lung cancer (NSCLC) tumors [4]. This allowed us to study gene expression in a cancer progression model, measuring how gene expression changes from normal lung to adenocarcinoma in situ to invasive carcinoma within the same patients.

One of the major disadvantages of LCM is that it requires an expensive piece of equipment that is not available to many researchers. Other techniques for isolation of areas of interests include macrodissection, typically just removing areas that are not useful, such as surrounding normal tissue, using a scalpel or similar instrument. These techniques may be useful for many studies, but are not as accurate as LCM, and their utility wanes as the area of interest becomes smaller.

Importantly, regardless of dissection technique, it is necessary to include a well-experienced pathologist who can assist with correctly identifying these areas of interest on the FFPE specimens. When LCM is utilized, including someone well trained on the equipment and areas of isolation is important to a rigorous study.

3 Developments and Challenges in Extraction of Nucleic Acids from FFPE

The use of FFPE samples for sequencing analyses requires overcoming several challenges. Formalin fixation results in nucleic acid degradation and fragmentation, as well as modification to nucleotides, all of which decrease the yield of good quality RNA or DNA extracted from such samples [5, 6]. These issues can be exacerbated in studies using small tumors where only small amounts of tissue may be available and every fragment counts. In a nice review and comparison of DNA extraction techniques from FFPE, Heydt et al. [7] demonstrated the variability in results using different methods. This highlights the needs to test the results of the DNA preparation for downstream analyses prior to those analyses. However, they also conclude that all methods are good for mutation identification.

RNA can be particularly sensitive to degradation in FFPE. However, as a result of numerous recent technological advances in addressing these challenges, many products are available from a number of companies that can extract good quality RNA from FFPE for downstream sequencing. In fact, several new studies utilizing real-time PCR (RT-PCR) comparing fresh frozen samples to FFPE samples as old as 40 years have demonstrated the utility of RNA extracted from FFPE samples for mRNA and miRNA expression [8–12].

3.1 DNA Sequencing from FFPE

The sequencing of DNA from FFPE has come a long way in recent years. Indeed, a number of groups have successfully performed DNA sequencing from FFPE tumors. A recent comprehensive study comparing FFPE sequencing to sequencing from fresh frozen specimens with newer technology showed very comparable sequencing results [13]. In addition, a large-scale study using a panel of variants (including mutations, insertions-deletions, and fusions) found “actionable” variants, that is, variants with a potential drug target, in a vast majority of the cancer patients [14] using FFPE samples. This highlights the potential clinical utility of sequencing in FFPE.

3.2 RNA Sequencing from FFPE

RNA-Seq from FFPE is an emerging area. RNA-Seq is highly desired as the gold standard for measure of RNA expression but also is not limited by arrays to known transcripts. RNA-Seq can also measure alternative splicing, identify gene fusions, and identify novel transcripts. It can also measure non-coding gene expression. Non-coding genes may represent a new avenue for treatment as our understanding

of this traditionally ignored part of our genome becomes better known. The role of non-coding genes in cancer is becoming established; thus, these genes should not be ignored [15].

Traditionally, RNA-Seq in FFPE has been challenged by fragmented RNA. However, recent studies have utilized new technological developments and have been successful at retrieving good quality RNA-Seq data from FFPE [13, 16], with very good correlations to fresh frozen samples [13]. In our recent study, we were able to successfully perform RNA sequencing on FFPE samples with excellent reproducibility [4]. In this study, we performed whole transcriptome RNA-Seq in a total of 18 samples from six patients, representing normal, preinvasive, and invasive lung cancer specimens from the same tumor. RT-PCR confirmed findings of the top upregulated mRNAs and lincRNAs, with excellent concordance.

Further, newer technologies are emerging with strong results. These new technologies hold promise for even higher-quality sequencing and/or ability to measure additional transcripts. Ribo-Zero-Seq eliminates the need for removal of non-polyadenylated RNA transcripts, which is typically done to remove all ribosomal RNAs. However, traditionally this process also removes all non-ribosomal RNAs that are not polyadenylated. Ribo-Zero-Seq was developed to overcome this and has been shown to have similar ribosomal RNA removal rates as well as coverage and sequencing efficiency from FFPE-derived samples [17]. Norton et al. had similar success for RNA-Seq from FFPE using the Ribo-Zero Gold ScriptSeq V2 library preparation [18].

4 Bioinformatic Considerations When Using Sequencing Data Derived from FFPE Specimens

The potential effect of nucleic acid degradation in FFPE tissues necessitates the adoption of more stringent quality control and statistical criteria in the bioinformatics analyses of sequencing data. In this section, we outline some of the considerations for both DNA and RNA sequencing data analyses that are typically performed on sequencing data and highlight the variations specific to the context of FFPE tissues.

4.1 DNA Sequencing

The principal considerations for the analyses of DNA sequencing data are summarized in Fig. 1.

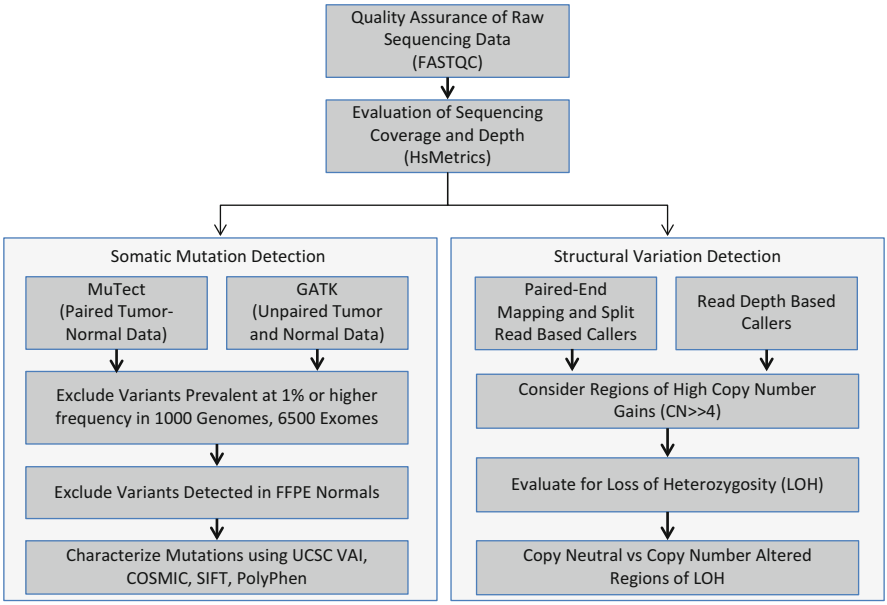


Fig. 1 Bioinformatics considerations for DNA-Seq data processing. This figure shows the suggested pipeline for bioinformatics analysis of DNA-Seq data

4.1.1 Quality Assurance of Raw Sequencing Data

Due to the inherent variability in sequence quality for FFPE-derived sequencing data, especially the likelihood of over-fragmentation of the sequenced DNA, comprehensive evaluation of read quality is necessary before further processing is performed on the data. A popular tool for sequencing data quality evaluation is FastQC (http://www.bioinformatics.bbsrc.ac.uk/projects/fastqc/). FastQC allows evaluation of the distribution of per base quality scores across length of the reads, which helps evaluate the rate of degradation of sequence quality over long runs due to sequencing chemistry. Poor base quality along the length of the read and issues of adapter read-through need to be addressed using read-trimming. Additionally, FastQC also allows evaluation of the sequence diversity of the run and quantifies overrepresented sequences that may arise due to contamination and high levels of sequence duplication that may indicate enrichment biases such as PCR over-amplification. Sequencing data that passes these quality control and assurance steps can then be aligned using a standard aligner [19].

4.1.2 Evaluation of Sequencing Coverage and Depth

Subsequently, the sequencing depth and coverage of the sequenced genomic regions can be derived from the aligned reads using the HsMetrics component of the *picard* package (http://picard.sourceforge.net/). In particular, the capture efficiency of the

exome or targeted sequencing panel can be evaluated using metrics such as the percentage of in-target and off-target reads that pass alignment quality filters, the median read depth and the percentage of bases covered by at least *N* reads. It is preferable that FFPE tumor samples are sequenced at a median depth of at least 100×, with ≥80 % of targets with at least 20× coverage. These considerations will enable more robust mutation calling given the low amounts of input DNA typically available from FFPE tissues as well as due to the potential of normal tissue admixture and intra-tumor heterogeneity. FFPE samples that pass these quality metrics can then be processed for genomic alteration detection.

4.1.3 Somatic Mutation Detection

Given that the identification of somatic mutations in tumor samples is the primary goal of most DNA sequencing endeavors, it is particularly important to consider and account for low allelic fraction nucleotide transition artifacts that are known to result from the FFPE fixation process [20]. These artifacts, when present in FFPE-derived matched normal tissue, will lead to false-negative calls and conversely lead to false-positive calls when present in the tumor tissue. It is therefore useful to use both paired tumor-normal-based somatic mutation callers such as MuTect [21] and unpaired callers such as the Unified Genotyper module in GATK. The union of the potentially somatic mutation calls from the independent detection approaches can then be further filtered using the 1,000 genomes and 6,500 exome databases to eliminate potential germ line variations. In addition, putative somatic mutations that are also found in FFPE-derived normal samples may be filtered as potential sequencing artifacts. The resulting high-confidence somatic mutations can then be annotated for their potential functional effects using the UCSC Variant Annotation Integrator (http://genome.ucsc.edu/cgi-bin/hgVai). In addition, querying these mutations within the COSMIC [22] and TARGET [23] databases can help delineate the clinical relevance of these mutations.

4.1.4 Copy Number Alteration Detection

Although the detection of somatic mutations using next-generation sequencing is increasingly becoming common, the simultaneous detection of somatic copy number alterations in FFPE samples is particularly challenging due to significant coverage and read-depth variations across the genome. Copy number estimation algorithms using next-generation sequencing data can be generally classified into three main approaches: read-depth-based CN estimation [24], combination of discordant read pairs and split reads to identify large structural variants [25], and assembly-based approaches that detect copy number alterations by mapping contigs to the reference genome [26, 27]. However, a recent comprehensive survey of copy number estimation algorithms from next-generation sequencing data reveals that there is a dearth of algorithms that can deal with sequencing data derived from

FFPE tissues [28]. This is primarily because significant sequencing depth variations occur in FFPE samples, and we have recently shown that these effects occur even in FFPE normal samples (manuscript under preparation). While it is possible that genes targeted by large copy number alterations may be detectable [23], one needs to exercise caution while identifying sCNAs using FFPE whole-genome or whole-exome sequencing.

4.2 RNA Sequencing

While significant precautions need to be taken in analyzing DNA sequencing data derived from FFPE tissues, RNA sequencing data poses even more challenges, therefore requiring additional steps as summarized in Fig. 2.

4.2.1 Quality Assurance of Raw Sequencing Data and Transcriptome Coverage

Evaluation of read quality can be performed on the raw sequencing just as outlined for DNA sequencing from FFPE tissues. Here too, the presence of overrepresented sequences could help flag deficiencies in sample or library quality and base quality distribution along the length of the read can identify the need for read trimming. The ENCODE project recommends at least 30 M paired-end reads per sample in order to evaluate transcriptional changes across samples. However, a minimum of

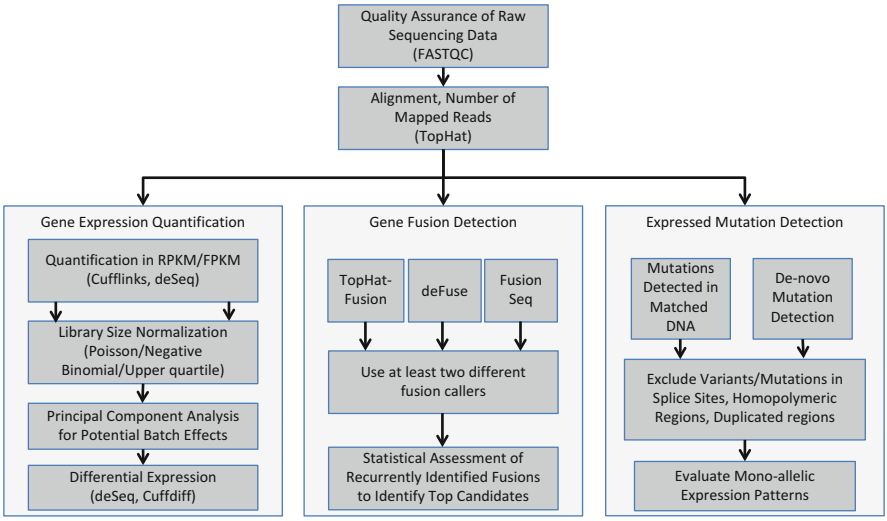


Fig. 2 Bioinformatics considerations for RNA-Seq data processing. An example pipeline for the data processing and analysis of RNA-Seq data

100–200 M paired-end reads are recommended for more in-depth characterization of alterative splicing, gene fusion detection, or evaluation of expressed mutations. Transcriptome-specific alignment tools such as TopHat [29] or Trinity [30] can be used to estimate known transcript abundance as well as identify novel transcripts due to their de novo transcript assembly capabilities. After alignment, additional quality assurance checks [31] such as duplication rates, GC bias, 3′/5′ bias, ribosomal RNA contamination, and coverage statistics for exons, introns, and intragenic regions are particularly important as the fragmentation inherent in FFPE samples could significantly bias the transcriptomic readout.

4.2.2 Transcript Assembly and Quantification

Transcript-level expression quantification after alignment of the RNA-Seq data is typically performed using Cufflinks/Cuffdiff [32] or DeSeq [33]. Since FFPE samples can suffer from excessive degradation of RNA that results in the attenuation of transcript abundance estimates, principal component analysis on the expression data can help determine if any samples in the cohort are particularly impacted, thus necessitating either exclusion or independent analysis of these samples. The potential for FFPE-induced degradation also impacts the estimation of differential transcript expression, and therefore, more stringent FDR and fold change thresholds need to be considered while analyzing such datasets.

4.2.3 Gene Fusion Detection

Multiple computational approaches have been proposed in the literature to detect gene fusions using RNA-Seq data: TopHat-Fusion [34] generates an initial set of alignments using Bowtie [19] and focuses on the discordant read pairs to derive candidate gene fusions that are then filtered using multiple strategies to account for potential false positives. In a different strategy focusing on exonic regions, ShortFuse [35] applies a graphical model using reads mapping to exons of each potential fusion gene pair to determine the most likely fusion junction. FusionSeq [36] uses strict filtering strategy to eliminate false positives and therefore falls within the same category as TopHat-Fusion. One of the most comprehensive fusion-calling algorithms developed is called deFuse [37] that first uses read mate pairs that are discordantly aligned to the genome to detect potential fusion candidates. Subsequently, deFuse comprehensively characterizes the sequences around each candidate fusion junction using BLAT [38] and builds a classifier using these sequence summaries to derive the probability that a candidate fusion call is likely to be true. SplitSeek [39] is a spliced aligner that uses ends of reads to identify fusion events, whereas Trans-ABySS [40] and ABySS [41] use de novo assembly to generate full-length transcripts and then use the assembled transcripts to detect fusion events. Each of these algorithms suffers from significant false-positive rates and is not easily translatable across studies due to the multiple parameters that need to be specified by the user.

Therefore, it is essential that one employs at least two qualitatively different algorithms to reliably detect a fusion in a given sample. The final list of candidate fusions can be determined by filtering based on whether an open reading frame is maintained in the fused transcript and if there is substantial coverage of the fusion junction. Furthermore, potential filtering strategies may include elimination of fusions between a gene and its pseudogene and fusions among paralogous genes as potential sequencing alignment artifacts. Such filtering strategies may be even more stringently applied in the context of FFPE tissue samples due to the higher potential of sequencing errors and artifacts. In summary, gene fusion detection using RNA-Seq data derived from FFPE tissues presents a significant challenge, and significant improvements in specificity of fusion-calling algorithms are needed before this technology can be widely used to screen for fusions.

4.2.4 Detection of Expressed Mutations

While mutations can be relatively easily detected using DNA sequencing as detailed above, identification of mutations that are expressed can provide additional insights into their functional role in cancers. Specifically, potentially functional oncogenic mutations, such as missense mutations in tumor suppressors leading to high expression (e.g., TP53 mutations), activating mutations in oncogenes (e.g., PIK3CA) or mutations in therapeutically targeted genes (ESR1), can be detected using RNA-Seq data. However, not all expressed mutations may be detectable. This is often due to nonsense-mediated decay of the transcript, thus resulting in lack of sufficient coverage for detection. Ideally, mutations in a particular sample would first be identified using DNA sequencing, and only the confirmed somatic mutations would be evaluated using RNA-Seq. De novo detection of expressed variants and mutations is also possible but requires stringent filtering due to frequent false-positive calls close to splice sites, homopolymeric regions, or duplicated regions [42]. Finally, the combination of expressed mutations with patterns of mono-allelic expression [43] would allow for the detection of potentially functional mutated alleles given that the wild-type allele has been silenced.

5 Translating Discoveries into the Clinic

Ultimately the goal of medical research is to lead to improvements in clinical practice. Indeed, FFPE DNA sequencing is already in trials in some clinics, either internally or through external services, such as Foundation One [14]. Sequencing presents a unique opportunity to unveil a wealth of extremely important insight into tumor biology that has previously been unheard of. These insights are likely to drive physicians toward the most efficacious treatments for that particular patient. However, using sequencing in the clinic is not without challenges as well.

One important challenge is physician understanding of the technology. Most physicians were not trained in interpretation of sequencing studies in medical school and may only have a basic understanding of the technology, including limitations. Indeed, when faced with a sequencing report, many physicians would find them very challenging to interpret. Some sequencing centers have opted to mitigate this issue by simplifying the report into "actionable" findings. However, this is not without loss of information which may be important, particularly in a rapidly evolving field. At our institution, as well as others, we have established multidisciplinary genomic tumor boards (MGTB) to discuss patients whose tumor has been sequenced. Importantly, beyond the traditional ensemble of physicians representing the entire oncology treatment team, these boards include scientists, including bioinformaticians, basic scientists, bioethicists, and clinical researchers whose role is to help bridge this gap.

Physicians are constantly required to practice "evidence-based" medicine. This means that therapies must bring with them a specific trial. In the case of sequencing, often the "actionable" variants that are identified do not have evidence supporting them as a therapeutic. For example, there may be a trial that suggests the efficacy of a therapeutic for treating a given cancer, but if the variant is found in a patient with a different cancer, for which the drug was not specifically approved, it is not necessarily clear if it is appropriate to treat that patient with that drug, and this can be a major struggle for oncologists [44], and thus, these treatments, while potentially the most effective, may be reserved for only those patients that fail other treatments. This issue will need to be addressed in clinical trials that take into account the comprehensive nature of NGS [45]. Some proposed sequencing technologies have limited the sequencing to known variants to alleviate some of this issue [46]. However, of course, this is at the expense of the comprehensive nature of NGS.

5.1 Using FFPE Tumors for Clinical Applications

FFPE samples represent a huge archive of available clinical specimens. Thus, their utility for research is immense. However, the use of FFPE tumors for clinical care of cancer patients could potentially be very high as well. Freezing fresh tumors is the current gold standard. However, this is not routinely done in a vast majority of clinical settings, whereas the protocol for preserving tumor tissue via FFPE is well established and routinely done by most pathologists. Further, FFPE is easier, cheaper, and more feasible at most institutions. This is an especially important consideration for resource-poor or remote areas, which are extremely important not to exclude from important medical advances. Thus, being able to use FFPE samples for clinical tests and decision making may make the tests more widely utilized and have a broader impact.

With any research that ultimately has the goal of being translated into a clinical application, it is important to replicate how the test would be used in a clinical setting. Thus, a rigorous study with the correct patient population, defined collection and

storage protocol, as well as a clearly defined processing and analysis pipeline, that is also feasible in a clinical setting, will be necessary in order to create a test that can be used to direct patient care.

5.2 Bioethical Concerns for Using NGS in the Clinic or in Clinical Trials

Sequencing of tumors, be it DNA or RNA sequencing, also brings with it quite a number of ethical challenges. Among these challenges are informed consent, return of incidental findings, and the potential to increase disparities.

Informed consent is an important concept in modern medical practice that requires all patients to have an understanding of any proposed medical test or procedure and only afterward provide consent to that test or procedure. Previously, we discussed the importance of consideration of incomplete physician knowledge of NGS. The understanding of basic genetic concepts, even less NGS, by the general public is very low. Thus, explaining NGS of a tumor to a patient and how it would be used for their treatment can be very challenging to physicians. Using persons trained in communication of genetic information to patients, such as genetic counselors, and/or significant training to oncologists may alleviate some of these communication challenges.

Another important consideration is the return of incidental findings. When we sequence a tumor, it is not common for "incidental" findings to arise. These are not part of the intent of the test, but are as a result of this wealth of data on the patient that is returned through NGS. For example, when sequencing a non-breast tumor, one might discover a mutation in a BRCA1 or BRCA2, which predispose to breast or ovarian cancer. Should this be conveyed to the patient if not relevant for the current cancer? What if the patient is a male, but has daughters? Another example is it might be discovered that the patient is a carrier for a mutation causing Tay-Sachs disease. If the patient is still in childbearing years, is it important to tell the patient about this mutation? What about a variant that causes a relatively small increase risk of developing a disease? Does it matter if the disease is potentially preventable (such as heart disease) or not (such as Alzheimer's)? Further, it is unclear how best to convey these findings to the patient. There may be differences, as well, in the context of a research study vs. clinical care [47]. Further, if a patient is a pediatric cancer patient, these may have entirely different meanings and consequences. These questions remain to be answered, and there is no consensus among professionals about what is the best policy [48].

Through massive advances in sequencing technologies, the cost of exome sequencing and RNA-Seq has reduced dramatically over the past couple decades. However, even today, NGS is not inexpensive and is not often covered by insurance. In addition, we previously discussed the advantages of having an MGTB to interpret findings from NGS. However, while MGTBs are relatively easy to develop at large

university-affiliated comprehensive cancer centers, MGTBs are not likely to be feasible for most patients, particularly those in more rural areas or resource-poor areas of the world. This limits the ability to bring this important new technology to these patients. As physicians and scientists, we need to find ways to reduce disparities and not to advance science in a way that increases economic disparities.

6 Conclusions

In conclusion, NGS from FFPE is currently feasible for research, and emerging technologies are improving the performance of the nucleic acid isolation as well as the sequencing itself. The large repositories of FFPE tumors could prove to be an extremely valuable resource to researchers investigating cancer genomics. However, there are still a number of considerations that must be made when doing NGS from FFPE. When completing the bioinformatical analyses, some precautions must be taken to ensure quality analysis, particularly for RNA-Seq, which is more prone to errors from nucleic acid degradation compared to DNA-Seq. Further, the goal of all research is to advance the clinical care of patients. However, before translating findings from sequencing from FFPE into the clinic, a number of validation studies as well as bioethical considerations must be made.

References

1. Burgemeister R. Nucleic acids extraction from laser microdissected FFPE tissue sections. Methods Mol Biol. 2011;724:117–29.
2. Joseph A, Gnanapragasam VJ. Laser-capture microdissection and transcriptional profiling in archival FFPE tissue in prostate cancer. Methods Mol Biol. 2011;755:291–300.
3. Park JY, Kricka LJ, Fortina P. Next-generation sequencing in the clinic. Nat Biotechnol. 2013;31(11):990–2.
4. Morton ML, Bai X, Merry CR, Linden PA, Khalil AM, Leidner RS, Thompson CL. Identification of mRNAs and lincRNAs associated with lung cancer progression using next-generation RNA sequencing from laser micro-dissected archival FFPE tissue specimens. Lung Cancer. 2014;85(1):31–9.
5. Farragher SM, Tanney A, Kennedy RD, Paul Harkin D. RNA expression analysis from formalin fixed paraffin embedded tissues. Histochem Cell Biol. 2008;130(3):435–45.
6. Gnanapragasam VJ. Unlocking the molecular archive: the emerging use of formalin-fixed paraffin-embedded tissue for biomarker research in urological cancer. BJU Int. 2010;105(2):274–8.
7. Heydt C, Fassunke J, Kunstlinger H, Ihle MA, Konig K, Heukamp LC, Schildhaus HU, Odenthal M, Buttner R, Merkelbach-Bruse S. Comparison of pre-analytical FFPE sample preparation methods and their impact on massively parallel sequencing in routine diagnostics. PLoS One. 2014;9(8):e104566.
8. Liu A, Xu X. MicroRNA isolation from formalin-fixed, paraffin-embedded tissues. Methods Mol Biol. 2011;724:259–67.
9. Lu X, van der Straaten T, Tiller M, Li X. Evidence for qualified quantitative mRNA analysis in formalin-fixed and paraffin-embedded colorectal carcinoma cells and tissue. J Clin Lab Anal. 2011;25(3):166–73.

10. Ludyga N, Grunwald B, Azimzadeh O, Englert S, Hofler H, Tapio S, Aubele M. Nucleic acids from long-term preserved FFPE tissues are suitable for downstream analyses. Virchows Arch. 2012;460(2):131–40.
11. Stewart GD, Baird J, Rae F, Nanda J, Riddick AC, Harrison DJ. Utilizing mRNA extracted from small, archival formalin-fixed paraffin-embedded prostate samples for translational research: assessment of the effect of increasing sample age and storage temperature. Int Urol Nephrol. 2011;43(4):961–7.
12. Waldron L, Simpson P, Parmigiani G, Huttenhower C. Report on emerging technologies for translational bioinformatics: a symposium on gene expression profiling for archival tissues. BMC Cancer. 2012;12:124.
13. Hedegaard J, Thorsen K, Lund MK, Hein AM, Hamilton-Dutoit SJ, Vang S, Nordentoft I, Birkenkamp-Demtroder K, Kruhoffer M, Hager H, et al. Next-generation sequencing of RNA and DNA isolated from paired fresh-frozen and formalin-fixed paraffin-embedded samples of human cancer and normal tissue. PLoS One. 2014;9(5):e98187.
14. Frampton GM, Fichtenholtz A, Otto GA, Wang K, Downing SR, He J, Schnall-Levin M, White J, Sanford EM, An P, et al. Development and validation of a clinical cancer genomic profiling test based on massively parallel DNA sequencing. Nat Biotechnol. 2013;31(11): 1023–31.
15. Xue B, He L. An expanding universe of the non-coding genome in cancer biology. Carcinogenesis. 2014;35(6):1209–16.
16. Guo X, Zhu SX, Brunner AL, van de Rijn M, West RB. Next generation sequencing-based expression profiling identifies signatures from benign stromal proliferations that define stromal components of breast cancer. Breast Cancer Res. 2013;15(6):R117.
17. Zhao W, He X, Hoadley KA, Parker JS, Hayes DN, Perou CM. Comparison of RNA-Seq by poly (A) capture, ribosomal RNA depletion, and DNA microarray for expression profiling. BMC Genomics. 2014;15(1):419.
18. Norton N, Sun Z, Asmann YW, Serie DJ, Necela BM, Bhagwate A, Jen J, Eckloff BW, Kalari KR, Thompson KJ, et al. Gene expression, single nucleotide variant and fusion transcript discovery in archival material from breast tumors. PLoS One. 2013;8(11):e81925.
19. Langmead B, Trapnell C, Pop M, Salzberg SL. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biol. 2009;10(3):R25.
20. Williams C, Ponten F, Moberg C, Soderkvist P, Uhlen M, Ponten J, Sitbon G, Lundeberg J. A high frequency of sequence alterations is due to formalin fixation of archival specimens. Am J Pathol. 1999;155(5):1467–71.
21. Cibulskis K, Lawrence MS, Carter SL, Sivachenko A, Jaffe D, Sougnez C, Gabriel S, Meyerson M, Lander ES, Getz G. Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples. Nat Biotechnol. 2013;31(3):213–9.
22. Bamford S, Dawson E, Forbes S, Clements J, Pettett R, Dogan A, Flanagan A, Teague J, Futreal PA, Stratton MR, et al. The COSMIC (Catalogue of Somatic Mutations in Cancer) database and website. Br J Cancer. 2004;91(2):355–8.
23. Van Allen EM, Wagle N, Stojanov P, Perrin DL, Cibulskis K, Marlow S, Jane-Valbuena J, Friedrich DC, Kryukov G, Carter SL, et al. Whole-exome sequencing and clinical interpretation of formalin-fixed, paraffin-embedded tumor samples to guide precision cancer medicine. Nat Med. 2014;20(6):682–8.
24. Boeva V, Popova T, Bleakley K, Chiche P, Cappo J, Schleiermacher G, Janoueix-Lerosey I, Delattre O, Barillot E. Control-FREEC: a tool for assessing copy number and allelic content using next-generation sequencing data. Bioinformatics. 2012;28(3):423–5.
25. Chen K, Wallis JW, McLellan MD, Larson DE, Kalicki JM, Pohl CS, McGrath SD, Wendl MC, Zhang Q, Locke DP, et al. BreakDancer: an algorithm for high-resolution mapping of genomic structural variation. Nat Methods. 2009;6(9):677–81.
26. Nijkamp JF, van den Broek MA, Geertman JM, Reinders MJ, Daran JM, de Ridder D. De novo detection of copy number variation by co-assembly. Bioinformatics. 2012;28(24):3195–202.
27. Iqbal Z, Caccamo M, Turner I, Flicek P, McVean G. De novo assembly and genotyping of variants using colored de Bruijn graphs. Nat Genet. 2012;44(2):226–32.

28. Zhao M, Wang Q, Wang Q, Jia P, Zhao Z. Computational tools for copy number variation (CNV) detection using next-generation sequencing data: features and perspectives. BMC Bioinformatics. 2013;14 Suppl 11:S1.
29. Trapnell C, Pachter L, Salzberg SL. TopHat: discovering splice junctions with RNA-Seq. Bioinformatics. 2009;25(9):1105–11.
30. Grabherr MG, Haas BJ, Yassour M, Levin JZ, Thompson DA, Amit I, Adiconis X, Fan L, Raychowdhury R, Zeng Q, et al. Full-length transcriptome assembly from RNA-Seq data without a reference genome. Nat Biotechnol. 2011;29(7):644–52.
31. DeLuca DS, Levin JZ, Sivachenko A, Fennell T, Nazaire MD, Williams C, Reich M, Winckler W, Getz G. RNA-SeQC: RNA-seq metrics for quality control and process optimization. Bioinformatics. 2012;28(11):1530–2.
32. Trapnell C, Roberts A, Goff L, Pertea G, Kim D, Kelley DR, Pimentel H, Salzberg SL, Rinn JL, Pachter L. Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. Nat Protoc. 2012;7(3):562–78.
33. Anders S, Huber W. Differential expression analysis for sequence count data. Genome Biol. 2010;11(10):R106.
34. Kim D, Salzberg SL. TopHat-Fusion: an algorithm for discovery of novel fusion transcripts. Genome Biol. 2011;12(8):R72.
35. Kinsella M, Harismendy O, Nakano M, Frazer KA, Bafna V. Sensitive gene fusion detection using ambiguously mapping RNA-Seq read pairs. Bioinformatics. 2011;27(8):1068–75.
36. Sboner A, Habegger L, Pflueger D, Terry S, Chen DZ, Rozowsky JS, Tewari AK, Kitabayashi N, Moss BJ, Chee MS, et al. FusionSeq: a modular framework for finding gene fusions by analyzing paired-end RNA-sequencing data. Genome Biol. 2010;11(10):R104.
37. McPherson A, Hormozdiari F, Zayed A, Giuliany R, Ha G, Sun MG, Griffith M, Heravi Moussavi A, Senz J, Melnyk N, et al. deFuse: an algorithm for gene fusion discovery in tumor RNA-Seq data. PLoS Comput Biol. 2011;7(5):e1001138.
38. Kent WJ. BLAT–the BLAST-like alignment tool. Genome Res. 2002;12(4):656–64.
39. Ameur A, Wetterbom A, Feuk L, Gyllensten U. Global and unbiased detection of splice junctions from RNA-seq data. Genome Biol. 2010;11(3):R34.
40. Robertson G, Schein J, Chiu R, Corbett R, Field M, Jackman SD, Mungall K, Lee S, Okada HM, Qian JQ, et al. De novo assembly and analysis of RNA-seq data. Nat Methods. 2010;7(11):909–12.
41. Simpson JT, Wong K, Jackman SD, Schein JE, Jones SJ, Birol I. ABySS: a parallel assembler for short read sequence data. Genome Res. 2009;19(6):1117–23.
42. Piskol R, Ramaswami G, Li JB. Reliable identification of genomic variants from RNA-seq data. Am J Hum Genet. 2013;93(4):641–51.
43. Mayba O, Gilbert HN, Liu J, Haverty PM, Suchit J, Jiang Z, Watanabe Y, Zhang Z. MBASED: allele-specific expression detection in cancer tissues and cell lines. Genome Biol. 2014;15(8):405.
44. Garber K. Ready or not: personal tumor profiling tests take off. J Natl Cancer Inst. 2011;103(2):84–6.
45. Simon R, Roychowdhury S. Implementing personalized cancer genomics in clinical trials. Nat Rev Drug Discov. 2013;12(5):358–69.
46. Bourgon R, Lu S, Yan Y, Lackner MR, Wang W, Weigman V, Wang D, Guan Y, Ryner L, Koeppen H, et al. High-throughput detection of clinically relevant mutations in archived tumor samples by multiplexed PCR and next-generation sequencing. Clin Cancer Res. 2014;20(8):2080–91.
47. Jarvik GP, Amendola LM, Berg JS, Brothers K, Clayton EW, Chung W, Evans BJ, Evans JP, Fullerton SM, Gallego CJ, et al. Return of genomic results to research participants: the floor, the ceiling, and the choices in between. Am J Hum Genet. 2014;94(6):818–26.
48. Yu JH, Harrell TM, Jamal SM, Tabor HK, Bamshad MJ. Attitudes of genetics professionals toward the return of incidental results from exome and whole-genome sequencing. Am J Hum Genet. 2014;95(1):77–84.

ChIP-BS-Sequencing in Cancer Epigenomics

Karthikraj Natarajan and Fei Gao

Abstract DNA methylation and histone modifications are crucial epigenetic modifications that involved in transcriptional regulatory network. Due to environmental cues, distortion in epigenomic landscape—in DNA methylation and histone modification—might be considered as a reason for aberrant gene expression in cancer. A confounding puzzle in cancer epigenetics is to decipher whether a significant mechanism between DNA methylation and histone modification triggers tumorigenesis initiation and progression. ChIP-BS-seq is a technique that combines chromatin immunoprecipitation and bisulfite conversion followed by high-throughput sequencing to study genome-wide cross talk between DNA methylation and histone modification. In this chapter, we have explored background, technological advancement in epigenomics research and its future developments. We also have summarized our latest findings on using ChIP-BS-seq in cancer cell lines.

1 Introduction

After the discovery of DNA structure, the Human Genome Project was proposed to identify the role of long-coiled DNA sequences. With higher expectation, the Human Genome Project was successfully completed on 2004. However, researchers understood that the genome project provides only the read sequence of the entire genome, but it does not provide information on gene regulation, protein and mRNA function, and their relation. Subsequently, it led to development of OMICS

K. Natarajan
Science & Technology Department, BGI-Shenzhen, Shenzhen 518083, China
e-mail: genekarthik@gmail.com

F. Gao (✉)
Science & Technology Department, BGI-Shenzhen, Shenzhen 518083, China

Section of Comparative Paediatrics and Nutrition, Department of Veterinary Clinical and Animal Sciences, Faculty of Medical and Health Sciences, University of Copenhagen, Dyrlægevej 16, Frederiksberg C, Copenhagen 1870, Denmark
e-mail: flys828@gmail.com

W. Wu, H. Choudhry (eds.), *Next Generation Sequencing in Cancer Research, Volume 2*, DOI 10.1007/978-3-319-15811-2_12

fields such as epigenomics, transcriptomics, and proteomics, which are believed to solve gene and genome mysteries.

Epigenomics is a booming hot spot, which is growing promisingly for a couple of decade to understand complex mechanism in gene regulatory system. The complex network of DNA and proteins mass tangled inside a capsule-like structure called nucleus. It is difficult to understand comprehensive epigenomic network with available technology. The development of the field is bottlenecked by lack of apt epigenetic tools or inefficient available techniques. To unravel molecular mechanism either in normal biological processes or on understanding complex diseases—cancer, diabetics, and Alzheimer's—new efficient interdisciplinary methods or tools are inevitably needed for the hour.

Epigenetic landscape of a genome is maintained by crucial mechanisms such as DNA methylation and histone modification both structurally and functionally (Fig. 1) [1]. The malfunction of these mechanisms profoundly distorts the normal cellular function, which may lead to development of cancer. Aberrant hypermethylation at CpG island (CGI) is involved in gene silencing in X-inactivation, imprinting genes, aging process [2, 3], and frequent loss of tumor suppressor gene (TSG) function in cancer [4, 5]. Likewise, change in DNA methylation is also associated with aberrant histone modification patterns [6, 7]. Generally, in normal cells, promoter CGI

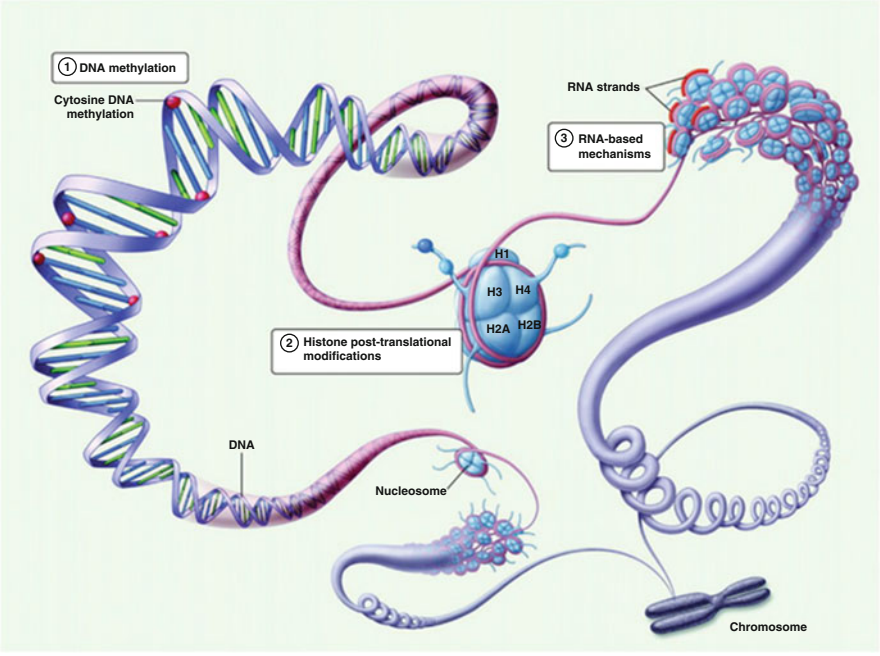


Fig. 1 Three primary epigenetic mechanisms: (1) DNA methylation, (2) histone posttranslational modifications, and (3) RNA-based mechanisms, including miRNAs and large noncoding RNAs (lncRNAs). [Reprinted with permission from Macmillan Publishers Ltd: Laboratory Investigation [81], copyright (2014)]

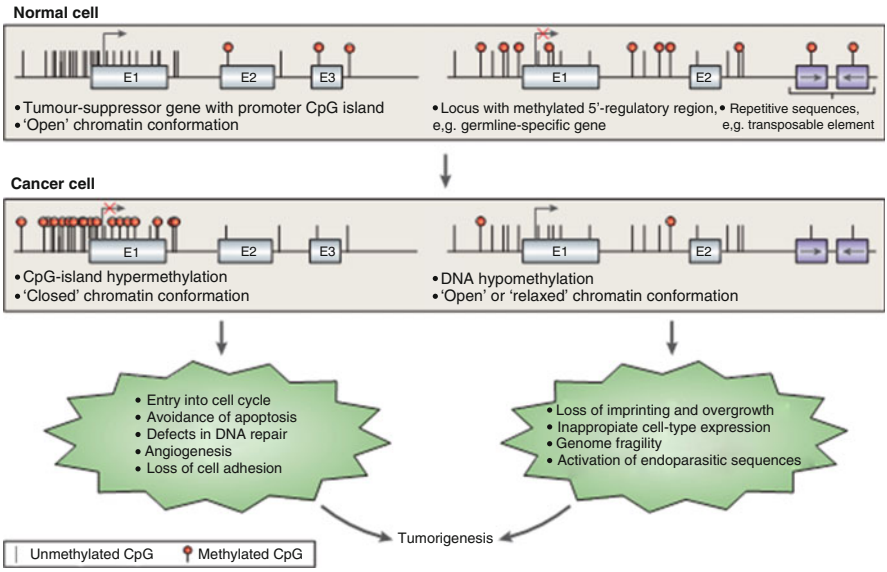


Fig. 2 Aberrant DNA methylation patterns in tumorigenesis. Aberrant hypermethylation at CGI promoter of tumor suppressor genes in cancer cells leads to transcriptional repression of these genes (Table 1), whereas in normal cells, CGI promoter is free from methylation. Contrastingly, the genome undergoes global hypomethylation in cancer cells, which seems to contribute genomic instability. [Reprinted with permission from Macmillan Publishers Ltd: Nature review genetics [9], copyright (2007)]

regions are not methylated; hence, the genes are transcribed without any hindrance. But in the cancer cells, many genes are repressed by aberrant methylation at promoter CGI (Fig. 2) [8, 9]. An intriguing puzzle in cancer epigenetics is to understand the underlying mechanism between DNA methylation and histone modification alterations in tumorigenesis. In other words, what factor triggers the normal epigenomic function onto tumor initiation and progression?

The available techniques are limited to study mechanism between DNA methylation and diverse histone marks on its effect on gene regulation. The newly developed ChIP-BS-seq technique can help to better understanding the study of DNA methylation and histone modification on cancer context [10, 11]. In this chapter, we aim to shed light on the advantage of ChIP-BS-seq in cancer epigenomics research and its future prospects.

2 Chromatin Modifications

DNA is wrapped around histone octamer, namely, H3, H4, H2A, and H2B to form nucleosome—a bead-on-string structure of length 147 bp [12, 13]. The histone proteins consist of N-amino tails that are subject to posttranslational modification

which subsequently leads to gene activation or repression by alteration in chromatin structure. There are about 60 types of histone modifications that have been reported to be modified at histone tails and main class of posttranslational modification (PTMs) such as histone methylation, acetylation, phosphorylation, ubiquitination, sumoylation, and ADP ribosylation [14]. PTMs are carried by recruiting proteins and its complex to perform unique enzymatic activity for gene regulation. Apart from change in chromatin architecture, remodeling enzymes are also involved in nucleosome repositioning by ATP hydrolysis. Because of its diversity and modification at multiple sites, histone modifications show high-level complexity on understanding its epigenetic regulation. Out of all modifications, histone methylation and acetylation are shown to be involved in tumor alterations [9, 15].

2.1 Histone Methylation and Demethylation

Histone methylation predominately modifies at different sites of lysine and arginine amino chain of histone tails. Further, the methyl group at lysine may be mono-, di-, or tri-methylated, whereas arginine may be mono-, symmetrical, asymmetrical, or di-methylated [16]. In case of histone methylation, modification doesn't alter the charge of histones. Histone lysine methyltransferase (HKMT) is an enzyme which methylates lysine by transfer of methyl group at S-adenosylmethionine (SAM) to lysine's ε-amino group. Most of the HKMTs contain SET domain for its enzymatic activity, and SET domain-containing HKMTs are considered to be controlling type of methylation. HKMT was first identified as SUV39H1 that methylates at H3K9 [17]. In case of arginine methylation, the addition of methylation is controlled by arginine methyltransferase by transfer of methyl group from SAM to arginine's ω-guanidino group. A group of arginine methyltransferase is generally referred to as protein arginine methyltransferases (PRMTs). PRMTs catalyze methylation at nitrogen of specific arginine residues, and the arginine methylation has three types: monomethylation and two types of dimethylation [18].

Methylation of lysine residue at histone tail was considered to be an irreversible reaction. Histone lysine demethylase (KDM) is called as eraser enzyme, which removes methyl group from histone tails. LSD1 (lysine-specific demethylase 1) was the first identified KDM protein that demethylates mono-/di-lysine 4 at histone 3 (H3K4me1/2) [19]. A second KDM was identified as LSD2 which also demethylates H3K4me1/2 in mammals [20]. Later, a group of KDM eraser proteins has been identified; they have structurally different JmjC domain than previously identified proteins. LSD1 and LSD2 are related to DNMTs for de novo methylation at DNA imprinting stage, and a notion is that feasibility of connection between histone methylation and DNA methylation might exist [21]. Aberrant expression or mutation in KDM has been reported in many types of cancer, and it is considered to be an implication of tumor development [22, 23]. Thus, the enzyme is not only considered as marker for cancer but also represents as targeting tool for novel anticancer therapeutic targets.

3 DNA Methylation

DNA methylation is a covalent modification, commonly called as cytosine methylation by addition of methyl ($–CH_3$) group at the fifth position of the cytosine nucleotide base [3]. Addition of methyl group on the cytosine commonly occurs within CG dinucleotides (CpG), and in mammals, about 60–90 % of all the CpGs are methylated [24]. DNA methylation is a well-studied epigenetic mechanism, and it is essential for embryonic development, X-chromosome inactivation, genomic imprinting, tissue-specific differentiation, aging, and chromosomal instability. It plays vital roles in gene regulation and decides on-off gene expression which determines the activity of cells. DNA methylation is regulated by sophisticated machineries—DNA methyltransferase (DNMTs), which is responsible for establishing and maintaining methylation patterns, and methyl-CpG binding proteins (MBDs), which "read" DNA methylation marks.

The dynamic change in methylation pattern is considered as onset of cancer and its progression. Many findings have reported that the interplay between DNA methylation and histone modification malfunction related to onset of cancer [9, 25]. Our understanding of epigenetic information and its complexity has been significantly enlightened with the help of high-throughput sequencing. Interestingly, understanding epigenetic regulation during mammalian development and stem cell differentiation and applying the knowledge in parallel to unravel cancer alterations and also substantial evidence show that DNA methylation and histone modifications work together to regulate gene expression [25].

DNA methylation pattern is unique among different types of cells and tissues; alteration in methylation pattern may affect the cell properties. The specificity at DNA methylation could be due to tissue-dependent differentiated methylated regions (T-DMRs) present in gene sequences and other regulatory elements in the genome [26]. T-DMRs are defined as C-DMRs if the patterns of T-DMR in cancer tissues are different from normal tissues [27].

3.1 CpG Island

A CpG island (CGI) is frequently present at gene promoter or exons, and it is unmethylated in normal cells [28]. The methylation of CGI promoters is essential for normal developmental in genomic imprinting, X-chromosome inactivation, and tissue-specific genes. In the mammalian genome, approximately 70 % of CpG dinucleotides are methylated. The methylation of CGI at TSG retinoblastoma (RB) gene in human cancer was first discovered in 1989 [29]. In human cancer, the inactivation of TSG $p16^{INK4a}$ by CGI hypermethylation was the first epigenetic gene silencing mechanism that was shown by Stephen Baylin and Peter A. Jones [30–32]. After this discovery, aberrantly methylated TSG and other genes at a CGI were identified (Table 1) [33]. Previous works have shown that CGIs have been found located closer to the overlapping transcription starting site, and it enunciates the relation

Table 1 A list of genes that are silenced by aberrant hypermethylation at CpG island in human cancer

Gene	Function	Tumor type
AR	Androgen receptor	Prostate
BRCA1	DNA repair, transcription	Breast, ovary
CRBP1	Retinol-binding protein	Colon, stomach, lymphoma
DAPK	Pro-apoptotic	Lymphoma, lung, colon
ER	Estrogen receptor	Breast
GATA4	Transcription factor	Colon, stomach
GATA5	Transcription factor	Colon, stomach
GSTP1	Conjugation to glutathione	Prostate, breast, kidney
HOXA9	Home box protein	Neuroblastoma
ID4	Transcription factor	Leukemia, stomach
IGFBP3	Growth-factor-binding protein	Lung, skin
Lamin A/C	Nuclear intermediate filament	Lymphoma, leukemia
LKB1/STK11	Serine-threonine kinase	Colon, breast, lung
MGMT	DNA repair of 06-alkyl-guanine	Multiple type mutations
MLH1	DNA mismatch repair	Colon, endometrium, stomach
NORE1A	Ras effector homologue	Lung
$p14^{ARF}$	MDM2 inhibitor	Colon, stomach, kidney
$p15^{INK4b}$	Cyclin-dependent kinase inhibitor	Leukemia
$p16^{INK4a}$	Cyclin-dependent kinase inhibitor	Multiple types
p73	p53 homologue	Lymphoma
PR	Progesterone receptor	Breast
PRLR	Prolactin receptor	Breast
RARβ2	Retinoic acid receptor-β2	Colon, lung, head and neck
RASSF1A	Ras effector homologue	Multiple types
Rb	Cell cycle inhibitor	Retinoblastoma
RIZ1	Histone/protein methyltransferase	Breast, liver
SLC5A8	Sodium transporter	Glioma, colon
SOCS1	Inhibitor of JAK-STAT pathway	Liver, myeloma
SOCS3	Inhibitor of JAK-STAT pathway	Lung
SRBC	BRCA1-binding protein	Breast, lung
SYK	Tyrosine kinase	Breast
THBS1	Thrombospondin-1, Anti-angiogenic	Glioma
TMS1	Pro-apoptotic	Breast
TPEF/HPP1	Transmembrane protein	Colon, bladder
TSHR	Thyroid-stimulating hormone receptor	Thyroid
VHL	Ubiquitin ligase component	Kidney, hemangioblastoma
WIF1	Wnt inhibitory factor	Colon, lung
WRN	DNA repair	Colon, stomach, sarcoma

(Adapted from Nature Review Genetics [9] 2007)

between CGIs and transcriptional initiation [34, 35]. The density of CpG overlapped at promoters is varied from different gene promoters. A study has shown that low CpG-density promoters are associated with tissue-specific genes where high

CpG-density promoters are present in housekeeping and developmental genes [36]. During somatic cell reprogramming, the repressive mark H3K27me3 was replaced with activating mark H3K4me2 by Parp1 and Tet2 in pluripotency genes [37, 38]. Most of the studies have showed that DNA methylation alteration in promoters or CpG could be responsible for cancer initiation or progression (Fig. 2). In addition, a group has shown that methylation alterations are not only at promoter or CGI in colon cancer, but also at CGI shore, which is 2 kb distant from CGI [27].

4 Interplay Between DNA Methylation and Histone Modifications at CpG

The specific histone modification H3K27me3 may associate with DNA methylation at CGI promoter and has higher H3K27me3 enrichment in cancer cells [39]. Also, in colon cancer cells, most of the promoter genes are enriched with higher H3K27me3, whereas in the controls, unmethylated CGI lacks H3K27me3 mark [40]. The similarity between DNA methylation and H3K27me3 are both repressive in function. But DNA methylation is a stable form of repression, whereas H3K27me3 is not [41]. The coordination of these two marks also plays a critical role in early mammalian development and a model has been proposed [25]. In the early developmental stage, the active promoters with high CG content are enriched with H3K27me3, which silence the non-expressing lineages [42]. Also, another group has shown that H3K7me3 enrichment sites exhibit higher methylation level in undifferentiated cells [43]. On the contrary, some groups have also argued that DNA methylation and H3K27me3 are mutually exclusive in normal and cancer cells [11, 44]. Nevertheless, co-occurrence of DNA methylation and H3K27me3 has been shown to be involved in early mammalian development and tumor progression. The interplay between DNA methylation and H3K27me3 and/or other histone modifications is yet to be revealed.

5 PRC-Mediated H3K27me3 Action: During Developmental Stages and in Cancer Cells

During developmental stages, embryonic cells are regulated by epigenetic factors such as DNA methylation and histone modifications by change in gene expression for cell differentiation and maintenance. Apart from these modifications, PcG proteins also play a vital role in early mammalian development [45, 46]. PcG repressive complex was first identified in *Drosophila*, which is involved in gene silencing to maintain cell fate [47]. PcG proteins function as two protein complexes into polycomb repressive complex 1 (PRC1) and polycomb repressive complex 2 (PRC2). EED, SUZ12, and EZH2 are the part of PRC2 which catalyzes H3K4me2/me3 [48]. SET domain of histone methyltransferase EZH2, as part of PRC2 complex, was shown to catalyze H3K27me3 [49], and the repressive mark H3K27me3 is considered as the hallmark for PcG-mediated silencing (Fig. 3) [50–52].

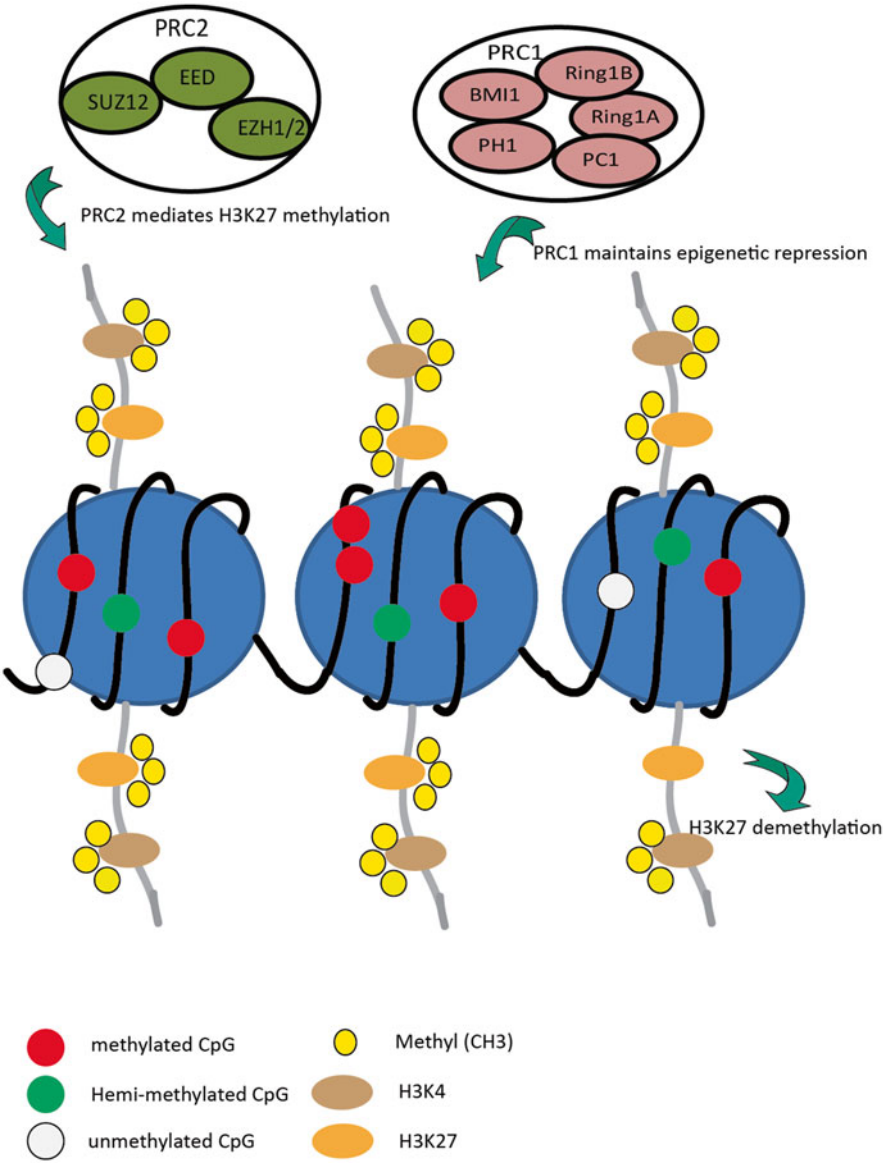


Fig. 3 Histone activation mark (H3K4me3) and repressive mark (H3K27me3) act as bivalent modifications in regulating developmental genes. Epigenetic repression of pro-differentiation genes is maintained by PRC1 and PRC2 complexes. H3K27me3 is catalyzed by EZH2 methyltransferase of PRC2 complex which leads to transcriptional repression of developmental genes. PRC1 is believed to coordinate with PRC2 during the process. CGI methylation is also contributed to gene expression regulation. Loss of H3K27me3 mark at specific promoter is considered to be transcriptional activation of genes. In cancer cells, H3K27me3 and DNA hypermethylation coexist together, and this PcG-mediated H3K27me3 silence machinery appears to be the hallmark of cancer. (The figure is adapted from the article [82])

PcG-mediated gene silencing plays an essential role in maintaining normal state of stem cells and progenitor cells and controlling specific-tissue types [51, 52].

Interestingly, other evidences are also suggested that PRC1 complex is also recruited along with PRC2 complex, which catalyze H3K27me3 for chromatin compaction [50, 53]. Chromatin landscape of ES cells revealed that PcG target sites contain large region of H3K27me3 repressive mark and activation mark H3K4me3 at transcriptional start site (TSS) [54, 55]. The genomic region that consists of opposing modifications is called as "bivalent domains" that believed to drive potential genes either into active or inactive state (Fig. 3). The plasticity of chromatin is maintained by the bivalent histone marks with low gene expression in developmental genes and depends on the differentiation signal; shifting in monovalent state leads to transcriptional activation by open chromatin conformation [54]. In the case of PcG target genes, CpG-containing promoter is protected from de novo methylation during implantation [56]. However, during the development process from stem cell to differentiated state, numbers of gene sequences undergo de novo methylation by PcG complex [57, 58]. Some groups showed that PcG-deficient ES cells enter differentiation stage but fail to maintain differentiated phenotypes [59, 60].

6 Aberrant DNA Methylation: Hypermethylation and Hypomethylation

The rapidly growing evidences corroborate the notion that malfunction in epigenetic network may cause aberrant DNA methylation [9]. DNA hypermethylation and global hypomethylation alterations are associated with neoplastic transformation [61]. In mouse ES cells, promoter region of some genes has a combination of repressive mark H3K27me3 and active mark H3K4me2 [62], which is a bivalent state of these histone modifications in ES cells that have been observed before [54, 63]. A group has found that in adult cancer cells, aberrant DNA methylation was observed at an H3K27me3-enriched region. H3K27me3 is also associated with hypomethylation at DNA promoters. Aberrant promoter methylation was frequently observed in many types of cancers, and in some cancers such as gastric cancer, aberrant methylation was involved in silencing tumor suppressor genes [64]. Feinberg and colleagues have shown that global DNA hypomethylation might associate with cancer initiation just like DNA hypermethylation [65]. There are growing evidences that link global DNA hypomethylation with genomic instability, and it might have long-term consequences in cancer [66, 67]. Activation of oncogene (cMYC) is stimulated by global hypomethylation [68]. Despite more results are reported on hypermethylation in cancer, hypomethylation is also considered to have strong correlation in many types of cancer (Fig. 2).

7 Technological Development in Cancer Epigenomics

A set of methodology is available for studying epigenetic modifications in normal and cancer cells (Fig. 4). The existing technologies can be used to assess different cells and tissues in either a gene-specific or genome-wide manner. Cancer epigenomics research is slowed down behind cancer genetics due to nonavailability of novel techniques and technical limitation to study epigenetic mechanism. The development of bisulfite treatment is a boon to DNA methylation research, which converts unmethylated cytosine to uracil and leaves the methylated cytosine. The bisulfite conversion then can be coupled with PCR amplification to analyze DNA methylome [69]. A challenging problem in conventional bisulfite sequencing is how to measure repeated sequence region like CGI region or heterochromatin region. Genome-wide bisulfite sequencing technologies can overcome this

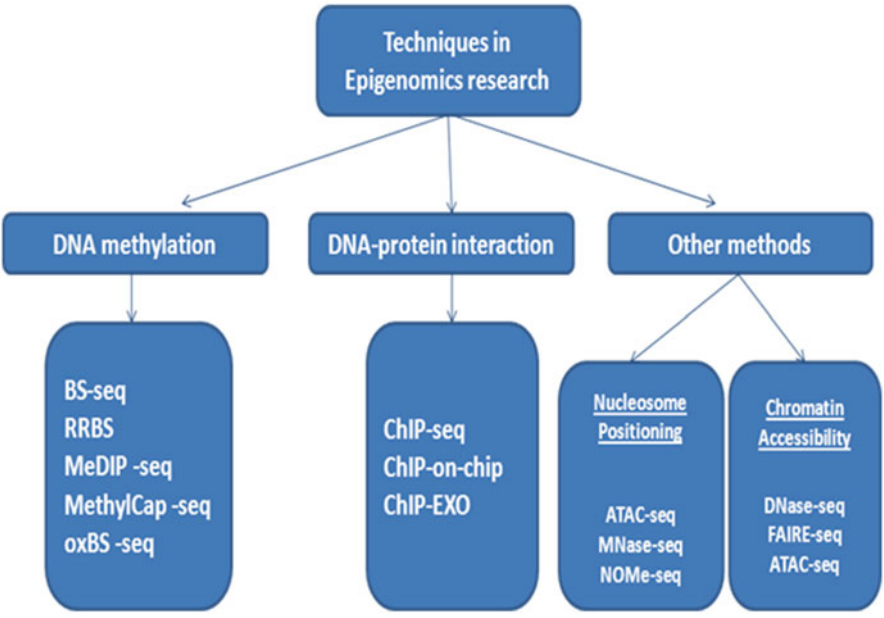


Fig. 4 Technique in epigenomics research. A series of techniques are shown in the figure, which are used to study DNA methylation, DNA-protein interactions, and chromatin remodeling—bisulfite sequencing (BS-seq), reduced representation bisulfite sequencing (RRBS), methylated DNA immunoprecipitation sequencing (MeDIP-seq), whole-genome shotgun bisulfite sequencing (MethylCap-seq), micrococcal nuclease digestion sequencing (MNase-seq), oxidative bisulfite sequencing (oxBS-seq), chromatin immunoprecipitation sequencing (ChIP-seq), chromatin immunoprecipitation exonuclease digestion (ChIP-exo), assay for transposase-accessible chromatin sequencing (ATAC-seq), nucleosome occupancy and methylome sequencing (NOMe-seq), DNase I hypersensitive sites sequencing (DNase-seq), formaldehyde-assisted isolation of regulatory elements sequencing (FAIRE-seq)

obstacle, for instance, reduced representation bisulfite sequencing (RRBS) was developed to mainly target high CpG genomic regions by combining MspI restriction digestion and bisulfite conversion [70]. RRBS method can be used to study aberrant hypermethylation and hypomethylation at CpG islands in different types of cancer cells [71]. To study diverse histone modification patterns, chromatin immunoprecipitation sequencing (ChIP-seq) method was developed to study DNA-protein interaction to decipher the role of histone mark, transcription factors, and other epigenetic regulators followed by high-throughput sequencing that provides genome-wide information. Bisulfite sequencing and ChIP-seq are the mostly used techniques in epigenomics research to study DNA methylation and histone modifications. Moreover, other techniques were developed such as MNase-seq, ATAC-seq, and NOMe-seq to study nucleosome positioning, DNase-seq and FAIRE-seq to study chromatin accessibility, and RNA-seq and SAGE-seq to study RNA level. However, these techniques have its unique purpose to study different mechanism individually. Despite the benefits of these robust techniques, studying the interaction between epigenetic modifications such as DNA methylation and histone modifications is limited.

8 ChIP-BS-seq and Summary

ChIP-BS-seq is a method to study DNA methylation directly at enriched histone marks region. ChIP-BS-seq is a fusion of chromatin immunoprecipitation and bisulfite conversion techniques and coupling with high-throughput sequencer for epigenetic profiling. The main advantage is that it can provide more information on DNA methylation at enriched histone marks in cell lines or tissues. Particularly it can be utilized to study interplay between DNA methylation and histone modifications followed by epigenetic profiling that can provide better insight on complex epigenetic regulation (Fig. 5).

ChIP-BS-seq can be carried out either in cells or tissues. First, cells or tissues are cross-linked with formaldehyde. Then, the input material is lysed and chromatin is sheared by sonication or MNase digestion method. The sheared chromatin is immobilized with beads followed by immunoprecipitation for histone modifications or transcription factors of our interest. ChIPed DNA is purified and bisulfite reaction is carried out on ChIPed DNA. After bisulfite conversion DNA library is prepared by addition of adapters and then purified DNA is sent for high-throughput sequencing. The sequenced DNA is analyzed by bioinformatics tools for epigenomic profiling.

In our study [72], we used ChIP-BS-seq to study H3K4me3 and H3K27me3 marks in a normal cell line (YH lymphoblastoid) and three cancer cell lines (one cervical cancer cell line (HeLa) and two gastric cancer (GC) cell lines BGC-823 and AGS). One of our main goals was to study the cross talk between DNA methylation and histone modification at CpG island—especially to compare aberrant methylation at CpG island in normal and cancer cell lines at different genomic elements.

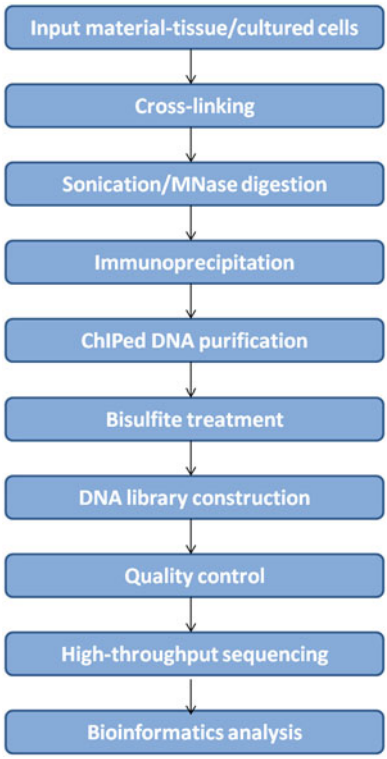


Fig. 5 ChIP-BS-seq workflow**.** ChIP-BS-seq is a method to study interplay between DNA methylation and histone modification. In ChIP-BS-seq, DNA methylation is studied at enriched histone modification. First, cells or tissues are cross-linked and lysed, and the chromatin is sheared and immunoprecipitated with antibody of our interest. Bisulfite reaction is carried out on ChIPed DNA, DNA library is prepared by addition of adapters and then purified DNA is sent for high-throughput sequencing. The sequenced DNA is analyzed by bioinformatics tools

With ChIP-BS-seq, we first determined H3K27me3-enriched regions for normal and three cancer cell lines. And we obtained the methylation status for these cell lines at an H3K27me3-enriched region by modifying ChIP protocol by using 50-bp pair-end sequencing instead of regular single-end sequencing. One hundred million mapping reads were obtained for the individual cell lines, and reads were aligned in the SOAP genome analyzer. From the aligned mapping reads, we found that CpG sites, transcription sites (±500), and exon elements are enriched with H3K27me3 mark at gastric cancer (GC) cell line (BGC-823 and AGS), but in the case of YH cell line, CpG sites are enriched with H3K27me3 mark. In contrast, in HeLa cells, a reverse pattern was observed that is the highest H3K27me3 enrichment at the intergenic region. Similar to variable histone modifications patterns, we also observed the different DNA methylation patterns for different cell populations. To verify this result with different sets of cell populations, we used ChIP-BS-seq data profiling of H3K27me3 marks in three different cell lines from TCGA database—a prostate cancer cell line LNCaP, a normal prostate epithelial cell line PrEC, and a colon cancer cell line HCT116. From the downloaded ChIP-BS-seq data, H3K27me3 pattern of HCT116 cell lines was similar to HeLa cell line, but it was different in LNCaP and PrEC. Hence, H3K27me3-enriched regions have variable DNA methylation patterns. To emphasize further, we want to highlight our findings concisely in the following sections.

8.1 Variable Patterns of H3K27me3 and DNA Methylation in Normal and Cancer Cell Lines

Because of variable patterns in H3K27me3 region and DNA methylation at enriched region, we raised an interesting question whether these variable patterns are due to cell-specific epigenetic signature. To explore further, we focused on promoter overlapped CGI regions, because most of these regions are enriched with H3K27me3 mark. We did cluster analysis for seven cell lines based on average values of methylation at H3K27me3-enriched regions. In cluster analysis, two normal cell lines (YH and PrEC) clustered together from the rest of the five cancer cell lines. Even though these cells are orientated from a unique cellular pathway, these results suggested that DNA methylation at an H3K27me3-enriched region might have onco-epigenomic signature.

8.2 Co-occurrence of Hypermethylation and H3K27me3 at CGI Promoter in Cancer Cell Lines

We screened for an H3K27me3-enriched genomic region that is overlapped among seven cell lines, and we found 223 H3K27me3-enriched genes and 5143 H3K27me3-deficient genes. We categorized the genes further based on whether CGIs are present in their promoter that 64 % of 228 H3K27me3-enriched genes containing CGI at promoter whereas only 45 % of 5143 H3K27me3-deficient genes containing CGI at promoter. Next, a comparison among five cancer cell lines and two normal cell lines shows that most of the CGI-containing promoter genes are hypermethylated in cancer lines than in normal cell lines. Most of these genes are highly methylated in HCT116 and AGS cell line, low methylation in LNCaP, and median methylation for BGC-823 and HeLa cell lines in the majority of these genes. Hence, the methylation at a CGI of H3K27me3 of enriched regions is varied among different cell populations. Because of cancer-specific methylation patterns, DNA-methylated region of an H3K27me3-enriched region could be an epigenetic signature for cancer studies.

8.3 Hypermethylation at CGIs of an H3K27me3-Enriched Region but Not at H3K4me3

ChIP-BS-seq was performed for H3K4me3 (activating mark) in all three cell lines (AGS, BGC-823, and YH) to compare hypermethylation between H3K27me3- and H3K4me3-enriched regions. DNA methylation level is extremely low at H3K27me3 (repressive mark)-enriched genomic region as expected. There are substantial evidences that have shown the negative correlation between DNA methylation and H3K4me3 level. Our results also suggested that DNA methylation level is indirectly

proportional to H3K4me3 enrichment. H3K27me3 and H3K4me3 co-occurred as a bivalent mark in stem cells, in differentiated cells, and in cancer cells [73]. These marks are also present together in PcG-expressed genes [55, 74, 75]. To understand the bivalence nature of these marks in our cell lines, we categories the genes into four—H3K27me3 high enriched, H3K4me3 high enriched, H3K4me3 and H3K27me3 enriched (bivalent), and neither H3K4me3 nor H3K27me3 enriched. We found that DNA methylation level at TSS is hypomethylated for "H3K4me3 high" or "bivalent" categories. But in cancer cells, "H3K27me3 high" shows a higher level of methylation than in two categories, whereas the methylation level is low in YH cells.

The H3K27me3-enriched region was categorized into CGI-containing promoter genes and CGI-deficient promoter genes, and the methylation level was compared among normal and cancer cells. CGI-containing promoter genes are hypermethylated in cancer cells but hypomethylated in normal cells. Increased DNA methylation of H3K27me3 enrichment correlates with hypermethylation at a CGI in cancer cells but not in normal cells. Moreover, H3K27me3-bound genes with CGI promoter hypermethylation in cell lines were also hypermethylated in primary cancer tissues. To sum it all, due to cancer heterogeneity, DNA methylation level was varied among different cell population. By using ChIP-BS-seq technology, a correlation between H3K27me3 histone mark and DNA methylation was revealed in different cancer cells, and it showed that some genes enriched with H3K27me3 have hypermethylation at a CGI in cancer cells but not in normal cells (for detailed analysis and figures, please refer to our article [72]).

9 Conclusion and Future Prospects

In this chapter, we have emphasized the relationship between DNA methylation and histone modification and how it can be studied using ChIP-BS-seq on cancer context. It is evident that both these modifications have its own function in regulating transcriptional network. For instance, histone modifications may have an effect on DNA methylation, and alternatively, DNA methyl-binding protein may also be conversed with histone modification to maintain nucleosome function during gene regulation. Simply, this study was quite preliminary, and further analysis is to be done on different type of cancer tissues to understand mechanism/heterogeneity in tumor. However, a couple of group has shown that DNA methylation might also be mediated by microRNAs [76, 77]. It needs to be clarified whether DNA methylation and histone modifications have direct or indirect mechanism. Nevertheless, there are many mechanistic details of this epigenetic mechanism that should be resolved either in normal developmental process or in tumorigenesis.

One key point could be quite obvious that available technology is "bottleneck" to understand complex mechanism or it may mislead the underlying mystery. To overcome this drawback, development of single-cell technology is maturing to study biological mechanism in single cell. Single-cell technology is to be integrated

with epigenetic techniques to overcome technical limitation such as paucity of input material especially in rare stem cells and cancer tissue biopsies and to study heterogeneity among cells and tissues [78]. Recently, single-cell technology is developed for genome-wide epigenetic profiling to study DNA methylation and CpG island [79, 80]. To get a clear picture on DNA methylation, histone modifications and noncoding RNA mechanisms and their complex interactions and sophisticated technologies are required to embellish precise knowledge on cancer misregulation. Without a flinch of doubt, a load of work is to be done with the help of evolving new technologies, and novel strategies might question the prevailing dogma in epigenomics research.

References

1. Portela A, Esteller M. Epigenetic modifications and human disease. Nat Biotechnol. 2010;28(10):1057–68.
2. Bird AP, Wolffe AP. Methylation-induced repression—belts, braces, and chromatin. Cell. 1999;99(5):451–4.
3. Bird A. DNA methylation patterns and epigenetic memory. Genes Dev. 2002;16(1):6–21.
4. Baylin SB, Herman JG. DNA hypermethylation in tumorigenesis: epigenetics joins genetics. Trends Genet. 2000;16(4):168–74.
5. Jones PA, Laird PW. Cancer epigenetics comes of age. Nat Genet. 1999;21(2):163–7.
6. Nguyen CT, Gonzales FA, Jones PA. Altered chromatin structure associated with methylation-induced gene silencing in cancer cells: correlation of accessibility, methylation, MeCP2 binding and acetylation. Nucleic Acids Res. 2001;29(22):4598–606.
7. Fahrner JA, Eguchi S, Herman JG, Baylin SB. Dependence of histone modifications and gene expression on DNA hypermethylation in cancer. Cancer Res. 2002;62(24):7213–8.
8. Hanahan D, Weinberg RA. The hallmarks of cancer. Cell. 2000;100(1):57–70.
9. Esteller M. Cancer epigenomics: DNA methylomes and histone-modification maps. Nat Rev Genet. 2007;8(4):286–98.
10. Statham AL, Robinson MD, Song JZ, Coolen MW, Stirzaker C, Clark SJ. Bisulfite sequencing of chromatin immunoprecipitated DNA (BisChIP-seq) directly informs methylation status of histone-modified DNA. Genome Res. 2012;22(6):1120–7.
11. Brinkman AB, Gu H, Bartels SJ, Zhang Y, Matarese F, Simmer F, Marks H, Bock C, Gnirke A, Meissner A, et al. Sequential ChIP-bisulfite sequencing enables direct genome-scale investigation of chromatin and DNA methylation cross-talk. Genome Res. 2012;22(6):1128–38.
12. Richmond TJ, Luger K, Mäder AW, Richmond RK, Sargent DF. Crystal structure of the nucleosome core particle at 2.8 Å resolution. Nature. 1997;389(6648):251–60.
13. Kornberg RD, Lorch Y. Twenty-five years of the nucleosome, fundamental particle of the eukaryote chromosome. Cell. 1999;98(3):285–94.
14. Kouzarides T. Chromatin modifications and their function. Cell. 2007;128(4):693–705.
15. Fuks F. DNA methylation and histone modifications: teaming up to silence genes. Curr Opin Genet Dev. 2005;15(5):490–5.
16. Bannister AJ, Kouzarides T. Regulation of chromatin by histone modifications. Cell Res. 2011;21(3):381–95.
17. Rea S, Eisenhaber F, O'Carroll D, Strahl BD, Sun ZW, Schmid M, Opravil S, Mechtler K, Ponting CP, Allis CD, et al. Regulation of chromatin structure by site-specific histone H3 methyltransferases. Nature. 2000;406(6796):593–9.
18. Yang Y, Bedford MT. Protein arginine methyltransferases and cancer. Nat Rev Cancer. 2012;13(1):37–50.

19. Shi Y, Lan F, Matson C, Mulligan P, Whetstine JR, Cole PA, Casero RA, Shi Y. Histone demethylation mediated by the nuclear amine oxidase homolog LSD1. Cell. 2004;119(7): 941–53.
20. Karytinos A, Forneris F, Profumo A, Ciossani G, Battaglioli E, Binda C, Mattevi A. A novel mammalian flavin-dependent histone demethylase. J Biol Chem. 2009;284(26):17775–82.
21. Ciccone DN, Su H, Hevi S, Gay F, Lei H, Bajko J, Xu G, Li E, Chen T. KDM1B is a histone H3K4 demethylase required to establish maternal genomic imprints. Nature. 2009;461(7262):415–8.
22. Hayami S, Kelly JD, Cho H-S, Yoshimatsu M, Unoki M, Tsunoda T, Field HI, Neal DE, Yamaue H, Ponder BAJ, et al. Overexpression of LSD1 contributes to human carcinogenesis through chromatin regulation in various cancers. Int J Cancer. 2011;128(3):574–86.
23. Rotili D, Mai A. Targeting histone demethylases: a new avenue for the fight against cancer. Genes Cancer. 2011;2(6):663–79.
24. Bird AP. CpG-rich islands and the function of DNA methylation. Nature. 1986;321(6067): 209–13.
25. Cedar H, Bergman Y. Linking DNA methylation and histone modification: patterns and paradigms. Nat Rev Genet. 2009;10(5):295–304.
26. Ikegami K, Ohgane J, Tanaka S, Yagi S, Shiota K. Interplay between DNA methylation, histone modification and chromatin remodeling in stem cells and during development. Int J Dev Biol. 2009;53(2–3):203–14.
27. Irizarry RA, Ladd-Acosta C, Wen B, Wu Z, Montano C, Onyango P, Cui H, Gabo K, Rongione M, Webster M, et al. The human colon cancer methylome shows similar hypo- and hypermethylation at conserved tissue-specific CpG island shores. Nat Genet. 2009;41(2):178–86.
28. Cross SH, Bird AP. CpG islands and genes. Curr Opin Genet Dev. 1995;5(3):309–14.
29. Greger V, Passarge E, Höpping W, Messmer E, Horsthemke B. Epigenetic changes may contribute to the formation and spontaneous regression of retinoblastoma. Hum Genet. 1989;83(2):155–8.
30. Merlo A, Herman JG, Mao L, Lee DJ, Gabrielson E, Burger PC, Baylin SB, Sidransky D. 5′ CpG island methylation is associated with transcriptional silencing of the tumour suppressor p16/CDKN2/MTS1 in human cancers. Nat Med. 1995;1(7):686–92.
31. Herman JG, Merlo A, Mao L, Lapidus RG, Issa JP, Davidson NE, Sidransky D, Baylin SB. Inactivation of the CDKN2/p16/MTS1 gene is frequently associated with aberrant DNA methylation in all common human cancers. Cancer Res. 1995;55(20):4525–30.
32. Gonzalez-Zulueta M, Bender CM, Yang AS, Nguyen T, Beart RW, Van Tornout JM, Jones PA. Methylation of the 5' CpG island of the p16/CDKN2 tumor suppressor gene in normal and transformed human tissues correlates with gene silencing. Cancer Res. 1995;55(20):4531–5.
33. Esteller M, Corn PG, Baylin SB, Herman JG. A gene hypermethylation profile of human cancer. Cancer Res. 2001;61:3225–9.
34. Reik W, Illingworth RS, Gruenewald-Schneider U, Webb S, Kerr ARW, James KD, Turner DJ, Smith C, Harrison DJ, Andrews R, et al. Orphan CpG islands identify numerous conserved promoters in the mammalian genome. PLoS Genet. 2010;6(9):e1001134.
35. Maunakea AK, Nagarajan RP, Bilenky M, Ballinger TJ, D'Souza C, Fouse SD, Johnson BE, Hong C, Nielsen C, Zhao Y, et al. Conserved role of intragenic DNA methylation in regulating alternative promoters. Nature. 2010;466(7303):253–7.
36. Saxonov S, Berg P, Brutlag DL. A genome-wide analysis of CpG dinucleotides in the human genome distinguishes two distinct classes of promoters. Proc Natl Acad Sci U S A. 2006;103(5):1412–7.
37. Koche RP, Smith ZD, Adli M, Gu H, Ku M, Gnirke A, Bernstein BE, Meissner A. Reprogramming factor expression initiates widespread targeted chromatin remodeling. Cell Stem Cell. 2011;8(1):96–105.
38. Doege CA, Inoue K, Yamashita T, Rhee DB, Travis S, Fujita R, Guarnieri P, Bhagat G, Vanti WB, Shih A, et al. Early-stage epigenetic modification during somatic cell reprogramming by Parp1 and Tet2. Nature. 2012;488(7413):652–5.
39. Viré E, Brenner C, Deplus R, Blanchon L, Fraga M, Didelot C, Morey L, Van Eynde A, Bernard D, Vanderwinden J-M, et al. The Polycomb group protein EZH2 directly controls DNA methylation. Nature. 2005;439(7078):871–4.

40. Schlesinger Y, Straussman R, Keshet I, Farkash S, Hecht M, Zimmerman J, Eden E, Yakhini Z, Ben-Shushan E, Reubinoff BE, et al. Polycomb-mediated methylation on Lys27 of histone H3 pre-marks genes for de novo methylation in cancer. Nat Genet. 2006;39(2):232–6.
41. Smith ZD, Meissner A. DNA methylation: roles in mammalian development. Nat Rev Genet. 2013;14(3):204–20.
42. Xie W, Schultz Matthew D, Lister R, Hou Z, Rajagopal N, Ray P, Whitaker John W, Tian S, Hawkins RD, Leung D, et al. Epigenomic analysis of multilineage differentiation of human embryonic stem cells. Cell. 2013;153(5):1134–48.
43. Gifford Casey A, Ziller Michael J, Gu H, Trapnell C, Donaghey J, Tsankov A, Shalek Alex K, Kelley David R, Shishkin Alexander A, Issner R, et al. Transcriptional and epigenetic dynamics during specification of human embryonic stem cells. Cell. 2013;153(5):1149–63.
44. Hahn MA, Hahn T, Lee DH, Esworthy RS, Bw K, Riggs AD, Chu FF, Pfeifer GP. Methylation of polycomb target genes in intestinal cancer is mediated by inflammation. Cancer Res. 2008;68(24):10280–9.
45. O'Carroll D, Erhardt S, Pagani M, Barton SC, Surani MA, Jenuwein T. The polycomb-group gene Ezh2 is required for early mouse development. Mol Cell Biol. 2001;21(13):4330–6.
46. Wang J, Rao S, Chu J, Shen X, Levasseur DN, Theunissen TW, Orkin SH. A protein interaction network for pluripotency of embryonic stem cells. Nature. 2006;444(7117): 364–8.
47. Lewis EB. A gene complex controlling segmentation in Drosophila. Nature. 1978;276(5688): 565–70.
48. Müller J, Hart CM, Francis NJ, Vargas ML, Sengupta A, Wild B, Miller EL, O'Connor MB, Kingston RE, Simon JA. Histone methyltransferase activity of a drosophila polycomb group repressor complex. Cell. 2002;111(2):197–208.
49. Cao R. Role of histone H3 lysine 27 methylation in polycomb-group silencing. Science. 2002;298(5595):1039–43.
50. Margueron R, Reinberg D. The polycomb complex PRC2 and its mark in life. Nature. 2011;469(7330):343–9.
51. Ringrose L, Paro R. Epigenetic regulation of cellular memory by the polycomb and trithorax group proteins. Annu Rev Genet. 2004;38(1):413–43.
52. Sparmann A, van Lohuizen M. Polycomb silencers control cell fate, development and cancer. Nat Rev Cancer. 2006;6(11):846–56.
53. Francis NJ. Chromatin compaction by a polycomb group protein complex. Science. 2004;306(5701):1574–7.
54. Bernstein BE, Mikkelsen TS, Xie X, Kamal M, Huebert DJ, Cuff J, Fry B, Meissner A, Wernig M, Plath K, et al. A bivalent chromatin structure marks key developmental genes in embryonic stem cells. Cell. 2006;125(2):315–26.
55. Zhao XD, Han X, Chew JL, Liu J, Chiu KP, Choo A, Orlov YL, Sung W-K, Shahab A, Kuznetsov VA, et al. Whole-genome mapping of histone H3 Lys4 and 27 trimethylations reveals distinct genomic compartments in human embryonic stem cells. Cell Stem Cell. 2007;1(3):286–98.
56. Brandeis M, Frank D, Keshet I, Siegfried Z, Mendelsohn M, Names A, Temper V, Razin A, Cedar H. Spl elements protect a CpG island from de novo methylation. Nature. 1994;371(6496):435–8.
57. Mohn F, Weber M, Rebhan M, Roloff TC, Richter J, Stadler MB, Bibel M, Schübeler D. Lineage-specific polycomb targets and de novo DNA methylation define restriction and potential of neuronal progenitors. Mol Cell. 2008;30(6):755–66.
58. Meissner A, Mikkelsen TS, Gu H, Wernig M, Hanna J, Sivachenko A, Zhang X, Bernstein BE, Nusbaum C, Jaffe DB, et al. Genome-scale DNA methylation maps of pluripotent and differentiated cells. Nature. 2008;454(7205):766–70.
59. Pasini D, Bracken AP, Hansen JB, Capillo M, Helin K. The polycomb group protein Suz12 is required for embryonic stem cell differentiation. Mol Cell Biol. 2007;27(10):3769–79.
60. Leeb M, Pasini D, Novatchkova M, Jaritz M, Helin K, Wutz A. Polycomb complexes act redundantly to repress genomic repeats and genes. Genes Dev. 2010;24(3):265–76.
61. Herman JG, Baylin SB. Gene silencing in cancer in association with promoter hypermethylation. N Engl J Med. 2003;349(21):2042–54.

62. Ohm JE, McGarvey KM, Yu X, Cheng L, Schuebel KE, Cope L, Mohammad HP, Chen W, Daniel VC, Yu W, et al. A stem cell–like chromatin pattern may predispose tumor suppressor genes to DNA hypermethylation and heritable silencing. Nat Genet. 2007;39(2):237–42.
63. Andrews 3rd DF, Nemunaitis J, Tompkins C, Singer JW. Effect of 5-azacytidine on gene expression in marrow stromal cells. Mol Cell Biol. 1989;9(6):2748–51.
64. Ushijima T, Sasako M. Focus on gastric cancer. Cancer Cell. 2004;5(2):121–5.
65. Feinberg AP, Vogelstein B. A technique for radiolabeling DNA restriction endonuclease fragments to high specific activity. Anal Biochem. 1983;132(1):6–13.
66. Eden A. Response to comment on "chromosomal instability and tumors promoted by DNA hypomethylation" and "induction of tumors in mice by genomic hypomethylation". Science. 2003;302(5648):1153c.
67. Yang AS. Comment on "chromosomal instability and tumors promoted by DNA hypomethylation" and "induction of tumors in mice by genomic hypomethylation". Science. 2003; 302(5648):1153b–1153.
68. Das PM. DNA methylation and cancer. J Clin Oncol. 2004;22(22):4632–42.
69. Herman JG, Graff JR, Myohanen S, Nelkin BD, Baylin SB. Methylation-specific PCR: a novel PCR assay for methylation status of CpG islands. Proc Natl Acad Sci U S A. 1996;93(18): 9821–6.
70. Meissner A. Reduced representation bisulfite sequencing for comparative high-resolution DNA methylation analysis. Nucleic Acids Res. 2005;33(18):5868–77.
71. Versteeg R. Aberrant methylation in cancer. Am J Hum Genet. 1997;60(4):751–4.
72. Gao F, Ji G, Gao Z, Han X, Ye M, Yuan Z, Luo H, Huang X, Natarajan K, Wang J, et al. Direct ChIP-bisulfite sequencing reveals a role of H3K27me3 mediating aberrant hypermethylation of promoter CpG islands in cancer cells. Genomics. 2014;103(2–3):204–10.
73. Blagosklonny MV, Ke X-S, Qu Y, Rostad K, Li W-C, Lin B, Halvorsen OJ, Haukaas SA, Jonassen I, Petersen K, et al. Genome-wide profiling of histone H3 lysine 4 and lysine 27 trimethylation reveals an epigenetic signature in prostate carcinogenesis. PLoS One. 2009;4(3):e4687.
74. Mikkelsen TS, Ku M, Jaffe DB, Issac B, Lieberman E, Giannoukos G, Alvarez P, Brockman W, Kim T-K, Koche RP, et al. Genome-wide maps of chromatin state in pluripotent and lineage-committed cells. Nature. 2007;448(7153):553–60.
75. Pan G, Tian S, Nie J, Yang C, Ruotti V, Wei H, Jonsdottir GA, Stewart R, Thomson JA. Whole-genome analysis of histone H3 lysine 4 and lysine 27 methylation in human embryonic stem cells. Cell Stem Cell. 2007;1(3):299–312.
76. Benetti R, Gonzalo S, Jaco I, Muñoz P, Gonzalez S, Schoeftner S, Murchison E, Andl T, Chen T, Klatt P. A mammalian microRNA cluster controls DNA methylation and telomere recombination via Rbl2-dependent regulation of DNA methyltransferases. Nat Struct Mol Biol. 2008;15(3):268–79.
77. Sinkkonen L, Hugenschmidt T, Berninger P, Gaidatzis D, Mohn F, Artus-Revel CG, Zavolan M, Svoboda P, Filipowicz W. MicroRNAs control de novo DNA methylation through regulation of transcriptional repressors in mouse embryonic stem cells. Nat Struct Mol Biol. 2008;15(3):259–67.
78. Picelli S, Faridani OR, Björklund ÅK, Winberg G, Sagasser S, Sandberg R. Full-length RNA-seq from single cells using Smart-seq2. Nat Protoc. 2014;9(1):171–81.
79. Smallwood SA, Lee HJ, Angermueller C, Krueger F, Saadeh H, Peat J, Andrews SR, Stegle O, Reik W, Kelsey G. Single-cell genome-wide bisulfite sequencing for assessing epigenetic heterogeneity. Nat Methods. 2014;11(8):817–20.
80. Guo H, Zhu P, Wu X, Li X, Wen L, Tang F. Single-cell methylome landscapes of mouse embryonic stem cells and early embryos analyzed using reduced representation bisulfite sequencing. Genome Res. 2013;23(12):2126–35.
81. Lee JJ, Murphy GF, Lian CG. Melanoma epigenetics: novel mechanisms, markers, and medicines. Lab Invest. 2014;94(8):822–38.
82. Kashyap V, Rezende NC, Scotland KB, Shaffer SM, Persson JL, Gudas LJ, Mongan NP. regulation of stem cell pluripotency and differentiation involves a mutual regulatory circuit of the nanog, OCT4, and SOX2 pluripotency transcription factors with polycomb repressive complexes and stem cell microRNAs. Stem Cells Dev. 2009;18(7):1093–108.

Integrative Analysis Identifies Transcription Factor–DNA Methylation Relationships and Introduces New Avenues for Translating Cancer Epigenetics into the Clinic

Matthew H. Ung, Shaoke Lou, Frederick S. Varn, and Chao Cheng

Abstract DNA methylation is essential in the regulation of gene expression and its misregulation has been implicated in a vast array of cancer types. The causal relationships between DNA methylation, transcription factor binding, chromatin structure, and gene expression are not well elucidated. Regardless, recent research has shown DNA methylation to be a key component in these regulatory modules, suggesting that dissecting the mechanisms underlying the formation of DNA methylation patterns can provide insight into cancer regulomes. In addition, DNA methylation information can potentially serve as novel biomarkers that indicate cancer type, predict patient prognosis, or be used to identify drug targets. Because transcription factors are key players in transcriptional regulation, there is reason to suspect they influence or are affected by genome-wide alterations in DNA methylation. This, coupled with the recent accumulation of large-scale genomic data, has allowed for high-resolution in silico dissection of transcription factor–DNA methylation relationships. In this chapter, we present an integrative analysis of ENCODE (Encyclopedia of DNA Elements) ChIP-seq and DNase I hypersensitivity data, coupled with TCGA (The Cancer Genome Atlas) breast cancer DNA methylation and gene expression data to study the interconnection between TFs with DNA methylation. Our results suggest that identifying DNA methylation patterning within transcription factor binding sites reveals information regarding transcription factor binding activity in breast cancer patients. From this, we discuss the translational potential of these novel findings and the power and flexibility of in silico analysis.

M.H. Ung • S. Lou • F.S. Varn
Department of Genetics, Geisel School of Medicine at Dartmouth, Hanover, NH 03755, USA
e-mail: matthew.h.ung.gr@dartmouth.edu; shaoke.lou@dartmouth.edu; Frederick.S.Varn.JR.GR@dartmouth.edu

C. Cheng (✉)
Department of Genetics, Geisel School of Medicine at Dartmouth, Hanover, NH 03755, USA

Institute for Quantitative Biomedical Sciences, Geisel School of Medicine at Dartmouth, Lebanon, NH 03766, USA

Norris Cotton Cancer Center, Geisel School of Medicine at Dartmouth, Lebanon, NH 03766, USA
e-mail: chao.cheng@dartmouth.edu

W. Wu, H. Choudhry (eds.), *Next Generation Sequencing in Cancer Research, Volume 2*, DOI 10.1007/978-3-319-15811-2_13

1 Introduction

Carcinogenesis arises from the acquisition of genetic aberrations that corrupt the instructional program of the cell and confer upon it, a drastic and uncontrollable proliferative advantage [1]. Cancer development is a mechanistically complex process that commences with an initial driver mutation(s) that causes further downstream genetic alterations (i.e., genomic instability, epigenetic alteration) with the majority of these being passenger mutations [1]. Alternatively, initial driver mutations may capitalize on existing mutations that have accumulated over many cycles of cellular division as in the case of leukemia [2]. The human genome encodes approximately 20,000 protein products and a multitude of noncoding regulatory elements such as miRNA. Therefore, it is imaginable that there exists a plethora of different combinations of mutations that can result in tumorigenesis in multiple tissue types. This heterogeneity presents a major obstacle in the development of cancer therapies because each cancer patient essentially suffers from a different disease [3]. In the clinic, subtyping of cancer into categories using molecular markers has yielded varied and partially effective results [4]. Thus, it is the goal of precision medicine to increase the customization of therapy to the level of individual patients.

To facilitate the discovery of new clinical technologies that could be used in precision medicine, there has recently been considerable interest and focus on understanding the epigenetic states of cancer. It has been postulated that alterations in epigenetic patterning contribute to and may accelerate the progression of cancer by perturbing normal transcriptional regulation. Ultimately, variation in epigenetic states among different cancer types introduces an additional layer of heterogeneity [5]. In general, DNA methylation is a biochemical process involving the addition of a methyl group to the 5th carbon of cytosine residues. Primarily, focus has been placed on DNA methylation that occurs in the context of CpG dinucleotides (CpGs). DNA methylation aberrations have been implicated in many cancer types; it has been observed that global hypomethylation accompanied by regional hypermethylation is a hallmark of many cancers [6]. The exact mechanisms by which changes in DNA methylation patterning contribute to dysfunctional transcriptional programs are not well elucidated. In most cases, DNA methylation is a passive process and is not the direct cause of misregulation, but is merely a consequence of transcription factor (TF) binding (i.e., CpGs not bound by transcription factors are methylated) [7, 8]. However, in some cases methylated CpGs possess functional relevance and may affect transcription factor binding or the three-dimensional conformation of chromatin [9]. Methylated CpGs can physically impede the binding of transcription factors or may actually be required for transcription factor binding [7, 10–13]. Furthermore, methylated CpGs may recruit docking proteins that facilitate site-specific binding of chromatin remodeling proteins such as histone deacetylases and histone methylases to modulate chromatin structure [14, 15]. Whether or not methylated CpGs play a causal role, identifying their patterns can serve as biomarkers in determining the transcriptional regulatory architecture of a cancer [16]. In summary, epigenetic markers can provide valuable information about the regulatory architecture of a biological system and provide new opportunities for developing

precision therapy in the clinic. In this chapter, we focus mainly on the relationships between transcription factors and the methylation states of their associated CpGs.

In order to interrogate CpG methylation and extract useful and interpretable data about genome-wide epigenetic states, advanced technology that exceeds that of basic molecular biology techniques is required. Fortunately, the postgenomic era has ushered in a new set of technologies that can probe the genome-wide molecular characteristics of a disease in a high-throughput and cost-effective manner. These ubiquitous genomic tools include, but are not limited to, DNA microarrays, protein arrays, DNA methylation arrays, and high-throughput sequencing (e.g., DNA-seq, RNA-seq, ChIP-seq). The data generated by these technologies have allowed for a data-driven approach to cancer biology. Much of these data are available in the public domain for biomedical researchers to download and apply to their own research. First, The Cancer Genome Atlas (TCGA) [17] is the most comprehensive cancer data repository to date and provides mutation, gene expression, DNA methylation, protein expression, copy number, and clinical information for thousands of cancer patients encompassing 32 cancer types. These data allow researchers to identify the molecular features most relevant for determining patient outcome and treatment strategies. Second, the ENCODE [18] project has generated a plethora of data on the functional elements in the human genome. One important set of data includes ChIP-seq data that is available for ~100 transcription factors in various cancer cell lines and treatment conditions. Furthermore, ENCODE provides data on the genomic coordinates of DNase I hypersensitive sites (DHS sites) which are regions of open chromatin most likely to be transcriptionally functional (high transcriptional activity) [8]. The applicability of these datasets increases significantly when integrated since it allows for more statistical power and provides more information to extract. In this chapter, we describe the integration of ENCODE ChIP-seq, ENCODE DHS, and TCGA DNA methylation datasets to analyze the relationships between transcription factor activity (i.e., ERα) and CpG methylation. More specifically, this integrative approach allows us to identify the methylation statuses of CpGs within transcription factor binding sites and DHS regions, and characterize how they fluctuate with respect to transcriptional activity within these functional regions of the genome. Ultimately, this will provide key insight into TF–DNA methylation relationships that reveal how DNA methylation is linked to transcriptional regulation.

Understanding transcriptional regulation is essential for determining treatment regimens for cancer patients. For example, the expression of ERα is always measured for each incidence of breast cancer and the decision to provide the patient with hormone therapy is partially dependent on ER status. As more information regarding the mechanisms of altered DNA methylation patterning becomes available, the relationship between epigenetics and transcription factor binding may facilitate the development of epigenetic testing in the clinic. Here, we present a proof of concept analysis that shows how integrative analysis can provide a high-resolution portrait of DNA methylation patterns within transcriptionally functional genomic regions. We present our analysis in breast cancer but this framework may also be applied to other cancer types or human diseases. In this study, we adopted breast cancer as a disease model because there are vast amounts of breast cancer data available in the public domain.

Additionally, breast cancer classification is well established in the clinic compared to other cancers allowing us to explore subtype-specific epigenetic changes. To summarize, we integrate ENCODE ChIP-seq, ENCODE DHS, DNA methylation, and TCGA gene expression data to analyze methylation levels of CpGs located within transcriptionally active regions and show how functional CpGs are distributed spatially across genes.

2 Results

2.1 Overview

In this chapter, we present an analysis strategy involving the integration of TCGA and ENCODE datasets to computationally dissect the intricacies of DNA methylation in the regulation of cancer transcriptomes (Fig. 1). Our study attempts to address

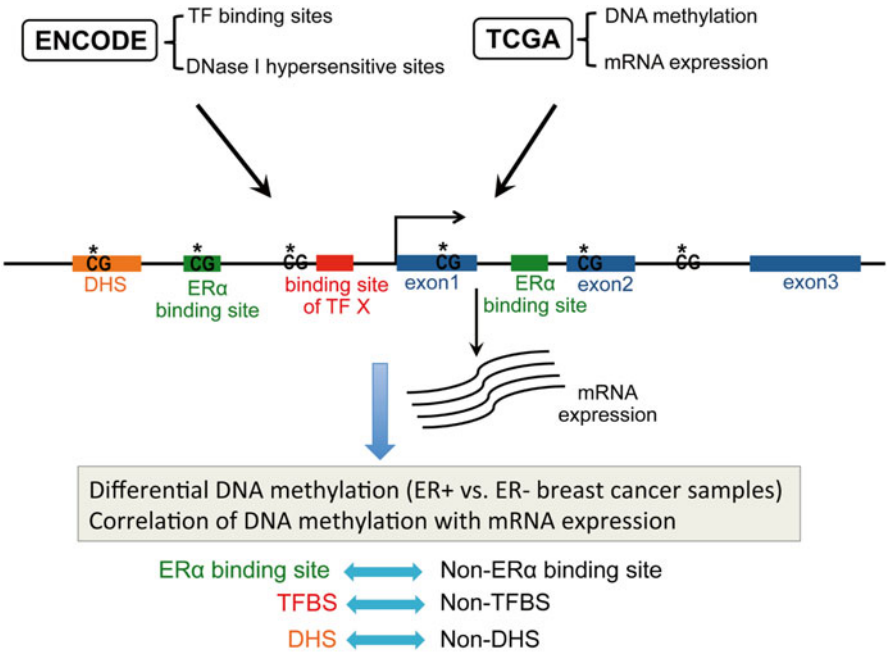


Fig. 1 Overflow of our analysis procedure. We utilized TCGA breast cancer DNA methylation and gene expression data and ENCODE DNase I hypersensitivity data. Differentially methylated CpGs between ER+ and ER− breast cancer samples were identified, and the correlation between their methylation levels and expression of their associated genes were calculated. We focused on addressing the question of whether genomic features of CpGs (bound by ER alpha or other TFs, or located in DNase I hypersensitive sites) impacted their differential methylation and their correlation with gene expression [23]

several key biological questions related to transcriptional regulation and DNA methylation. First, we aim to decipher the causal relationships between ERα binding and DNA methylation. We also extend this analysis to identify the causal relationships between DNA methylation and other TFs. Second, we want to address the extent to which TF–DNA methylation relationships could be generalized across all TF–DNA methylation associations (i.e., is DNA methylation correlated or anticorrelated with TF binding). Third, we want to discover whether DNA methylation signals are stronger and more informative in DHS regions of the genome. Lastly, we strove to develop a sense of how CpGs that are correlated with transcriptional activity are spatially distributed across all genes and in what gene elements do they occur most frequently.

To address these questions, gene expression and DNA methylation data for 222 breast cancer patients were downloaded from TCGA. Next, all ENCODE ChIP-seq data containing TF binding site information identified by the peak-calling algorithm PeakSeq [19] were downloaded. Finally, we acquired all ENCODE DHS data from DNase-seq generated in T47d cell lines. All ENCODE data were downloaded from the UCSC Genome browser. Our strategy utilized two popular statistical methods: differential methylation analysis and Spearman's rank correlation. We show that integrating these data sources allows for the identification of each transcription factor's relationship with CpG methylation in the context of its binding activity.

2.2 Correlation Between CpG Methylation and ESR1 Expression

According to the passive DNA methylation model, if DNA methylation passively "fills in" the DNA regions that are not protected by TFs, we would expect to see an inverse correlation between CpG methylation within TF binding sites and TF abundance. Thus, we investigated the relationship between CpG methylation in ERα binding sites and ERα activity in breast cancer samples. Specifically, we correlated the β-values of each of the ~450,000 CpGs interrogated by Illumina's HumanMethylation450K arrays (used by TCGA) with *ESR1* expression (gene encoding ERα) across 222 patient samples. This yielded an average Spearman's correlation coefficient (SCC) of −0.056. This result suggests that methylation of the majority of CpGs have no statistically significant association with ERα activity. Thus, to identify CpGs whose methylation levels do in fact exhibit functional relevance, we repeat the analysis considering only CpGs located within ERα binding sites. To accomplish this, ENCODE ChIP-seq data for ERα was integrated into the analysis allowing CpGs to be stratified into functional and nonfunctional subsets depending whether they fall within or out of ERα binding peaks, respectively (Fig. 1). Once this integration procedure was implemented, the average SCC of functional CpG methylation with *ESR1* expression of CpGs rose to −0.20 (Fig. 2a). Conversely, the average SCC of nonfunctional CpG methylation with *ESR1* expression was −0.083 (Fig. 2a). Furthermore, the percentage of functional CpGs yielding SCC < −0.4 is approximately 30 % (Fig. 2b). As mentioned previously, ER-based

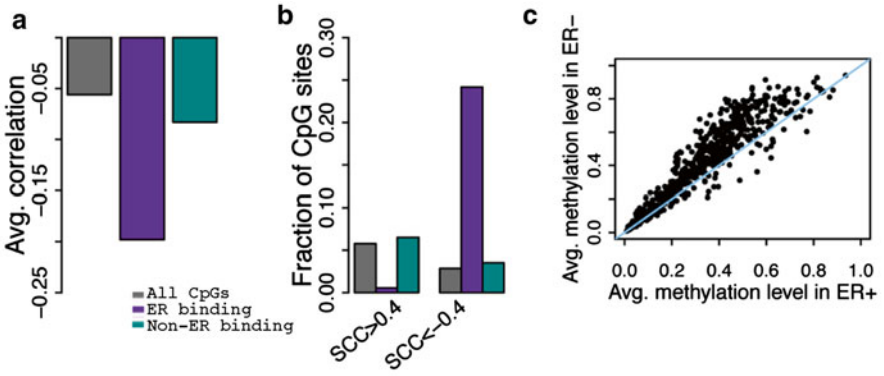


Fig. 2 Correlation between CpG methylation and ESR1 expression levels. (**a**) The methylation levels of CpGs located within ER binding peaks exhibit much more extreme negative correlations with ESR1 expression. (**b**) Comparison of the fraction of CpGs highly correlated with ESR1 expression considering all CpGs and CpGs within and out of ER binding peaks. (**c**) CpGs in ER− samples have higher mean methylation values than those in ER+ samples. *SCC* Spearman correlation coefficient [23]

stratification of breast cancer is standard protocol in the clinic so we analyzed CpG methylation within these classified samples and determined that DNA methylation levels tend to be higher in ER− breast cancer compared to ER+ breast cancer, presumably due to the lack of ERα activity (Fig. 2c).

2.3 Distribution of CpGs with Differential Methylation Levels Between ER+ and ER− Breast Samples

In this study, we also aimed to characterize where and how frequently differentially methylated CpGs between ER+ and ER− are located. Doing so would give us an idea about what genomic regions harbor informative CpGs and thus allow us to infer transcriptional activity in those regions. To approach this analysis, we stratified breast cancer samples into ER+ and ER− samples to identify differentially methylated CpGs and determine how they are spatially distributed. After identifying differentially methylated CpGs, using $P<0.001$ as a cutoff, we then mapped each differentially methylated CpG relative to the transcription start site (TSS) of its associated gene (included in TCGA dataset). We then calculated the fraction of differentially methylated CpGs to the number of CpGs present at that specific genomic location across all representative genes. We made the unexpected discovery that there is actually a higher fraction of differentially methylated CpGs located in regions distal to gene TSSs (Fig. 3b). This suggests the effect of DNA methylation on gene regulation may have a more pronounced effect when it occurs in locations distal to the gene TSS. Additionally, these results show that there is an enrichment of CpGs proximal to gene TSSs (Fig. 3a). In summary, these results indicate that the functional relevance of CpG methylation is dependent on genomic context (i.e., location).

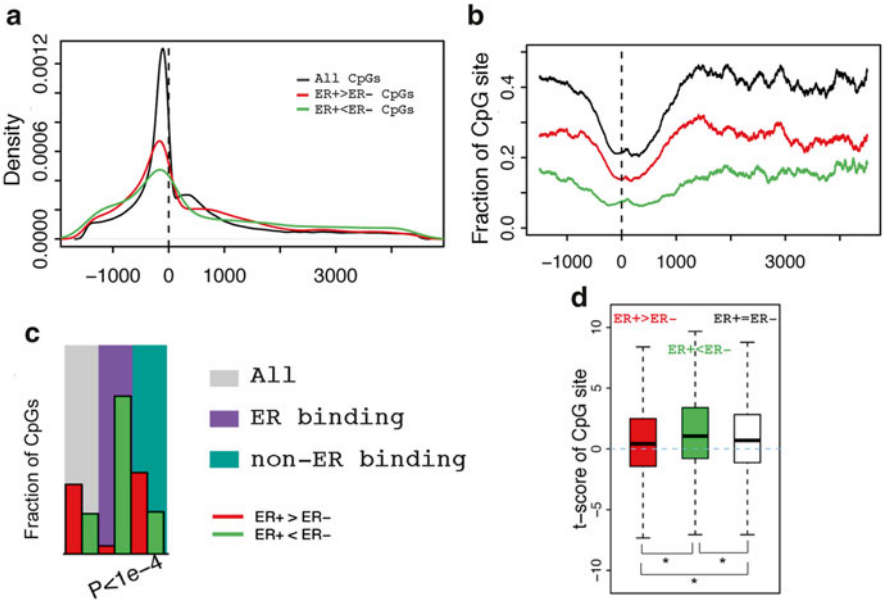


Fig. 3 Distribution of CpGs with differential methylation levels between ER+ and ER− breast samples. (**a**) Distribution of differentially methylated CpGs between ER+ and ER− samples ($P < 1e-6$). (**b**) Fraction of CpGs with significant differential methylation levels at different positions. Fraction corresponds to the number of significant CpGs to the total number of CpGs in a DNA window. Differentially methylated CpGs in ER+ (*red*) and in ER− (*green*) samples are examined separately. (**c**) The fraction of differentially methylated CpGs between ER+ and ER− samples. CpGs in ER binding regions have higher methylation levels in ER− samples, while CpGs not in ER binding regions tend to have higher methylation levels in ER+ samples. (**d**) Distribution of *t*-scores (ER+ vs. ER−) of methylation levels for CpGs. Genes were categorized into three groups based on their relative mRNA expression levels in ER+ versus ER− samples: ER+>ER− (up-regulated in ER+, *red*), ER+<ER− (down-regulated in ER+, *green*) and ER+=ER− (nonsignificant differential expression between ER+ and ER−, *white*). Distributions of CpGs associated with the three gene classes are shown separately [23]

2.4 *Impact of ERα Binding on DNA Methylation*

In order to understand the effect of ERα binding on DNA methylation, we tested whether or not CpGs within ERα binding sites tended to be more differentially methylated between ER+ and ER− samples. TCGA datasets provide immunohistochemical subtyping of samples, which allows us to conduct a high-resolution analysis of ERα binding since its activity can essentially be controlled for by stratifying patient samples on ER status. Thus, we first identified differentially methylated CpGs between ER+ and ER− CpGs and further classified them into two CpG sets. The first set consisted of CpGs that were located within ERα binding sites and the second set consisted of CpGs that were located out of ERα binding sites. The rationale behind this analysis was to determine if ERα activity had an effect on DNA

methylation near its binding sites. We determined that 31 % of CpGs located in ERα binding sites exhibited lower methylation levels in ER+ samples, whereas only 1.1 % of CpGs showed higher methylation levels (Fig. 3c). These results suggest that ERα activity causes hypomethylation within its binding sites. The novelty of this analysis lies in the fact that the effect of ERα binding on DNA methylation was determined purely by data integration and in silico calculation. In addition, we also aimed to understand the relationship between DNA methylation and gene expression. Our first approach was to first identify all differentially expressed genes between ER+ and ER− samples. We then further categorized these differentially expressed genes into up-regulated and down-regulated, and defined the remaining genes as nondifferentially expressed (Fig. 3d). In other words, there will be genes whose expression will increase, decrease, or remain constant upon loss of ERα. Once we established the category of these genes, we calculated the change in methylation levels of CpGs that are associated or located vicinal to these genes upon ERα loss (Fig. 3d). Our results show that CpGs in up-regulated genes had a tendency to decrease in methylation upon loss of ERα and conversely, CpGs in down-regulated genes tended to increase in methylation (Fig. 3d). This suggests that DNA methylation, in most cases, is associated with gene silencing or decrease in gene expression. After showing the relationships between DNA methylation, transcription factor binding activity, and gene expression in our breast cancer samples, we aimed to characterize the local effects of transcription factor binding. From our previous analysis of differential methylation, we calculated *t*-scores for all CpGs which is a statistical metric indicating the magnitude of change in methylation intensity for each CpG between ER+ and ER− samples. If these *t*-scores are plotted as a function of genomic coordinate, it is clear that *t*-scores tend to be lower near the center of ERα binding locations (Fig. 4b). These results are striking because they show a high-resolution DNA methylation "footprint" of ERα binding. To be clear, these footprints vary across different TFs. For example, in the case of SUZ12, the *t*-scores of CpGs tend to increase near the center of SUZ12 binding sites (Fig. 4c). From these observations, we were able to identify the local effects of transcription factor binding on DNA methylation for a variety of transcription factors.

2.5 ERα Is Not the Only Transcription Factor That Leaves a DNA Methylation "Footprint"

We previously explored the impact of ERα binding on DNA methylation in ER+ and ER− patient samples. However, stratifying patient samples into ER+ and ER− subgroups also allows us to identify the DNA methylation footprint of other transcription factors, especially those whose activity is highly correlated with that of ERα. In addition to ERα, we also identified FOXA1 and GATA3 binding sites to be enriched in hypomethylated CpGs. In other words, CpG methylation levels in FOXA1 and GATA3 binding sites are significantly lower in ER+ samples compared to ER− samples (Fig. 4a). This is in accordance with previous experimental

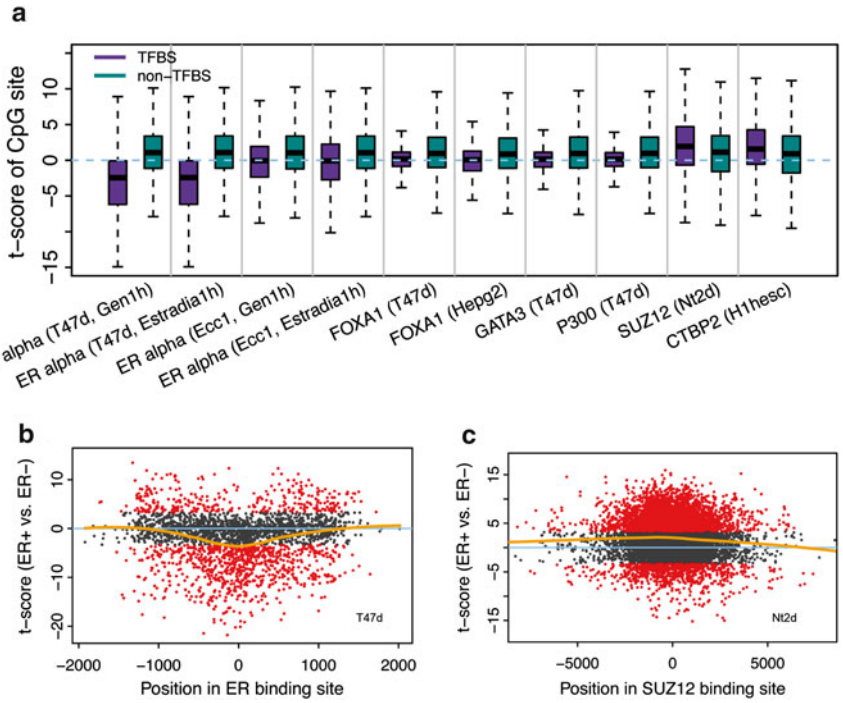


Fig. 4 Relationship between differential methylation of CpGs and TF binding. (**a**) *t*-score distribution of CpG methylation levels between ER+ and ER− samples. Binding of some TFs is associated with reduced CpG methylation, while binding of others (SUZ12 and CTBP2) is associated with increased methylation levels. (**b**) CpGs proximal to ER binding center exhibit lower methylation levels in ER+. (**c**) CpGs proximal to SUZ12 binding center have higher methylation levels in ER+ (*t*-scores reflect ER+ vs. ER− comparison) [23]

literature reporting that the two transcription factors function upstream of ERα. The fact that FOXA1 and GATA3 were identified to contain significantly more hypomethylated CpGs is a testament to the power and versatility of in silico analyses of large-scale integrative analysis of genomic data. One essential concept when dissecting transcription factor–DNA methylation relationships is that each transcription factor and its relationship with CpG methylation must be considered case by case. For instance, SUZ12 binding sites are hypermethylated in ER+ than in ER− samples, suggesting that binding of SUZ12 increases CpG methylation (Fig. 4c).

2.6 Correlation Between DNA Methylation and Gene Expression

Because the regulatory functions of DNA methylation are of considerable interest, understanding its association with gene expression is a major focus in the field. It has been postulated that DNA methylation can silence, promote, or have no effect

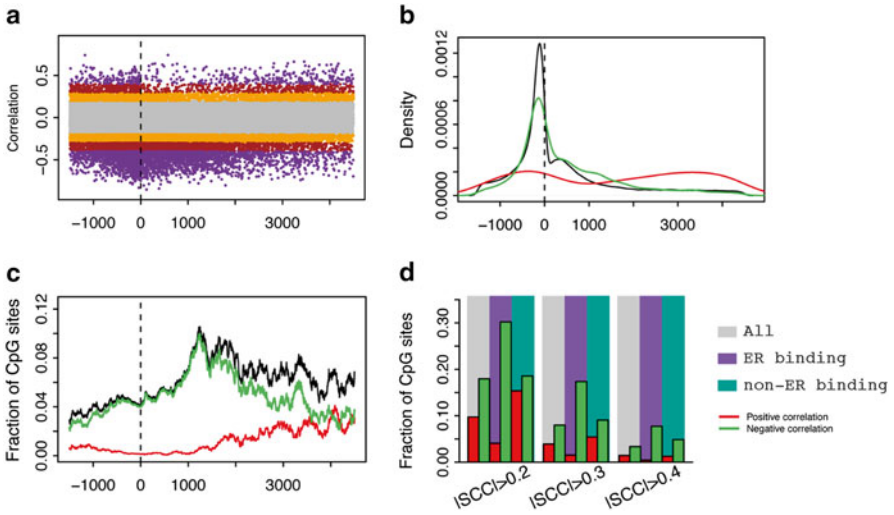


Fig. 5 Correlation of CpG methylation level with expression level of the associated genes. (**a**) Correlation of CpG methylation with gene expression as a function of CpG position relative to transcription start site (from −1,500 upstream to 4,500 downstream of TSS). (**b**) Distribution of CpGs that exhibit strong correlations in their methylation levels with expression of associated genes across genomic region. Positive correlation (*red*, $r>0.4$) and negative correlation (*green*, $r<-0.4$) are examined separately. (**c**) Fraction of CpGs strongly correlated with expression of the associated genes at different positions. (**d**) CpG methylation in ER binding regions exhibit larger negative correlation with expression of their associated genes [23]

on gene expression depending on genomic context. Hence, we correlated DNA methylation levels of each CpG with the expression level of its associated gene across samples. Our results show that there are many more CpGs that are anticorrelated, rather than correlated, with gene expression (Fig. 5a). We also analyzed the spatial distribution of correlated CpGs (those with $r>0.4$ or <-0.4) and discovered that anticorrelated CpGs exhibited a peak in DNA regions upstream from the TSS, whereas correlated CpGs exhibited peaks in the promoter region and in the gene body (Fig. 5b). However, the distribution of CpGs across genes is not uniform but exhibits greater density near gene promoters. Therefore, if the fraction of high correlation CpGs ($r>0.4$ or $r<-0.4$) is calculated relative to the overall number of CpGs at the site, then there exists a peak approximately 1,000 bp downstream of the TSS with a high percentage of strong correlation CpGs (Fig. 5c). This striking result suggests that CpGs distal to gene TSSs may possess more important functional roles than previously thought. Furthermore, we correlated CpG methylation levels with gene expression for CpGs that lie within ERα binding sites and show that the fraction of anticorrelated CpGs is much larger in these sites than in regions located outside of ERα binding sites (Fig. 5d). Again, this suggests that restricting our analysis to transcription factor binding sites can drastically increase the informativeness of DNA methylation patterning.

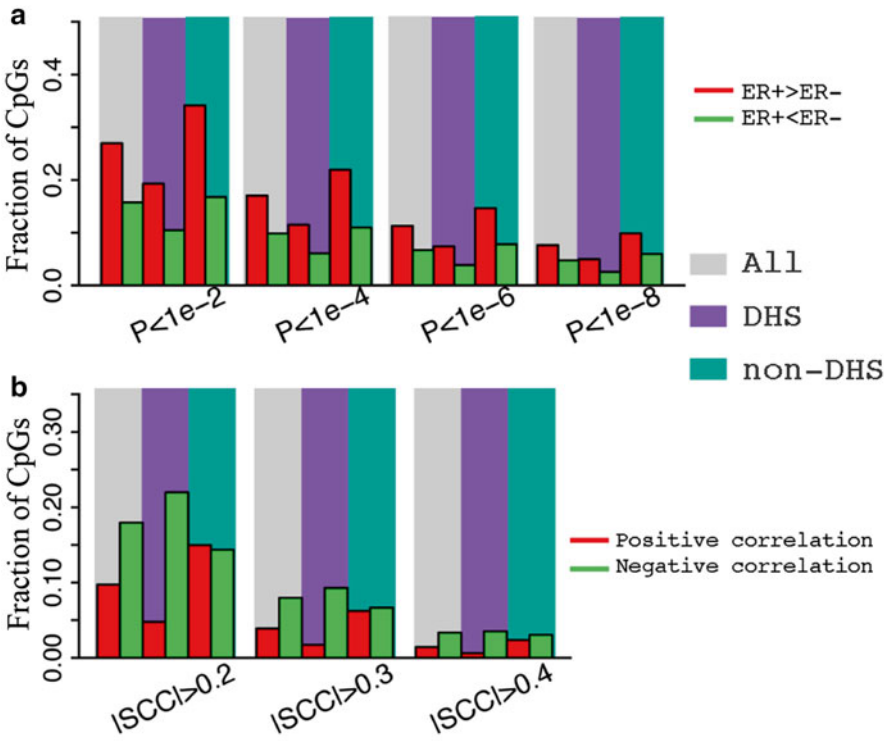


Fig. 6 Comparison of CpGs in and not in DNase hypersensitive sites. (**a**) CpGs in DHS and non-DHS exhibit no significant difference in differential methylation between ER+ and ER− breast cancer samples. (**b**) CpG methylation in DHS exhibits larger negative correlation with the expression of their associated genes [23]

2.7 CpGs in DNase Hypersensitive Sites Are More Informative

DNase hypersensitive regions are sections of open chromatin that are amenable to transcription factor binding due to its physically open state. Based on this, we hypothesized that DNA methylation patterns identified within these regions may be more informative and perhaps even functional in terms of their regulatory activity. Thus, we integrated ENCODE DNase I hypersensitivity data to identify differentially methylated CpGs between ER+ and ER− breast cancers and correlated CpGs with gene expression levels in these regions. We find that there is a depletion of differentially methylated CpGs in DNase hypersensitivity regions (Fig. 6a). From this, it can be deduced that there is increased transcription factor binding events that occur in these regions that leads to hypomethylation, which suggests that these DNase hypersensitivity regions are enriched in regulatory elements (i.e., promoters and enhancers). In addition, we also conducted correlation analysis with CpGs and expression of their associated genes within and out of DNase hypersensitivity

regions and reported that highly anticorrelated CpGs are enriched in DNase hypersensitive regions (Fig. 6b). Conversely, highly correlated CpGs are depleted in DNase hypersensitivity regions (Fig. 6b).

2.8 Integrative Analysis of Biological Datasets Must Take into Account the Underlying Biology

Despite the power of integrative analysis, extracting information from different sources does come with limitations. One essential consideration that must be taken before integrating biological datasets is deciphering the biological overlap between the two data sources. To provide a specific example, take into consideration ChIP-seq data; in our analysis, we integrated ChIP-seq data for ERα in both T47d and Ecc1 cell lines. Interestingly, the difference in methylation between ER+ and ER− within ERα binding sets was much more pronounced when using ERα ChIP-seq data derived from T47d cell lines (Fig. 4a). This is due to the fact that T47d cell lines are derived from breast epithelium whereas Ecc1 is derived from the epithelium of endometrium tissue. Since our DNA methylation datasets were derived from breast cancer patient samples, using a breast epithelium cell line provides much more precise and accurate results due to greater biological similarity. This example reflects the importance of considering tissue specificity when integrating datasets. Another consideration that must be taken into account is the treatment conditions under which ChIP-seq datasets were derived (Fig. 4a). In our dataset, the T47d cell line was also treated under two different hormone conditions before ChIP-seq experiments were performed. The first was without 17β-estradiol and the second was with 17β-estradiol. Since estradiol is an activating ligand of ERα, the hormone must be present for ERα to be active. Therefore, ChIP-seq experiments derived from T47d cell lines treated with estradiol provided the best results in our analysis. As more data becomes available, determining the biological overlap between datasets will become a major challenge. Since data is generated from a variety of technological platforms, in different labs, in different disease contexts, different conditions, etc., it is integral to a priori determine which datasets are appropriate for integration.

3 Discussion

3.1 Correlation Between DNA Methylation and Gene Expression

The association between DNA methylation and gene expression is highly complex, nuanced, and varies from case to case. In general, it has been suggested that DNA methylation in promoter regions represses gene expression [20]. Overall, there is a negative correlation between the expression of genes and the methylation levels in their promoter proximal regions. For example, we calculated the correlations between gene expression and promoter methylation (from TSS to 200 bp upstream)

across all transcribed genes in hESC and IMR90 cells using ENCODE data. In hESC, we observed a correlation coefficient of $r=-0.37$ and in IMR90 $r=-0.22$. Consistently, a weak negative correlation ($r=-0.24$) between gene expression and promoter methylation has been reported in H1 cell lines [21]. More recent studies have shown that the across individual methylation-gene expression associations can be either positive or negative, even for DNA methylation sites in promoter regions [22, 23]. Overall, the correlation between DNA promoter methylation and gene expression is complex and nonlinear. For instance, gene body methylation was observed to positively correlate with gene expression in some cell types [24, 25], but not in others [26]. Previous reports have proposed different possible mechanisms: the reduction of efficacy of transcription elongation [27, 28], regulation of alternative promoters [29], and the blocking of transcriptional repressors [28, 30–32]. In comparison with gene promoter regions, gene bodies have a relatively few number of CpG sites, but extremely high methylation levels. Quantitative models between DNA methylation and gene expression reveal that DNA methylation can partially determine gene expression, and only extremely high- and lowly expressed genes can be predicted well [33]. This study also found that gene body methylation is a stronger indicator of expression level than promoter methylation. Promoter methylation appears to affect a relatively small set of genes with extreme methylation levels by means of on/off switches, while the effect of gene body methylation on reducing transcriptional efficiency may operate under a more general mechanism; thus affecting more genes This provides a plausible explanation for the stronger modeling power of gene body methylation features. To reiterate, it should be noted that despite the correlation between gene expression and DNA methylation, it remains unclear whether DNA methylation is the cause or the consequence of altered gene expression.

3.2 DNA Methylation Within Distal Genetic Elements

One of the key findings that arose from our study was that there was a high fraction of CpGs that were highly correlated with expression of their associated genes, enriched 1,000 bp downstream from gene TSSs. These results suggest that CpGs across gene regions other than TSS may play a profound role in gene expression. However, our study only analyzed DNA methylation in 2 kb bins and did not explore the potential effects of DNA methylation in distal elements such as enhancers. Since these genetic elements are essential players in transcriptional regulation, much insight would be provided if information from distal element methylation were to be incorporated. Indeed studies have suggested regulatory roles for CpGs in gene bodies, intergenic regions, and in distal elements such as enhancers [26, 29, 34–36]. Ball et al. applied bisulfite padlock probes (BSPPs) and methyl-sensitive cut counting (MSCC) to profile methylation levels of genomic sites and discovered that gene body methylation was correlated with gene expression [24]. In another large-scale analysis, Aran et al. applied reduced representation bisulfite sequencing and Infinium HumanMethylation 450 BeadChip arrays to study the relationships

between DNA methylation in distal elements and gene expression levels across 58 human cancer cell types [34]. They analyzed 1911 distal methylation sites that showed a relationship with the expression levels of 486 genes between normal and cancer cell lines [34]. They determined that genes associated with hypermethylated enhancers exhibited decreased expression and conversely, genes associated with hypomethylated enhancers exhibited increased expression [34]. This comprehensive analysis revealed that enhancer methylation has an effect on gene expression. Overall, these results imply that functional DNA methylation is not restricted to a single genomic region but is a ubiquitous regulatory player; thus, there is still much to explore with respect to transcriptional regulation and DNA methylation.

3.3 Translational Potential of DNA Methylation Analysis

Recently, several attempts have been made to identify DNA methylation markers that can be used as biomarkers in the clinic to predict patient prognosis. For instance, a recent study Anjum et al. identified a BRCA1-mutation-associated DNA methylation signature that was predictive of breast cancer incidence and survival [37]. Despite these preliminary results, the potential of DNA methylation markers is still not fully realized. Our study provides insight into the molecular activity that results in these "signatures." By showing that transcription factors are intimately related to DNA methylation patterning in its binding sites, we demonstrate the potential use of these signatures to infer transcriptional activity in various disease contexts (i.e., ER+ and ER− breast cancer). As a result, we can identify the regulatory architecture underlying a specific disease and use that information to predict prognosis or treatment. For example, in our study we show that, upon loss of ERα, the binding sites of transcription factors such as ERα, FOXA1, and GATA3 become hypermethylated. Hypothetically, DNA methylation signatures that reflect the status of each transcription factor could be used to understand the drivers of disease. As a result, it is possible to develop drugs that target these regulators that exhibit altered activity in a disease. In addition, if a protein is known to contribute to oncogenesis, it may be possible to assess whether this protein's DNA methylation "footprint" is present in a disease. This provides an interesting new approach to biomarker development for patient prognosis. In conclusion, understanding transcription factor–DNA methylation interactions provides high-resolution insight into the regulatory crosstalk that occurs in cancer and provides numerous avenues for biomarker and drug design.

3.4 Computational Analysis in Modern Oncogenomics

In our study, we utilized data that was available in the public domain and integrated them to derive biological meaning. By properly integrating these datasets, we carried out a large-scale analysis that provided a high-resolution view of the potential

effects transcription factor binding may have on DNA methylation. In addition, computational analysis allowed us to discern the physical distribution of functional CpGs across all genes. Additionally, carrying out analysis in TF binding sites and DHS regions we were able to increase the resolution of our study by pinpointing CpGs that occur in functional regions of the genome. As more genomic data is generated, large-scale data integration and analysis will become mainstay procedures in biomedical research.

4 Methods

4.1 Datasets

Gene expression and DNA methylation data for breast cancer patients were downloaded from TCGA's data portal (accessed 30 May 2013). Gene expression data were derived from two-channel Agilent microarrays. Methylation levels of CpGs were measured with the Illumina HumanMethylation450 microarray technology. CpG intensities were outputted as β values which range from 0 (completely unmethylated) to 1 (completely methylated).

Genome-wide transcription factor binding data were generated from ChIP-seq experiments as part of the ENCODE project. Binding peaks for TFs are available from the UCSC Genome browser at http://genome.ucsc.edu/ENCODE/downloads.html. Binding peaks were identified using the PeakSeq software. Data from T47d and MCF7 breast epithelial cell lines were used in our analysis.

DNase hypersensitivity data were generated by the ENCODE project based on DNase-seq experiments and were downloaded from the UCSC genome browser. The data provide the genomic coordinates of DNA regions sensitive to DNase I treatment. Data from T47d and MCF7 cell lines were used in our analysis.

4.2 Differential DNA Methylation Between ER+ and ER− Breast Cancer Samples

TCGA provides methylation data for 485,577 CpGs in 630 ER+ and 187 ER− breast cancer samples. The majority of interrogated CpGs can be assigned to a particular gene based on its location: in the transcribed region or vicinal to the transcription start site of a gene. To identify differentially methylated CpGs, the β values of each CpG were compared between ER+ and ER− breast cancer samples using the Student's t-test. A significance cutoff of $P < 0.001$ was used to determine whether a CpG was differentially methylated. CpGs that exhibited a t-statistic greater than 0 (ER+ > ER−) were categorized into a hypermethylated set. Similarly, CpGs with a t-statistic less than 0 were labeled as hypomethylated. Different significance cutoffs were used and the results were stable.

4.3 Differential Gene Expression Between ER+ and ER− Breast Cancer Samples

To identify differentially expressed genes, we utilized the microarray gene expression profiles for 519 breast cancer samples. Samples were divided into ER+ (401) and ER− (118) samples, and the expression levels of genes between ER+ and ER− were compared using Student's *t*-test. Genes were considered differentially expressed if they yielded $P < 0.001$. Differentially expressed genes were categorized into up-regulated in ER+, down-regulated in ER−, and nondifferentially expressed based on their *t*-statistics and *P*-values.

4.4 Relating CpGs with ER Binding, TF Binding, and DNase I Hypersensitive Sites

Given the genomic coordinates of ERα binding sites in a cell line, we can determine which CpGs fall within these regions. We defined ER binding CpGs as those falling directly within an ERα binding peak and non-ER binding CpGs as those that are not located in any ERα binding peaks but were located in genes that contain these peaks. By considering only CpGs within peak-containing CpGs, we can restrict our analysis to a local genomic region. This procedure was used for all other TFs. When conducting analysis using DNase hypersensitivity data, we utilized the aforementioned procedure but instead limited it to DNase hypersensitive or non-DNase hypersensitive sites.

4.5 Correlation of DNA Methylation with Gene Expression

Both gene expression and DNA methylation data were available for 222 of the TCGA breast cancer samples. We applied this data to study the correlation of DNA methylation with gene expression. For each CpG, we correlated its β values with the mRNA expression values of its associated gene using Spearman's rank correlation. Fisher's transformation was used to determine significance of the outputted correlation coefficient.

References

1. Hanahan D, Weinberg RA. Hallmarks of cancer: the next generation. Cell. 2011;144:646–74.
2. Jan M, Snyder TM, Corces-Zimmerman MR, Vyas P, Weissman IL, Quake SR, Majeti R. Clonal evolution of preleukemic hematopoietic stem cells precedes human acute myeloid leukemia. Sci Transl Med. 2012;4:149ra118.

3. Burrell RA, McGranahan N, Bartek J, Swanton C. The causes and consequences of genetic heterogeneity in cancer evolution. Nature. 2013;501:338–45.
4. Schnitt SJ. Classification and prognosis of invasive breast cancer: from morphology to molecular taxonomy. Mod Pathol. 2010;23 Suppl 2:S60–4.
5. Easwaran H, Tsai HC, Baylin SB. Cancer epigenetics: tumor heterogeneity, plasticity of stem-like states, and drug resistance. Mol Cell. 2014;54:716–27.
6. Hansen K, Timp W, Bravo HC, Sabunciyan S, Langmead B, McDonald O, Wen B, Wu H, Liu Y, Diep D, et al. Increased methylation variation in epigenetic domains across cancer types. Nat Genet. 2011;43:768–75.
7. Medvedeva YA, Khamis AM, Kulakovskiy IV, Ba-Alawi W, Bhuyan MS, Kawaji H, Lassmann T, Harbers M, Forrest AR, Bajic VB, Consortium F. Effects of cytosine methylation on transcription factor binding sites. BMC Genomics. 2014;15:119.
8. Thurman R, Rynes E, Humbert R, Vierstra J, Maurano M, Haugen E, Sheffield N, Stergachis A, Wang H, Vernot B, et al. The accessible chromatin landscape of the human genome. Nature. 2012;489:75–82.
9. Vire E, Brenner C, Deplus R, Blanchon L, Fraga M, Didelot C, Morey L, Van Eynde A, Bernard D, Vanderwinden JM, et al. The Polycomb group protein EZH2 directly controls DNA methylation. Nature. 2006;439:871–4.
10. Nabilsi NH, Broaddus RR, Loose DS. DNA methylation inhibits p53-mediated survivin repression. Oncogene. 2009;28:2046–50.
11. Prendergast G, Lawe D, Ziff E. Association of Myn, the murine homolog of max, with c-Myc stimulates methylation-sensitive DNA binding and ras cotransformation. Cell. 1991;65:395–407.
12. Comb M, Goodman H. CpG methylation inhibits proenkephalin gene expression and binding of the transcription factor AP-2. Nucleic Acids Res. 1990;18:3975–82.
13. Hu S, Wan J, Su Y, Song Q, Zeng Y, Nguyen HN, Shin J, Cox E, Rho HS, Woodard C, et al. DNA methylation presents distinct binding sites for human transcription factors. Elife. 2013;2:e00726.
14. Hendrich B, Bird A. Identification and characterization of a family of mammalian methyl-CpG binding proteins. Mol Cell Biol. 1998;18:6538–47.
15. Fuks F, Hurd PJ, Wolf D, Nan X, Bird AP, Kouzarides T. The methyl-CpG-binding protein MeCP2 links DNA methylation to histone methylation. J Biol Chem. 2003;278:4035–40.
16. Laird PW. The power and the promise of DNA methylation markers. Nat Rev Cancer. 2003;3:253–66.
17. Cancer Genome Atlas Research Network, Weinstein JN, Collisson EA, Mills GB, Shaw KR, Ozenberger BA, Ellrott K, Shmulevich I, Sander C, Stuart JM. The Cancer Genome Atlas Pan-Cancer analysis project. Nat Genet. 2013;45:1113–20.
18. Consortium EP, Bernstein BE, Birney E, Dunham I, Green ED, Gunter C, Snyder M. An integrated encyclopedia of DNA elements in the human genome. Nature. 2012;489:57–74.
19. Rozowsky J, Euskirchen G, Auerbach R, Zhang Z, Gibson T, Bjornson R, Carriero N, Snyder M, Gerstein M. PeakSeq enables systematic scoring of ChIP-seq experiments relative to controls. Nat Biotechnol. 2009;27:66–75.
20. Jones PA. Functions of DNA methylation: islands, start sites, gene bodies and beyond. Nat Rev Genet. 2012;13:484–92.
21. Du X, Han L, Guo AY, Zhao Z. Features of methylation and gene expression in the promoter-associated CpG islands using human methylome data. Comp Funct Genomics. 2012;2012: 598987.
22. Gutierrez-Arcelus M, Lappalainen T, Montgomery SB, Buil A, Ongen H, Yurovsky A, Bryois J, Giger T, Romano L, Planchon A, et al. Passive and active DNA methylation and the interplay with genetic variation in gene regulation. Elife (Cambridge). 2013;2:e00523.
23. Ung M, Ma X, Johnson KC, Christensen BC, Cheng C. Effect of estrogen receptor alpha binding on functional DNA methylation in breast cancer. Epigenetics. 2014;9:523–32.
24. Ball MP, Li JB, Gao Y, Lee JH, LeProust EM, Park IH, Xie B, Daley GQ, Church GM. Targeted and genome-scale strategies reveal gene-body methylation signatures in human cells. Nat Biotechnol. 2009;27:361–8.

25. Rauch TA, Wu X, Zhong X, Riggs AD, Pfeifer GP. A human B cell methylome at 100-base pair resolution. Proc Natl Acad Sci U S A. 2009;106:671–8.
26. Lister R, Pelizzola M, Dowen RH, Hawkins RD, Hon G, Tonti-Filippini J, Nery JR, Lee L, Ye Z, Ngo QM, et al. Human DNA methylomes at base resolution show widespread epigenomic differences. Nature. 2009;462:315–22.
27. Rountree MR, Selker EU. DNA methylation inhibits elongation but not initiation of transcription in Neurospora crassa. Genes Dev. 1997;11:2383–95.
28. Lorincz MC, Dickerson DR, Schmitt M, Groudine M. Intragenic DNA methylation alters chromatin structure and elongation efficiency in mammalian cells. Nat Struct Mol Biol. 2004;11:1068–75.
29. Maunakea AK, Nagarajan RP, Bilenky M, Ballinger TJ, D'Souza C, Fouse SD, Johnson BE, Hong C, Nielsen C, Zhao Y, et al. Conserved role of intragenic DNA methylation in regulating alternative promoters. Nature. 2010;466:253–7.
30. Belfort M, Kass N, Oppenheim A, Katzir N, Oppenheim AB. Repressor and int synthesis of bacteriophage lambda in the E. coli host mutant ER437. Mol Gen Genet. 1977;155:347–9.
31. Bird A. DNA methylation patterns and epigenetic memory. Genes Dev. 2002;16:6–21.
32. Kass SU, Pruss D, Wolffe AP. How does DNA methylation repress transcription? Trends Genet. 1997;13:444–9.
33. Lou S, Lee HM, Qin H, Li JW, Gao Z, Liu X, Chan LL, Lam V, So WY, Wang Y, et al. Whole-genome bisulfite sequencing of multiple individuals reveals complementary roles of promoter and gene body methylation in transcriptional regulation. Genome Biol. 2014;15:408.
34. Aran D, Sabato S, Hellman A. DNA methylation of distal regulatory sites characterizes dysregulation of cancer genes. Genome Biol. 2013;14:R21.
35. Feng S, Cokus SJ, Zhang X, Chen PY, Bostick M, Goll MG, Hetzel J, Jain J, Strauss SH, Halpern ME, et al. Conservation and divergence of methylation patterning in plants and animals. Proc Natl Acad Sci U S A. 2010;107:8689–94.
36. Cokus SJ, Feng S, Zhang X, Chen Z, Merriman B, Haudenschild CD, Pradhan S, Nelson SF, Pellegrini M, Jacobsen SE. Shotgun bisulphite sequencing of the Arabidopsis genome reveals DNA methylation patterning. Nature. 2008;452:215–9.
37. Anjum S, Fourkala EO, Zikan M, Wong A, Gentry-Maharaj A, Jones A, Hardy R, Cibula D, Kuh D, Jacobs IJ, et al. A BRCA1-mutation associated DNA methylation signature in blood cells predicts sporadic breast cancer incidence and survival. Genome Med. 2014;6:47.

Differential Methylation Analysis with Next-Generation Sequencing

Hongyan Xu

Abstract DNA methylation is an important epigenetic modification of DNA sequences, which could potentially affect gene expression and final phenotypes. Abnormal methylation has been discovered in many types of cancers and other human diseases. Detecting differentially methylated loci (DML) and differentially methylated regions (DMRs) is critical in understanding the genetic mechanism of cancer and identifying biomarkers and treatment targets, which could be used for cancer diagnosis, prognosis, prevention, and treatment. Next-generation sequencing (NGS) has been widely used to generate genome-wide methylation data. These data provide unique challenges in the differential methylation analysis at genomic levels. In this paper, we discuss these challenges and some statistical and computational approaches for detecting DML and DMRs for NGS methylation data.

1 Introduction

With the completion of the Human Genome Project, we have a draft of the human genomic sequence. Yet we are far from fully understanding our genome. Genome-wide association study has been a popular approach in recent years toward this goal. Although many disease-related loci have been identified, in total they only explain a small proportion of phenotypic variation [1]. With the availability of whole-genome sequencing data through next-generation sequencing (NGS), many efforts have been put into sequence-based rare genetic variants association [2]. Besides rare genetic variants, epigenetic modifications, such as DNA methylation have been shown to be related to gene expression and therefore can affect phenotypic variations. Abnormal DNA methylation marks have been involved in many human diseases especially cancer [3, 4].

H. Xu, Ph.D. (✉)
Department of Biostatistics and Epidemiology, Medical College of Georgia, Georgia Regents University, Augusta, GA 30912, USA
e-mail: hxu@gru.edu

W. Wu, H. Choudhry (eds.), *Next Generation Sequencing in Cancer Research, Volume 2*, DOI 10.1007/978-3-319-15811-2_14

1.1 DNA Methylation

At molecular level, DNA methylation is the covalent addition of a methyl group to the cytosine nucleotides. This process typically occurs at the 5′ of cytosine in CpG dinucleotides. DNA methylation plays an important role in various cellular processes including X chromosome inactivation, genomic imprinting, and chromosome stability [5–7]. Several DNA methylation modifications involved in developing primordial germ cells and fertilized occytes are inheritable. Ultimately, these methylation modifications can produce stable alterations of gene expression. DNA methylation is also associated with histone modifications and plays a crucial role in the basis of chromatin structure [8, 9].

Aberrant DNA methylations, both hypermethylation (gain of methylation) and hypomethylation (loss of methylation) have been associated with human diseases such as cancer [10]. Hypermethylation within the promoter regions is commonly known to silence certain tumor suppressor genes. In many types of cancer, hypermethylation occurs in genomic regions with a high frequency of CpG sites (CpG Islands) [11, 12]. In contrast, hypomethylation frequently occurs in the early stages of neoplasm and is linked to chromosomal instability and loss of imprinting [13]. Further, methylation has been shown to be prognostic for tumor progression, disease severity, and metastatic potential [14].

1.2 DNA Methylation Profiling with NGS

Next-Generation Sequencing (NGS) is a major platform to extract methylation information from biological systems. NGS has the advantage of huge sequencing capacity, cost-effectiveness, and broad applications, all of which may enhance our knowledge of how genetics affects health and disease. Various sequencing methods based on bisulfite conversion have been applied to determine genomic methylation patterns [15]. Bisulfite Treatment of DNA converts unmethylated cytosines to uracils, but leaves methylated cytosines intact. Therefore, after bisulfite conversion, methylated and unmethylated cytosines can be distinguished by DNA sequencing, with methylated sites appearing as cytosine and unmethylated sites appearing as thymine.

With NGS platforms, DNA methylation measurements are represented by the counts of methylated and unmethylated molecules. The total count of molecules that covers a CpG site is called read depth or coverage. There are considerable variations in coverage for CpG sites across the genome and between individuals. An important difference between methylation data from NGS platforms and the data from methylation array is that the methylation proportion, called β-value, has to be estimated from the methylation counts data and the accuracy of the estimate is affected by read depth.

2 Methods for Detecting Differentially Methylated Loci (DML)

2.1 *Student's* t-*Test*

Student's *t*-test is a simple statistical approach to detect differences in mean values of methylation levels using the estimated methylation proportions from NGS counts data [16, 17]. This approach converts methylation count information to an estimated proportion by taking the ratio of methylation count and read depth at a particular CpG site. Thus, it removes problems associated with the unequal read depth among individuals [18]. However, there are several disadvantages for this approach. First, since the methylation proportions are between 0 and 1 with unknown distributions, the normality assumption of *t*-test may not be valid for the methylation data, especially with small sample size or outliers. Second, the methylation proportion is estimated as the ratio of the count of methylated molecules over the read depth. This approach is different from microarray experiments where the methylation level is directly measured. The variation of estimated methylation proportion is directly affected by the read depth and could easily be affected by factors such as sampling process, library preparation, and batch effect. Another complexity of the methylation study comes from the presence of other potential confounders such as age and sex. If the relationship between methylation level and potential confounders is substantial, Student's *t*-test may lose power or be invalid in practice.

2.2 *Cluster Analysis Approach*

Xu et al. [18] proposed to use the adjusted chi-square test [19] for differential methylation analysis with NGS data. The key principle of this approach is to treat the NGS reads as clusters within each individual and to adjust the methylation levels for clustered observations. Under this approach the overall methylation proportion for each group is estimated as the ratio of the total methylation count over the total read depth within group, and the overall variance of the estimated proportion for each group is estimated as the average of the squared difference of the methylation count and its expected values based on the calculated overall methylation proportion for all the individual within groups. This step ignores the clustering effect within individuals. The clustering effect is adjusted with the cluster counts within individuals by dividing the estimated proportion and its variance for each group by the "design effect." The adjusted proportions and their variances are used to derive a chi-square test statistic with one degree of freedom under the null hypothesis of no differential methylation. This approach takes the differences in reading depth into account so that it decreases the bias from the estimation of methylation proportion. However, this method suffers from potential bias and power loss from not considering the confounders.

2.3 Logistic Regression

Logistic regression is a commonly used approach to detect differential methylation levels with NGS data [20–22]. This method models the odds of methylation at each CpG site through a logistic regression model with group as a predictor:

$$\log\left(\frac{\pi_i}{1-\pi_i}\right) = \beta_0 + \beta_1 X_i$$

where π_i is the methylation probability for individual i and X_i denotes the treatment group indicator for individuals i. The parameter β_0 denotes the log odds of the normal group and β_1 denotes the log odds ratio between normal and disease group. The differential methylation can be decided by testing whether parameter β_1 for the group variable is significantly different from zero. The standard inference procedure for testing these regression coefficients is based on the empirical likelihood function, which eliminates the assumption on the distribution of methylation values. As usual, the Wald test for β_1 is calculated based on the maximum likelihood estimation and its estimated variance. Akalin et al. [23] provided an R package-methylKit for determining differential methylation across genomic regions through logistic regression model. Although the use of logistic regression offers great flexibility in the models that it can fit, the approach did not account for the differences in coverage between samples and the variability of the methylation proportion within a group.

2.4 Bayesian Framework

Bayesian models have been successfully applied in modeling NGS sequencing data, such as ChIP-Seq [24] and RNA-Seq data [25]. Bayesian hierarchical framework offers flexibility in modeling the complex process of generation sequencing counts. BayesPeak used a hidden Markov model (HMM) and modeled the counts with a Gamma-Poisson mixed distribution for ChIP-Seq data [24]. Markov chain Monte Carlo algorithm (MCMC) was performed for posterior samples to detect enriched locations in the genome. Wu et al. [26] proposed two methods for NGS data using methyl-seq approach and reduced representation bisulfite sequencing (RRBS) approach. One method is a maximum likelihood estimation and the other is a Bayesian estimation with Gamma-Poisson mixture distribution model. They demonstrated that the maximum likelihood method yielded biased estimation at extreme methylation levels while Bayesian hierarchical model could adjust this bias flexibly. However, this paper just provided the statistical approach for parameters estimation without formal statistical approach for Bayesian hypothesis testing to detect differential methylation.

A number of empirical Bayesian models have been developed to analyze differential DNA methylation proportion using microarrays. Down et al. developed a Bayesian tool for methylation analysis (Batman) with oligonucleotide arrays

(MeDIP-chip) data [27]. Batman modeled a group of CpGs in 50- or 100-bp windows assuming same methylation state for the CpG sites and then used MCMC to derive posterior samples.

McCallum et al. [53] provided a Bayesian framework with a beta-binomial hierarchical model to detect differentially methylation loci between normal group and disease group. The maximum likelihood estimated priors were chosen and the posterior distribution of the methylation proportion was derived using MCMC from the combination of the likelihood function and the estimated priors. To test the hypothesis, H_0: $\beta_0 = \beta_1$ versus H_a: $\beta_0 \neq \beta_1$, where β_0 and β_1 are the methylation proportions in the two groups respectively, they used the posterior log odds Δ. The null hypothesis is rejected if $|\Delta| > \sigma_\alpha$, where σ_α is the cutoff value at significance level α. They demonstrated that this Bayesian framework approach is more powerful to handle small sample size than Fisher's exact test at many sites across the genome and has a well control of false discovery rate (FDR).

Hardcastle and Kelly [28] developed an empirical Bayesian framework under the assumption that the data follows a Negative Binomial (NB) distribution to model paired data from high-throughput sequencing platforms. The same authors [29] developed another Bayesian framework based on the Beta-Binomial distribution. Comparison between Bayesian framework and generalized linear modeling approach indicates that Bayesian framework offers better performance on both simulated data and real data. Both Bayesian models are implemented in the baySeq R/Bioconductor package.

Most of these Bayesian methods rely on MCMC. MCMC algorithm is a popular method to estimate the Bayesian posterior densities. This algorithm offers a sequence of samples whose stationary distribution is the target posterior distribution. Normally, the quality of the Markov chain improves as the number of iteration increases. However, it is difficult to determine how many iterations are needed for a Markov chain to converge to its stationary distribution. It is important to make sure that all parameters are converged in order to get accurate MCMC posterior samples. In a hierarchical model, we have high-dimensional posterior parameters. It is not easy to solve the convergence problem even if we choose very large number of "burn-in" iterations [30]. Currently, NGS offers methylation measurement at over 2 million CpG sites. It is impossible to check the convergence for each CpG site and the application of MCMC methods in methylation analysis with over 2 million CpG sites imposes tremendously computational burden. A more efficient method is needed in this regard.

2.5 Nonparametric Methods

Nonparametric statistics are often used when certain distribution assumptions about the underlying population are unknown or questionable. These methods do not require a distribution assumption, thus they can be used when the data may deviate from the assumed distribution in parametric analysis, even for nominal or ordinal data. The lack of distribution assumption makes nonparametric methods very flexible.

In most cases, they are easy to compute and understand. They use less information than the parametric tests, such as sign test, which is based on the ranks of the observations. However, in the case where a distributional assumption is reasonable, nonparametric methods are less powerful than their parametric counterparts.

Chen et al. [31] proposed an age-adjusted nonparametric method for detecting differential methylation for case–control designs. To adjust the age effect, this method divided the samples into several age groups and then combined the *p*-values from Student's *t*-test in each group to generate a new test statistic. This new statistic was assumed to follow a chi-square distribution. Huang et al. [32] adjusted Chen's method with *p*-value calculation based on Neuhaeuser's one-sided test [33] for each age group. Both methods have been demonstrated to have improved detection power and a good control of the type I error for detecting differential methylation for case–control design.

3 Identifying Differentially Methylated Regions (DMRs)

When performing any test for differential methylation between two groups of samples, the null hypothesis is that the mean methylation level as a function of position is no different for two groups or that it is independent of some continuous covariate. This test can be carried out by either performing an individual test for each site and then using a correction for multiple comparisons, or it can be done via smoothing in conjunction with multiple testing correction. Site-by-site tests can offer better resolution but will lead to correlated *p*-values and can miss functional differences that are large over a region but small at any individual site. Functional tests from smoothed data can better detect differences over a region that site-by-site tests may not be able to detect. The downside is deciding how to define regions and determining an appropriate functional test to detect different methylation profiles.

3.1 t-*Test Like Approach*

For microarray or next-generation sequencing methylation data, the simplest and crudest way to test for DMR are to logit-transform the beta values at each site for each sample and compute a *t*-statistic for each site. Wilcoxon rank-sum tests can also be performed [34]. A false discovery rate or Bonferroni correction can be used for multiple comparisons. These methods are not advisable because a type I error or false discovery that occurs at a given site can also occur in nearby sites. Having *p*-values that are correlated are problematic when attempting to compute a false discovery rate. This method is generally not used in practice. Some scientific publications have used the Wilcoxon Rank Sum Test [35]. Pei et al. [36] use bisulfite sequencing data and average methylation values for windows of length 200 bp and perform a *t*-test for each window. They identify genes lying in DMRs for chronic lymphocytic leukemia.

3.2 Fisher's Exact Test

The most basic way to test for DMR with next-generation sequencing data is Fisher's Exact Test [37, 38]. This can be done on a site-by-site basis or can be used by pooling read counts for nearby sites together. Lister et al. [37] used a sliding window approach. Each window is 1 kb. If a region contained at least four differentially methylated CpG sites, it was extended 1 kb at a time until a 1 kb region is reached that does not contain at least four differentially methylated sites. Use of Fisher's Exact Test in this fashion is also not desirable because the results can be heavily affected by the variation in read depth/coverage.

3.3 Methods Based on Differential Variability

Another possible way to identify differentially methylated locations is to find either sites or regions that have more variability in methylation levels than would be expected when there is no differential methylation. Jaffe et al. [39] developed a method based on some smoothed measure of variation according to genomic position. The authors chose a measure of mean absolute deviation as their measure of data spread.

Zhang et al. [40] also developed a method with this variability approach. They devised a test statistic for CpG sites within a region based upon Shannon entropy. Within a group of samples in one particular region where there was differential methylation, the value of the estimated entropy should be higher. A Monte Carlo estimate of the distribution of the test statistic was obtained from samples with uniform methylation. This is realized by simulating regions with uniform methylation level with small deviations about the average.

3.4 Smooth-Based Approach

Bsmooth [41] used a signal-to-noise ratio statistic similar to a function *t*-test to test for DMRs. The variance as a function of position was floored at the 75th percentile and then transformed as a running average. DMRs were defined as regions with the *t*-statistics above the critical value *c*, which was established from the empirical distribution.

Jaffe et al. [42] utilized a method for microarray data called bump hunting which can be used to possibly detect differential methylation with respect to a continuous covariate. For reach CpG site in the microarray, a regression model was fit with the covariate on the logit-transformed *M*-values to obtain the slope for each genomic location. The result was an estimate of the regression slope as a function of position. These values were smoothed using local polynomial regression fitting. The assumption was that the slope function was zero except for certain "bumps" which indicated

regions containing sites where methylation was associated with the covariate. The sum of the absolute values of the regression coefficient over a given region can be the test for differential methylation with a continuous covariate. Permutation tests were performed by permuting the value of the covariate and estimating the distribution of the test statistic.

4 Summary

Various methods have been developed to identify DML or DMRs. The challenges for these methods come from the complexity of the NGS methylation data. Among them, how to make full use of the NGS information is a prominent one. Methods based on Bayesian frameworks have the potential of modeling the complex process involved in NGS. Yet, the standard MCMC algorithm poses big computational burden for genome-wide analysis. An efficient algorithm has to be developed to use these methods for epigenomic analysis.

One key limitation with methods specifically designed for next-generation sequencing datasets is that the sample sizes are generally small due to the lack of availability of bisulfite sequencing data. Hebestreit et al. [43], for example, only uses two sets of six samples because that is a realistic number of samples in any kind of study. Next-generation sequencing data also has the complication of variable coverage across different CpG sites. This can make smoothing more complicated but can also cause some CpG sites to be removed from consideration because the coverage is low. Using Illumina Metylation microarray data could alleviate the problem with the sample sizes, but for many methods the microarray data are too sparse and read counts are required, which one cannot obtain from the microarray data because it only provides signal intensity values for methylated and unmethylated signal intensities for each CpG site that is covered in the microarray.

Using functional data analysis methods appears to be a very good method to deal with the high dimensionality, the autocorrelation, and the missing values within both microarray and next-generation sequencing data. The use of functional data allows for smoothing to represent the data as being a Gaussian process in the clustering, which can reveal categories within the high-dimensional data. Functional clustering seems to be a natural way to extend functional data methods such as the functional *T* tests and *F* tests.

References

1. Maher B. Personal genomes: the case of the missing heritability. Nat News. 2008;456:18–21. doi:10.1038/456018a.
2. Eichler EE, Flint J, Gibson G, et al. Missing heritability and strategies for finding the underlying causes of complex disease. Nat Rev Genet. 2010;11:446–50. doi:10.1038/nrg2809.
3. Handel AE, Ebers GC, Ramagopalan SV. Epigenetics: molecular mechanisms and implications for disease. Trends Mol Med. 2010;16:7–16. doi:10.1016/j.molmed.2009.11.003.

4. Egger G, Liang G, Aparicio A, Jones PA. Epigenetics in human disease and prospects for epigenetic therapy. Nature. 2004;429:457–63. doi:10.1038/nature02625.
5. Robertson KD. DNA methylation and human disease. Nat Rev Genet. 2005;6:597–610. doi:10.1038/nrg1655.
6. Ferguson-Smith AC, Surani MA. Imprinting and the epigenetic asymmetry between parental genomes. Science. 2001;293:1086–9. doi:10.1126/science.1064020.
7. Lee JT. Molecular links between X-inactivation and autosomal imprinting: X-inactivation as a driving force for the evolution of imprinting? Curr Biol. 2003;13:R242–54. doi:10.1016/S0960-9822(03)00162-3.
8. Ehrlich M, Gama-Sosa MA, Huang L-H, et al. Amount and distribution of 5-methylcytosine in human DNA from different types of tissues or cells. Nucleic Acids Res. 1982;10:2709–21. doi:10.1093/nar/10.8.2709.
9. Esteller M, Herman JG. Cancer as an epigenetic disease: DNA methylation and chromatin alterations in human tumours. J Pathol. 2002;196:1–7. doi:10.1002/path.1024.
10. Feinberg AP, Tycko B. The history of cancer epigenetics. Nat Rev Cancer. 2004;4:143–53. doi:10.1038/nrc1279.
11. Bird AP. CpG-rich islands and the function of DNA methylation. Nature. 1986;321:209–13. doi:10.1038/321209a0.
12. Issa J-P. CpG island methylator phenotype in cancer. Nat Rev Cancer. 2004;4:988–93. doi:10.1038/nrc1507.
13. Müller HM, Oberwalder M, Fiegl H, et al. Methylation changes in faecal DNA: a marker for colorectal cancer screening? The Lancet. 2004;363:1283–5. doi:10.1016/S0140-6736(04)16002-9.
14. Nakamura N, Takenaga K. Hypomethylation of the metastasis-associated S100A4 gene correlates with gene activation in human colon adenocarcinoma cell lines. Clin Exp Metastasis. 1998;16:471–9. doi:10.1023/A:1006589626307.
15. Zilberman D, Henikoff S. Genome-wide analysis of DNA methylation patterns. Development. 2007;134:3959–65. doi:10.1242/dev.001131.
16. Yan PS, Shi H, Rahmatpanah F, et al. Differential distribution of DNA methylation within the RASSF1A CpG island in breast cancer. Cancer Res. 2003;63:6178–86.
17. Vucetic Z, Kimmel J, Totoki K, et al. Maternal high-fat diet alters methylation and gene expression of dopamine and opioid-related genes. Endocrinology. 2010;151:4756–64. doi:10.1210/en.2010-0505.
18. Xu H, Podolsky RH, Ryu D, et al. A method to detect differentially methylated loci with next-generation sequencing. Genet Epidemiol. 2013;37:377–82. doi:10.1002/gepi.21726.
19. Rao JNK, Scott AJ. A simple method for the analysis of clustered binary data. Biometrics. 1992;48:577–85. doi:10.2307/2532311.
20. Akalin A, Garrett-Bakelman FE, Kormaksson M, et al. Base-pair resolution DNA methylation sequencing reveals profoundly divergent epigenetic landscapes in acute myeloid leukemia. PLoS Genet. 2012;8:e1002781. doi:10.1371/journal.pgen.1002781.
21. Bediaga NG, Acha-Sagredo A, Guerra I, et al. DNA methylation epigenotypes in breast cancer molecular subtypes. Breast Cancer Res BCR. 2010;12:R77. doi:10.1186/bcr2721.
22. Yang Y, Nephew K, Kim S. A novel k-mer mixture logistic regression for methylation susceptibility modeling of CpG dinucleotides in human gene promoters. BMC Bioinformatics. 2012;13 Suppl 3:S15. doi:10.1186/1471-2105-13-S3-S15.
23. Akalin A, Kormaksson M, Li S, et al. methylKit: a comprehensive R package for the analysis of genome-wide DNA methylation profiles. Genome Biol. 2012;13:R87. doi:10.1186/gb-2012-13-10-r87.
24. Spyrou C, Stark R, Lynch AG, Tavaré S. BayesPeak: Bayesian analysis of ChIP-seq data. BMC Bioinformatics. 2009;10:299. doi:10.1186/1471-2105-10-299.
25. Zheng S, Chen L. A hierarchical Bayesian model for comparing transcriptomes at the individual transcript isoform level. Nucleic Acids Res. 2009;37:e75. doi:10.1093/nar/gkp282.
26. Wu G, Yi N, Absher D, Zhi D. Statistical quantification of methylation levels by next-generation sequencing. PLoS One. 2011;6:e21034. doi:10.1371/journal.pone.0021034.

27. Down TA, Rakyan VK, Turner DJ, et al. A Bayesian deconvolution strategy for immunoprecipitation-based DNA methylome analysis. Nat Biotechnol. 2008;26:779–85. doi:10.1038/nbt1414.
28. Hardcastle TJ, Kelly KA. baySeq: empirical Bayesian methods for identifying differential expression in sequence count data. BMC Bioinformatics. 2010;11:422. doi:10.1186/1471-2105-11-422.
29. Hardcastle TJ, Kelly KA. Empirical Bayesian analysis of paired high-throughput sequencing data with a beta-binomial distribution. BMC Bioinformatics. 2013;14:135. doi:10.1186/1471-2105-14-135.
30. Kruschke JK. What to believe: Bayesian methods for data analysis. Trends Cogn Sci. 2010;14:293–300. doi:10.1016/j.tics.2010.05.001.
31. Chen Z, Liu Q, Nadarajah S. A new statistical approach to detecting differentially methylated loci for case control Illumina array methylation data. Bioinformatics. 2012;28:1109–13. doi:10.1093/bioinformatics/bts093.
32. Huang H, Chen Z, Huang X. Age-adjusted nonparametric detection of differential DNA methylation with case-control designs. BMC Bioinformatics. 2013;14:86. doi:10.1186/1471-2105-14-86.
33. Neuhäuser M. One-sided nonparametric tests for ordinal data. Percept Mot Skills. 2005;101:510–4.
34. Wang D, Yan L, Hu Q, et al. IMA: an R package for high-throughput analysis of Illumina's 450K Infinium methylation data. Bioinformatics. 2012;28:729–30. doi:10.1093/bioinformatics/bts013.
35. Choufani S, Shapiro JS, Susiarjo M, et al. A novel approach identifies new differentially methylated regions (DMRs) associated with imprinted genes. Genome Res. 2011;21:465–76. doi:10.1101/gr.111922.110.
36. Pei L, Choi J-H, Liu J, et al. Genome-wide DNA methylation analysis reveals novel epigenetic changes in chronic lymphocytic leukemia. Epigenetics Off J DNA Methylation Soc. 2012;7:567–78. doi:10.4161/epi.20237.
37. Lister R, Pelizzola M, Dowen RH, et al. Human DNA methylomes at base resolution show widespread epigenomic differences. Nature. 2009;462:315–22. doi:10.1038/nature08514.
38. Pelizzola M, Ecker JR. The DNA methylome. FEBS Lett. 2011;585:1994–2000. doi:10.1016/j.febslet.2010.10.061.
39. Jaffe AE, Feinberg AP, Irizarry RA, Leek JT. Significance analysis and statistical dissection of variably methylated regions. Biostatistics. 2012;13:166–78. doi:10.1093/biostatistics/kxr013.
40. Zhang Y, Liu H, Lv J, et al. QDMR: a quantitative method for identification of differentially methylated regions by entropy. Nucleic Acids Res. 2011;39:e58. doi:10.1093/nar/gkr053.
41. Hansen KD, Langmead B, Irizarry RA. BSmooth: from whole genome bisulfite sequencing reads to differentially methylated regions. Genome Biol. 2012;13:R83. doi:10.1186/gb-2012-13-10-r83.
42. Jaffe AE, Murakami P, Lee H, et al. Bump hunting to identify differentially methylated regions in epigenetic epidemiology studies. Int J Epidemiol. 2012;41:200–9. doi:10.1093/ije/dyr238.
43. Hebestreit K, Dugas M, Klein H-U. Detection of significantly differentially methylated regions in targeted bisulfite sequencing data. Bioinformatics. 2013;29:1647–53. doi:10.1093/bioinformatics/btt263.

Performance Comparison and Data Analysis Strategies for MicroRNA Profiling in Cancer Research

Erik Knutsen, Maria Perander, Tonje Fiskaa, and Steinar D. Johansen

Abstract MiRNA profiling generates a global view of the miRNA expression pattern in a biological or clinical sample, and there are currently several different profiling platforms available. However, correct and accurate measurements of miRNA have turned out rather challenging, and performance comparisons of available profiling technologies are of high importance in order to improve the correlation of results generated by different studies. In this chapter we discuss general considerations of miRNA profiling approaches and data analysis strategies, applied in cancer research. Performance comparisons of the SOLiD and Illumina high-throughput next-generation sequencing platforms, as well as hybridization-based methods and PCR-based methods are presented. Topics covered include platform sensitivity and accuracy, miRNA isoforms (isomiRs), and miRNA normalization procedures.

1 Introduction to MicroRNA Profiling

1.1 MicroRNAs

MicroRNAs (miRNAs) were first characterized in the early 1990s in *C. elegans* as small noncoding RNA (ncRNA) transcripts with the ability to regulate specific mRNAs by complementary base pairing [1]. Today, a number of small ncRNA classes and subgroups have been discovered, with the most studied classes being the miRNAs, the small interfering RNAs (siRNAs), and the piwi-interacting RNAs (piRNAs). These classes share important aspects in size, biogenesis, and associated

E. Knutsen (✉) • M. Perander • T. Fiskaa
Department of Medical Biology, Faculty of Health Sciences, University of Tromsø, 9037 Tromsø, Norway
e-mail: erik.knutsen@uit.no

S.D. Johansen
Department of Medical Biology, Faculty of Health Sciences, University of Tromsø, 9037 Tromsø, Norway

Marine Genomics group, Faculty of Biosciences and Aquaculture, University of Nordland, 8049 Bodø, Norway

W. Wu, H. Choudhry (eds.), *Next Generation Sequencing in Cancer Research, Volume 2*, DOI 10.1007/978-3-319-15811-2_15

proteins, but differ in function and origin [2, 3]. MiRNAs are now recognized as important modulators of gene expression at a posttranscriptional level in a wide variety of organisms, and they play pivotal roles in many cellular processes by targeting specific mRNAs for degradation and silencing.

The majority of miRNAs are processed from RNA polymerase II transcripts, named primary miRNAs (pri-miRNAs), that form characteristic hairpin structures (Fig. 1) [2]. The pri-miRNAs are cleaved by the RNase III endonuclease Drosha and

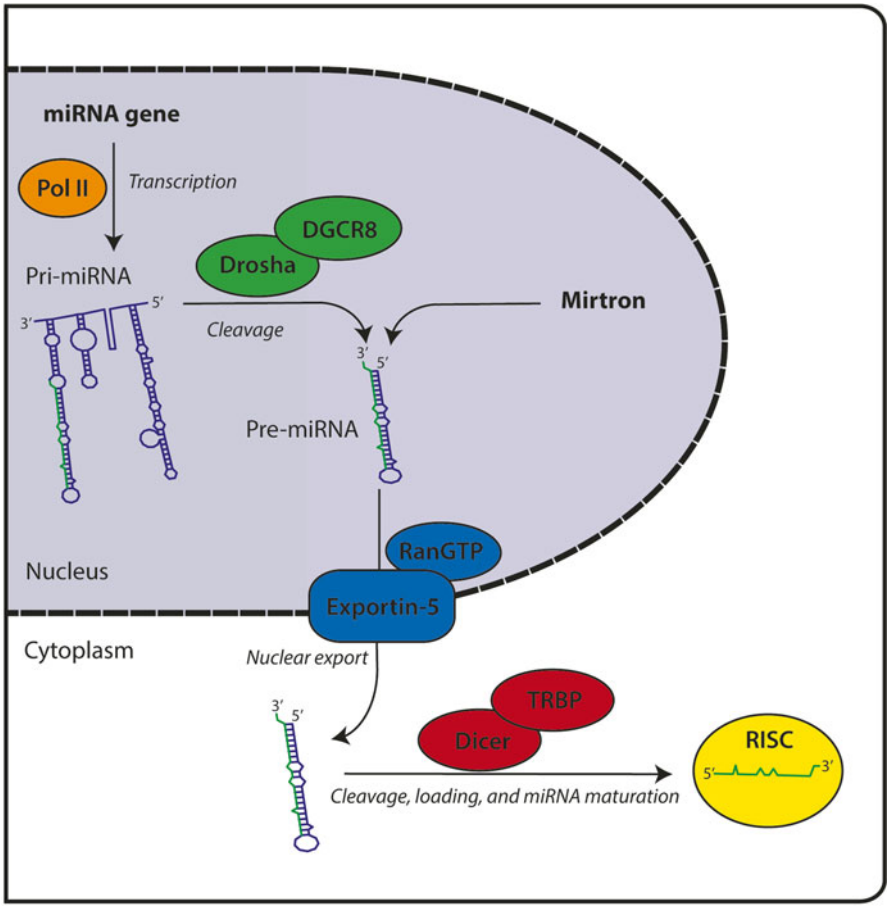


Fig. 1 Biogenesis of miRNAs. The majority of miRNAs are transcribed by RNA polymerase II (Pol II) into primary miRNA transcripts (pri-miRNAs). Pri-miRNAs are processed into precursor miRNAs (pre-miRNAs) by the enzyme Drosha and the double stranded RNA binding protein (dsRBP) DGCR8. Exportin-5 in complex with RanGTP recognizes pre-miRNAs and mirtrons, spliced-out introns sharing specific features of pre-miRNAs, and transports the transcript to the cytosol . Here they are further processed by Dicer and dsRBP TRBP into a miRNA duplex and loaded onto the RNA-Induced Silencing Complex (RISC). Maturation involves separation of one of the strands from the duplex, which directs posttranscriptional silencing. See main text for details and references

its cofactor DGCR8 into a 65–70 nt stem-loop precursor miRNAs (pre-miRNAs). Pre-miRNAs are then exported from the nucleus to the cytoplasm by Exportin 5 and Ran-GTP. In addition, a minor subset of miRNAs (mirtrons) is generated from spliced-out introns. Mirtrons share common features with pre-miRNAs and are therefore directly recognized by Exportin 5. Pre-miRNAs and mirtrons are further processed by Dicer, a cytoplasmic RNase III endonuclease, that in complex with TRBP generates a 22 nt duplex with 2-nt overhang at both 3′ ends. One of the strands of the duplex, often referred to as the guide strand, becomes the functional mature miRNA, and it is incorporated into the RNA Induced Silencing Complex (RISC). The other strand, referred to as the passenger strand, is in many cases degraded, but can under certain cellular circumstances be selected as the mature miRNA. This allows each miRNA hairpin structure to generate two distinct mature miRNAs with different target genes and presumable biological role. The two strands are named accordingly to their cleavage position at the 5′ end (miR-X-5p) or the 3′ end (miR-X-3p) of the hairpin. So far, 1881 precursors and 2588 mature miRNAs have been annotated in the human genome (miRBase v. 21, June 2014) ranging from 16 to 28 nt in length [4–8].

Mammalian miRNAs usually pair imperfectly with their mRNA targets at the 3′ untranslated region (UTR) [3]. Several computational and biochemical studies indicate that their specificity is dependent on position 2–8 of the guide strand, also known as the "seed" region. The mechanism used by miRNAs to regulate gene expression is still somewhat controversial. Some miRNAs destabilize their target mRNA by recruiting deadenylation factors that remove the poly-A tail, thereby making the mRNA susceptible to exonucleolytic degradation. Others suppress the initiation or elongation of protein synthesis, and some can promote posttranslational degradation of the newly synthesized peptide. Finally, a few miRNAs have even been shown to activate the expression of certain genes [3].

1.2 MiRNAs in Cancer

MiRNAs have crucial roles in the regulation of cellular processes that are altered during cancer initiation and progression [9–11]. The first link between miRNAs and cancer was established in 2002 when two miRNAs, miR-15 and -16, were found to be deleted in chronic lymphocytic leukemia [12]. Since then, whole-genome profiling of miRNAs in both solid and hematological tumors have demonstrated that abnormal miRNA expression signatures are a general trait of cancer [13, 14]. MiRNAs have been shown to negatively regulate the expression of both oncogenes and tumor suppressor genes, and thus can act as either tumor suppressors or oncogenes (oncomiRs) [9–11]. Interestingly, downregulation of miRNAs represents a more frequent event in cancer pathogenesis, suggesting that most miRNAs act as tumor suppressors [13]. In breast cancer, miR-103/107, whose expression is associated with metastasis and poor outcome, directly targets Dicer and causes a global downregulation of the miRNA network [15]. Furthermore, tumor suppressing

miRNA genes are often located in loci subjected to deletions, mutations, or epigenetic silencing, leading to their loss of function [16]. MiR-21, on the other hand, is one of the most frequently upregulated miRNAs in solid tumors and act as an oncomiR by targeting tumor suppressor genes encoding the proteins Tropomyosin 1 (TPM1), Phosphatase and tensin homolog (PTEN), and Programmed cell death 4 PDCD4 [17–19].

In general, specific cancer miRNA expression patterns are shown to be consistent in tumor samples deriving from the same developmental lineage or differentiation state, enabling a more precise classification of the tumor [11, 13, 14, 17, 20]. Therefore, miRNA expression profiling is not only important from a basic research point of view, but also in translational research as this can provide important information for correct diagnosis, progression, and prognosis of the disease. In addition, miRNAs are easily detected in tumor biopsies [21], and they have also been shown to exist with remarkable stability in various types of body fluids, including blood [22]. This has led to increased research focus on disease-related variations in serum and plasma miRNA expression and the possibility that such variations could serve as promising diagnostic, prognostic, and predictive cancer biomarkers in the future.

1.3 Definition of miRNA Profiling

MiRNA profiling is the measurement of the abundance of all miRNA species in a single sample. In contrast, studies of only a few selected miRNA species should be referred to as "expression analyses" as this is a different discipline especially in regard to data management. Acquiring a global overview of the miRNA abundance in a biological or clinical sample is often instrumental to get a full picture of the disease, and to select interesting candidates for further functional studies.

Correct and accurate measurements of miRNA contents in cells and tissues have historically turned out rather challenging, and a lack of significant agreement between different miRNA profiling studies has been reported. Chen and coworkers compared published miRNA expression profiles in epithelial ovarian cancer across a 10-year span [23]. Eight papers were included in the comparison, and out of 17 miRNAs that were reported in at least four profiling studies, only one third had a consistent change in expression. The inability to validate miRNA expression profiles across different studies is due to several factors, including the use of different profiling technologies, platforms, protocols, and methods for data analysis.

1.4 MiRNA Profiling Technologies

There are several methods used to investigate miRNA expression. In addition to traditional small-scale experiments using Northern blot analyses and in situ hybridization (ISH), developed methods for medium- or high-throughput analyses

including microarray analyses, quantitative real-time PCR (RT-qPCR), and next-generation sequencing (NGS) are now commonly used both in research and medical diagnosis. All methods can be categorized into one out of three main technologies: hybridization-based, PCR-based, or sequencing-based. The choice of technology and method depends on several factors such as the starting material, number of samples, time restrictions, manpower, cost, and if the aim is to discover novel miRNA species or isoforms of known miRNAs (isomiRs).

Precise identification and accurate quantification of miRNAs is challenging due to their small sizes and frequent modifications like 2′-*O*-methylation and adenosine to inosine (A-I conversions) editing. Furthermore, the analyses have to discriminate between mature and primary forms of a miRNA, between highly similar miRNA families, and between isoforms of specific miRNAs. Also, variations in GC-contents between different miRNAs can influence the global analyses.

1.4.1 Hybridization-Based Methods

Microarray analyses are widely used for miRNA profiling, especially when the number of samples to be analyzed is high [24]. Here, quantification implements fours steps: (1) Fluorescent labeling of miRNA; (2) hybridization to array of DNA-based capture probes; (3) washing and scanning of array; and (4) data extraction and processing. A major challenge for microarray in regard to miRNAs profiling is that the DNA-based capture probes used for hybridization often will have to cover the entire sequence of the mature miRNA (about 20 base pairs), resulting in a wide Tm range for the entire miRNA population and thereby decreased binding efficacy or fluorescent distortion. This is by some approaches corrected for by the addition of miRNA specific tags or single nucleotides to increase the melting temperature for each individual miRNA to a similar magnitude. Also the introduction of Locked Nucleic Acid (LNA)-modified oligonucleotide probes can eliminate the distortion of Tm by enhancing binding affinity for specific probes. Microarray will always face the issue of cross-hybridization of similar miRNAs, and this should be kept in mind as this can reduce the sensitivity.

In 2008 a novel hybridization-based technology, the NanoString nCounter system, was introduced for miRNA profiling [25]. This system has the advantage of detecting specific nucleic acid molecules from low amounts of starting material without the need for reverse transcription or cDNA amplification. The technology captures and counts nucleic acids in a complex mixture using sequence-specific molecular barcodes and single molecule imaging. The NanoString nCounter system circumvents the problem of probe hybridization with the short miRNA sequence by bridge ligation of miRNA specific tags.

1.4.2 PCR-Based Methods

RT-qPCR is considered as the "gold standard" for expression analysis of miRNA using either TaqMan or SYBR green assays [26]. In the TaqMan assay, miRNA-specific stem–loop primers are used for cDNA synthesis, thereby increasing the length of the miRNA DNA template. Probe sequences that fluoresce upon hydrolysis by a DNA polymerase with exonucleolytic activity are included in the amplification and are used for quantification. In the SYBR green assay, a poly(A) tail is added to all RNA transcripts before reverse transcription. MiRNA are detected by a specific forward primer and a reverse primer that anneals to both the 3′ miRNA sequence as well as to the poly(A) tail. The double stranded DNA-intercalating SYBR green dye is used for quantification. MiRNA primer design, unspecific binding, and nonlinearity in PCR amplification can be a challenge for RT-qPCR [27]. Synthesis of cDNA is also subject to errors with variations resulting from secondary structures, differences in priming efficiencies, and biases caused by the reverse transcriptase [28, 29]. As novel miRNAs are constantly discovered, the sizes of primer panels for RT-qPCR constantly need to be expanded, leading to an increase in the cost of the technology.

1.4.3 Sequencing-Based Methods

NGS is the only technology that simultaneously allows for the discovery of new miRNAs and quantification of annotated miRNAs. The technology includes some common library preparation steps: (1) Unspecific adaptor oligonucleotide RNA ligation; (2) reverse transcription through primer binding site included in adaptor; (3) size selection of fragments within the correct size range; and (4) PCR amplification. The method circumvents many of the problems faced by microarray and RT-qPCR, like variability in melting temperatures, co-expression of nearly identical miRNA family members, or posttranscriptional modifications. However, NGS can suffer from laborious and inherent biases in library preparation and extensive and intricate data analysis [30–32]. Further research and development within this technology, including launching of new medium-throughput benchtop sequencing platforms, might circumvent some of these challenges.

2 Platform Comparison

Several reports have been published the last few years comparing the performance of different miRNA profiling technologies (see Table 1 for an overview). All studies use different approaches both in regard to sample material, number of miRNAs profiled, normalization method, and statistical method to determine the sensitivity and accuracy of the technologies under investigation. Even though the overall correlation of profiles generated by different technologies is in concordance, there are

Table 1 Recent platforms comparison studies that include a minimum of two technologies

Technologies (platforms)	# samples	# miRNA species	Normalization methods	Sensitivity	Accuracy	Reference, Year
Microarray (Illumina) NGS (Illumina) RT-qPCR (SYBR Green)	20 samples (benign follicular thyroid adenoma and malignant follicular thyroid carcinoma)	1145 miRNAs used for correlation of NGS and microarray 6 miRNA used for correlation with qPCR	Illumina (microarray)—normalized, method unknown Illumina (NGS)—linear total count scaling SYBR Green—reference gene normalization (RNU6B)		– NGS and microarray has a satisfactory correlation of 0.67 (unknown method, log2 absolute values) – A very high correlation is seen for the 6 miRNAs validated with qPCR with the two other technologies (0.86–0.90 Pearson's correlation, log2 absolute values)	[33] 2014
Microarray (Illumina) NanoString (nCounter) NGS (SOLiD, Illumina) RT-qPCR (TaqMan)	Ten samples (five pairs of non-small-cell lung cancer cell lines and their corresponding xenograft models)	205 miRNAs used for interplatform comparison 68 miRNA used for the correlation with RT-qPCR	Illumina (microarray)—robust spline normalization algorithm nConter—geometric mean of five housekeeping mRNA genes (ACTB, B2M, GAPDH, RPL19, and RPL10) and thereafter to total miRNA count SOLiD4—linear total count scaling Illumina (NGS)—linear regression using the median count value of each miRNA across the samples as reference TaqMan—geometric mean of let-7 g, miR-191 and miR-335-3p		– The combination of the two NGS platforms had the highest correlation (0.75 Spearman's correlation, log2 relative expression) – A high correlation was also seen with the combination of NGS and microarray (0.73 and 0.69, SOLiD4 and Illumina respectively, Spearman's correlation, log2 relative expression) – The lowest correlation was seen when platforms were combined with the nCounter (0.49–0.52 Spearman's correlation, log2 relative expression) – For the comparison of RT-qPCR and the four other platforms, a high correlation was seen with all platforms, with the highest being SOLiD4 (0.86 Spearman's correlation, log2 relative expression) and the lowest being nCounter (0.72 Spearman's correlation, log2 relative expression)	[34] 2014

(continued)

Table 1 (continued)

Technologies (platforms)	# samples	# miRNA species	Normalization methods	Sensitivity	Accuracy	Reference, Year
NanoString (nCounter) NGS (SOLiD and Illumina) RT-qPCR (SYBR Green)	Four samples (human breast cell lines)	517 miRNAs included in correlation In addition all miRNAs screened for were included in additional sensitivity comparison and accuracy comparison (ranging from 631 to 1719)	All platforms were normalized by linear total count scaling	– NGS had the highest sensitivity (0.982 and 0.983, SOLiD and Illumina, respectively). Sensitivity was calculated by a miRNA being defined as a true positive if at least three out of the four platforms identified the miRNA – RT-qPCR came close to the NGS platforms with a sensitivity of 0.959 – NanoString, had the lowest sensitivity (0.645) – A more sophisticated method of calculating sensitivity where all miRNAs screened for were included gave the same result	– Overall correlation for all platforms was high (0.703–0.797 Pearson's correlation, log2 relative expression) with the highest being the combination of RT-qPCR and NanoString and the lowest being the combination of NGS (SOLiD) and NanoString	[35], 2013

Microarray (Affymetrix, Agilent, Illumina) NanoString (nCounter) NGS (Illumina) RT-qPCR (TaqMan)	Six samples (two fresh frozen tissues, two formalin fixated tissues from normal human lung and lung tumors, and two lung carcinoma cell lines)	484 miRNAs for interplatform comparison 39 miRNAs (average) were included for comparison with RT-qPCR	Affymetrix and Illumina (microarray)—no normalization Agilent—unknown nCounter—average count Illumina (NGS)—linear total count scaling TaqMan—unknown	– NGS and Illumina (Microarray) had a high number of commonly detected miRNAs with the other platforms (average of 1139 miRNAs for both platforms in the six samples). – Affymetrix (Microarray) showed a medium common detection (average of 977 miRNAs). – Agilent (Microarray) and NanoString showed a low common detection (average of 801 and 862 miRNAs, respectively).	– NGS had the highest average correlation (0.588 Spearman's correlation, relative expression) when correlated to RT-qPCR – Agilent (Microarray) had the lowest (0.423 Spearman's correlation, relative expression) when correlated to RT-qPCR	[36] 2013

(continued)

Table 1 (continued)

Technologies (platforms)	# samples	# miRNA species	Normalization methods	Sensitivity	Accuracy	Reference, Year
Microarray (Agilent and Affymetrix) NGS (Illumina)	Seven samples (human placenta RNA spiked in with 6 synthetic mature miRNA oligoes at different concentration)	Six synthetic mature miRNA oligoes	Agilent—quantile normalization Affymetrix—quantile normalization Illumina (NGS)—linear total count scaling		– Agilent (Microarray) was found to be the most accurate platform showing closed fold change means (10.7) and the smallest standard deviations (3.2) when compared to synthetic miRNAs that were introduced at concentrations with tenfold incremental differences (10 vs 1-fmol concentration) – NGS showed a higher difference in relative expression than the true tenfold incremental differences with a fold change means of 14.8 and standard deviation of 8.6 – Affymetrix (Microarray) underestimated the fold changes, and even had some which were not statistically significant (fold change means of 7.2 and standard deviations of 6.5) – For the lowest concentration (1 vs 0.1-fmol concentration), a similar performance was seen for the three platforms	[37] 2013

Microarray (Agilent, Exiqon, Illumina, Ambion, Combimatrix, and Invitrogen) NGS (Illumina) RT-qPCR (either TaqMan or SYBR Green)	Three samples (a pool of commercial RNAs from normal breast tissue, and two breast cancer cell lines)	204 miRNA included in correlation 89 miRNAs were included for correlation with RT-qPCR	Ambion, Combimatrix, and Invitrogen—loess spatial correction within arrays Agilent and Exiqon—quantile normalization between arrays Illumina (Microarray)—no normalization Illumina (NGS)—linear total count scaling TaqMan—reference gene normalization (RNU48) SYBR green—reference gene normalization (5S rRNA)		– A large difference in correlation was seen for the combination of NGS and the different microarray platforms (0.42–0.77 Pearson's correlation, log2 absolute values), with the best being Illumina (Microarray), and the worst being Invitrogen (Microarray) – Similar results were seen for correlation between the different microarray platforms (0.42–0.85-Pearson's correlation, log2 absolute values), where the best were Agilent (Microarray) and Ambion (Microarray), and the worst were the combination Invitrogen (Microarray) and Illumina (Microarray) and Invitrogen (Microarray) and Combimatrix (Microarray) – A high correlation was seen for RT-qPCR with the other platforms (0.68–0.92 Pearson's correlation, log2 relative values), with the best being NGS, and the worst being Combimatrix (Microarray)	[26] 2010
Microarray (Affymetrix, Agilent, and Illumina) NGS (Illumina) RT-qPCR (TaqMan)	Two samples (human heart and brain total RNA)	218 miRNAs for correlation of NGS and microarray with RT-qPCR 718 miRNAs for correlation of NGS with microarray	Affymetrix, Agilent, and Illumina (Microarray)—quantile-normalization Illumina (NGS)—quantile-normalization TaqMan—reference gene normalization (U6 and RNU48)		– A high correlation was seen between the technologies (0.66–0.90 unknown method, log2 relative values), with the best being RT-qPCR and NGS and RT-qPCR and Agilent (Microarray), and the worst being NGS and Affymetrix (Microarray)	[38] 2010

(continued)

Table 1 (continued)

Technologies (platforms)	# samples	# miRNA species	Normalization methods	Sensitivity	Accuracy	Reference, Year
Microarray (Agilent) NGS (SOLiD and Illumina) RT-qPCR (TaqMan)	Two samples (resting and activated NK cells)	134 miRNAs for sensitivity analysis 338 miRNAs for correlation of NGS platforms	Agilent—median normalization (75 % intensity) SOLiD—linear total count scaling Illumina—linear total count scaling TaqMan—reference gene normalization (U6)	– A high detection rate were seen for the NGS platforms and RT-qPCR (detected 93 %, 91 % and 87 % of profiled miRNAs, RT-qPCR, Illumina, and SOLiD, respectively), while a low detection rate was seen for microarray (detected 68 % of profiled miRNAs)	– High correlation of absolute values for the combination of the two NGS platforms (average correlation of 0.77 Spearman's correlation, log2 absolute values)	[39] 2009

Microarray (Exiqon) NGS (Illumina)	Two samples (two synthetic samples from 744 synthetic RNA oligos)	744 synthetic oligos	Unknown if any normalization was introduced	– Microarrays were more sensitive than sequencing, especially at the lowest concentration (94 % percent of RNAs detected in microarray compared to 70 % in NGS for the 47 lowest expressed RNAs), and had a lower false negative rate (0,97 % undetected RNAs in microarray, compared to 3.1 % in NGS)	– Microarray data correlated best with the known RNA concentration (0.69 Pearson's correlation, log2 absolute values), sequencing data were less correlated (0.50 Pearson's correlation, log2 absolute values) – A poor correlation of the two platforms were seen for absolute values (0.47 Pearson's correlation, log2 absolute values), while relative values had a very good correlation (0.93 Pearson's correlation, log2 relative values). Both technologies correlated well with the expected ratios (0.96 Pearson's correlation, log2 relative values)	[40] 2009

Key aspects, like type of platforms and technologies, sample material, number of miRNAs profiled, normalization method, and statistical method to determine the sensitivity and accuracy of the technologies under investigation are displayed

several important discrepancies and pitfalls that have to be considered. Importantly, reports that compare different platforms within the same technology conclude that the type of technology is not the major variability factor. In fact it is the sample preparations and pre-quantification steps that is most influential on the correlation of the results [38, 41–43].

2.1 Experimental Implications

Depending on the purpose of the study, miRNA expression profiling can be performed on different biological or pathological samples derived from cell cultures, fresh tissue, formalin-fixed paraffin-embedded (FFPE) tissue, or body fluids like plasma or serum, saliva, and urine. Compared to longer RNA transcripts, miRNAs are stable molecules even in body fluids that are known to contain substantial amounts of RNases [44–49]. This is probably because they are protected in either protein complexes or in membrane-enclosed exosomes [47, 49]. Also in FFPE samples, that display a varying degree of degraded RNA depending on fixation efficiency and storage time, miRNAs are relatively intact [36, 50–53]. MiRNA expression profiling can therefore be performed on large cohorts of clinical samples from biobanks of stored blood or FFPE samples. Custom-made protocols for isolation of RNA from different sources followed by enrichment steps for small RNAs have been developed, both by research groups and life science manufacturers [24, 36].

The abundance of mature miRNAs in clinical samples from both body fluids and tissues is however often low. In many cancers, a global downregulation of miRNAs is observed that could be due to chromosomal deletions, epigenetic silencing, or genetic loss of proteins involved in the biogenesis of miRNAs [9, 10]. Also the amount of starting material from different cohorts might be limited. Intra-tumor genomic heterogeneity is a common phenomenon in solid cancer, raising the issue of doing profiling in subsets of cells or even in single cells [54–58]. Therefore, miRNA profiling technologies face the challenge of generating reliable expression data on low-input samples. In general, amplification-based methods have a benefit in this regard, and compared to microarray, RT-qPCR stands out as more sensitive for analyses with only minute input of miRNA from for instance blood plasma [59]. NGS platforms are constantly improving their library-preparation protocols adjusting to low-input samples, and small RNA-sequencing is now performed on samples with a starting concentration as low as 1–100 ng small-RNA enriched RNA.

The number of samples, time, manpower, as well as the budget of the research project, will undoubtedly affect the choice of profiling method. The microarray platforms are still the cheapest alternative for miRNA profiling [24]. However, the new benchtop sequencers are getting within the range of the running cost of the microarray platforms [60]. RT-qPCR profiling is typically expensive as multiplexed array panels should be run in triplicates. A single run on an NGS platform is costly, and the sequencing takes several days to complete [61]. In addition, the library preparation protocols are laborious. However, the depth of the sequencing allows

multiple samples to be analyzed in one experimental set-up by barcoded samples. This contrasts microarray and RT-qPCR platforms where only one sample can be analyzed at a time (two samples for some microarray platforms).

To conclude, the purpose of the study and available resources will ultimately decide which technology to choose. A broad understanding about both the advantages and the pitfalls of the different technologies is essential for correct biological and clinical interpretation of the generated data.

2.2 Technical Implications

2.2.1 Platform Sensitivity

The sensitivity of a platform is defined as the ability of the platform to only detect miRNA species that are truly present in a sample, and not identifying true negatives. Sensitivity is best measured by profiling of a synthetic sample where the identity and composition of the different species are known. This however is usually performed by companies and rarely by the scientific community since no biological information is gained. Most sensitivity comparisons between platforms are for this reason performed on biological samples with unknown composition and concentration of specific miRNAs. In such analyses, a general assumption is that the likelihood for a detected miRNA to be a false positive decreases if other platforms also detects the same miRNA. This postulate can be hampered if most of the technologies under investigation in fact have a low sensitivity. Under such circumstances, the technology with a true high sensitivity would easily be regarded as the one identifying a high number of false positives. This important notion will have to be taken into considerations when interpreting data from performance experiments.

A general assumption is that microarray has a lower sensitivity than both NGS and RT-qPCR platforms [24], even though both lower and higher sensitivities have been reported (Table 1) [36, 39, 40]. Different microarray platforms differ in terms of oligo probe design, sample labeling, probe immobilization chemistry, microarray chip signal-detection methods, and most importantly the amount of RNA used for profiling and the inclusion of pre-amplification steps. All these factors can potentially influence on the sensitivity of the analyses. A major problem in microarray is cross-hybridization, both of similar miRNA species and of pre-miRNAs, generating false positive results, and thereby reducing the sensitivity of the technology.

The NanoString nCounter system has been reported to have a low sensitivity by two separate platform comparison studies (Table 1) [35, 36]. The fact that nCounter does not include any amplification steps of the miRNA prior to detection may explain the observed low sensitivity, as this potentially can reduce the window between a true-positive miRNA expression and the background signal. Similar to other hybridization based technologies, nCounter has limitations in distinguishing between highly similar target sequences, especially heterogeneities at the 5′ end of the miRNA. Such limitation is important to note since a number of miRNA species differ in sequence with only a single nucleotide.

RT-qPCR is often referred to as the "gold-standard" in miRNA profiling. In most platform comparison studies RT-qPCR is only included for verification of result generated by other platforms (Table 1) [26, 33, 34, 36]. RT-qPCR is highly sensitive [35, 39] and represents the method of choice if the RNA input concentration is low.

NGS is regarded less sensitive compared to RT-qPCR [24]. We recently detected a slightly higher concordance in miRNA species between the SOLiD and Illumina platforms than by the combination of either NGS platforms with RT-qPCR [35]. However, the level of detection by NGS and RT-qPCR technologies is similar in two independent platform comparison studies, and thus difficult to rank (Table 1) [35, 39]. Furthermore, the sequence depth of NGS will affect the technologies sensitivity since high depth will increase the numbers of miRNAs profiled.

The probability of a miRNA to be detected across the various platforms is higher for an extensively expressed miRNA than for a scarcely expressed miRNA. This was clearly visualized in our own platform comparison study, where the majority of miRNAs detected by only a single platform were expressed at low levels, whereas the majority of highly expressed miRNAs were detected by all four platforms [35]. As many of the platforms are dynamic in regard to the amount of starting material, pre-amplification, and detection depth, the sensitivity can therefore be increased for the individual platforms. However, the detection of specific miRNAs should be verified by additional methods because of the presence of highly similar miRNA species. Here, verification will be affected by the abundance of the miRNA under investigation, and this should be kept in mind when selecting candidate miRNAs for verification.

2.2.2 Platform Accuracy

The accuracy of a platform is defined as the ability to correctly measure the exact concentration of a specific miRNA specie in a sample. The expression levels can be presented in two different, but equally important manners: (1) The absolute value presented as a count number for each individual miRNA species within a single sample, or (2) the relative values presented as a change in expression of a specific miRNA between two samples (Fig. 2a). The two different presentations of expressions are of equal importance in regard to platform performance, but the biological aspects of the results have to be interpreted differently.

Absolute values result in a ranked list of miRNAs present in a sample. However, correct estimation of absolute values is challenging [39–41]. Linsen and coworkers reported a profiling study using two NGS platforms and RT-qPCR that included 473 synthetic human miRNAs at equal molarity. Interestingly, for all platforms a non-uniform distribution of miRNAs was observed with up to four orders of magnitude difference between the most and least frequently detected miRNA [41]. Furthermore, only about half of the miRNAs profiled varied within a single order of magnitude. For NGS, this discrepancy is probably due to library preparation methods, and not the different sequencing platforms. Several comparison studies have seen a higher concordance when the same library preparation has been subjected to profiling by

a

miRNA	Absolute values: Normal RPM	Absolute values: Tumor RPM	Relative values: Fold Change
miR-182-5p	21	1933	90,9
miR-378a-3p	15	359	23,5
miR-345-5p	19	417	21,3
miR-146a-5p	83	1494	17,8
miR-192-5p	195	3208	16,4
miR-196a-5p	159	2492	15,7
miR-30d-5p	52	424	8,1
miR-15b-5p	3315	26524	8,0
miR-25-3p	587	4643	7,9
miR-21-5p	3121	24472	7,8

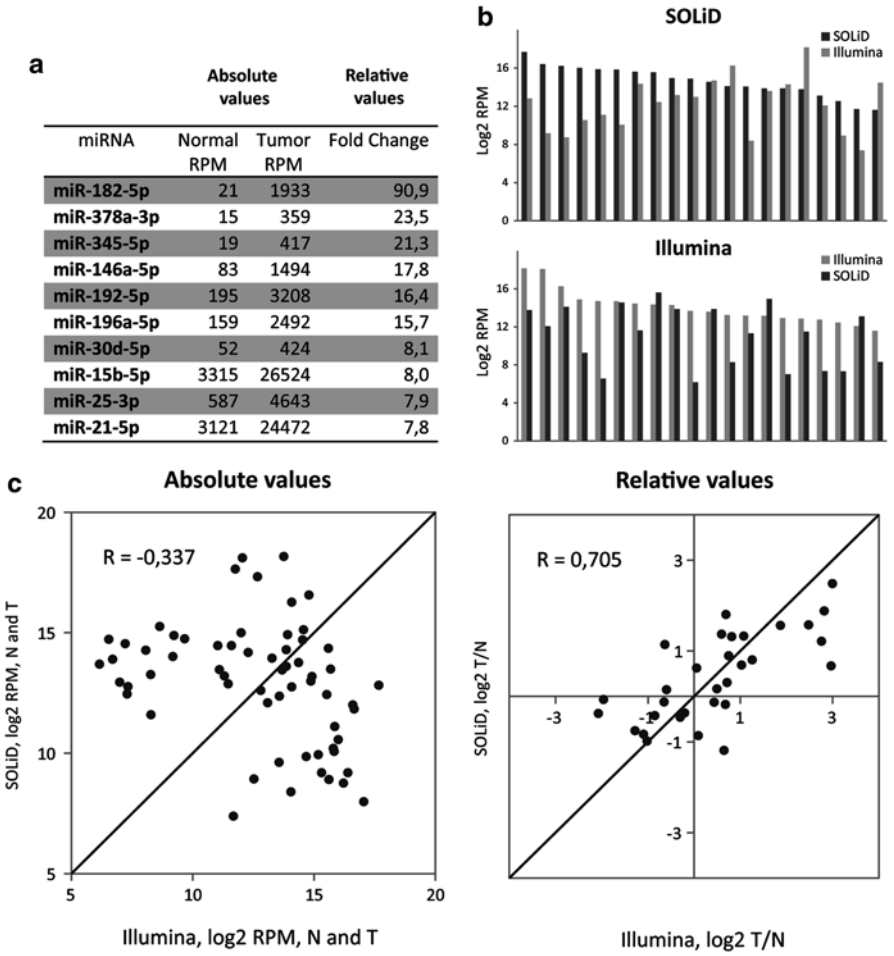


Fig. 2 Effect of absolute and relative miRNA quantification values for platform concordance. (**a**) Normal (Hs 578Bst) and breast tumor (Hs 578T) cell lines were sequenced by SOLiD small RNA-Seq, and top-ten upregulated miRNAs in tumor samples are displayed. Both relative values (fold change) and absolute values (RPM) are presented. (**b**) Comparison between SOLiD and Illumina NGS platforms for the 20 highest expressed miRNAs (absolute values) in Hs 578Bst. Disagreement between platforms in absolute values is clearly visualized. A similar trend is noted when comparing the same dataset with RT-qPCR and NanoString. (**c**) Same miRNA data set as presented in **b** visualized in a scatterplot. The absolute values (*left*) indicate that the data generated from SOLiD and Illumina platforms are not well correlated (Pearson's correlation of −0.337). However, relative values (*right*) show a much better correlation between the two NGS platforms (Pearson's correlation of 0.705). RPM, Reads Per Million; NGS, Next-Generation Sequencing; N, Normal; T, Tumor

different NGS platforms, than different type of libraries subjected to profiling by the same NGS platforms [41–43]. We show that different profiling approaches preferentially capture a distinct set of miRNAs (Fig. 2b), but as technical replicates are highly similar for each platform we infer that the biases are of a systematic nature.

These biases complicate the comparison of miRNA absolute expression, which currently cannot be exactly determent by any technology.

Relative values are however not affected by library preparation protocols since identical biases are introduced for both the control and the test sample. A higher platform correlation is usually observed in studies based on relative values compared to absolute values (Fig. 2c) [41, 42]. There are, however, important aspects of relative values that need to be addressed in order to better understand the data generated. First, relative values will be strongly affected by low expression levels in either one or both paired samples. This is clearly illustrated from our own data set of the ten most upregulated miRNAs detected from a comparison between normal and breast tumor cell lines (Fig. 2a). Here, miR-182-5p has a fold change of 90.9, but the difference in absolute values is only 1,912 reads per million (RPM). In comparison, miR-15b-5p has a fold change difference of only 8.0, but the absolute difference is 23,209 RPM. The second concern relates to which order of magnitudes calculated fold change values should have to confidently claim that the observed difference in miRNA expression is significant. As previously stated, agreement between different platforms implies that an observation has a higher likelihood of being a correct estimate. Importantly, if the fold change calculated for a specific miRNA between two samples is less than 2, the agreement between two platforms falls dramatically. This is clearly demonstrated in our study where the same direction of fold change values (upregulated or downregulated) is only observed for 81 % of miRNAs that display less than twofold differences in relative expression [35]. These observations strongly suggest that because of technical discrepancies, fold change values should be above 2 to conclude that a specific miRNA is differentially expressed in two samples. This notion does not take into consideration biological variations. Of note, agreement between different technologies on fold change directions is relatively constant across different miRNA concentrations [35]. This means that even though the expression level of a specific miRNA is low, observed differences in expression in paired samples can be trusted as correct as long as fold change values above 2 are obtained [35].

2.3 MiRNA IsomiRs

NGS platforms offer genome-wide approaches for profiling that are not restricted to known miRNA sequences provided by preestablished databases. This implies that novel miRNA species and variants can be discovered. Studies of miRNA populations in cells from different organisms have demonstrated that there exist several miRNA isoforms (Fig. 3a) [62–66]. These *isomiRs* contain either deletions or extensions at the 5′- or 3′-ends, or single nucleotide changes within the miRNA. IsomiRs might be a result of imprecise processing of the pri-miRNA and pre-miRNA, trimmed miRNA ends by exoribonucleases (“nibbling”), 3′ uridylatation or 3′ adenylation by nucleotidyl transferases (“tailing”), or RNA editing enzymes [63]. There is currently a debate whether some of the miRNA variants observed by NGS

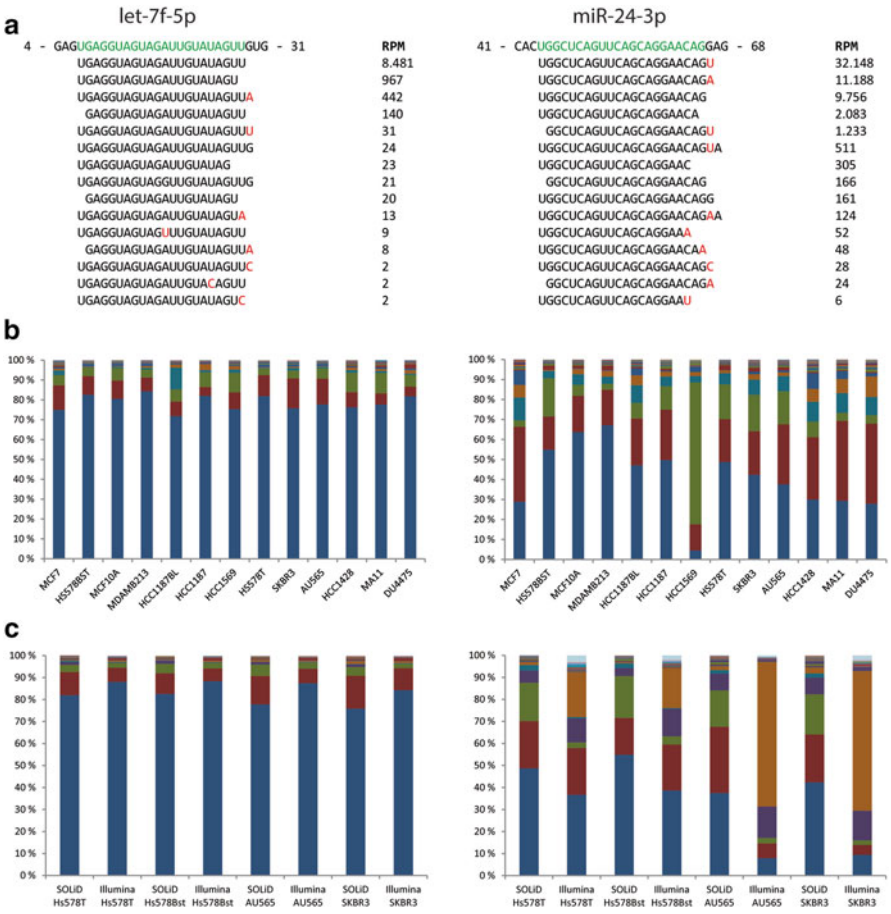


Fig. 3 Identification of isomiRs by SOLiD and Illumina NGS platforms. (**a**) Normal breast cell line (Hs 578Bst) was sequenced by SOLiD small RNA-Seq and the 15 highest expressed unique sequences mapping to two miRNAs species are displayed (let-7f-5p and miR-24-3p). While the most abundant sequence for let-7f-5p is the canonical miRNAs sequence, an isomiR containing a non-template U-nucleotide at the 3′ end is the most abundant miR-24-3p sequence. (**b**) The experiment above was extended to a total of thirteen human breast cell lines. Distribution of the same miRNA sequences and corresponding isomiRs is shown in bar plots. Identical color represents the same unique sequence in all cell lines. The expression has been scaled in order to accommodate for different expression levels in the cell lines. While let-7f-5p (*left*) shows a similar distribution pattern of canonical and isomiR sequences in all cell lines, miR-24-3p (*right*) variants varies among cell lines. (**c**) Comparison between SOLiD and Illumina platforms. Four of the cell lines were also profiled by the Illumina platform and let-7f-5p (*left*) and miR-24-3p (*right*) variants are presented. Let-7f-5p shows a similar distribution of the canonical and isomiR variants in all cell lines in both platforms. miR-24-3p, however, shows significant variation between platforms and cell lines. In general, a more uniformly distribution of variants is noted within a platform than between

might be library preparation artifacts or sequencing errors. However, there are reports that clearly demonstrate that isomiR expression patterns differ between cell types and developmental stages [67–73].

To assess the relevancy of performing isomiR profiling in cancer, we subjected 14 breast cancer cell lines to SOLiD small RNA-seq and compared the expression patterns of miRNA variants. Intriguingly, the distribution of isomiRs differs consistently and significantly among cell lines (Fig. 3b). We then compared isomiR profiles generated from two different NGS platforms, SOLiD and Illumina. The sensitivity and accuracy were found to be similar for isomiRs as for the canonical mature miRNAs in these experiments [35]. However, absolute values of the different isomiRs for a single miRNA are not constant between the two platforms (Fig. 3c). Consequently, different isoforms appear as the most abundant canonical isoform in the two platforms. This is an important issue for the classification of the true mature miRNA. If different isoform compositions in cancer cells affect the stability or mRNA targeting specificity of certain miRNAs, remain to be unraveled.

3 MiRNA Normalization

3.1 Introduction

Data normalization is essential for obtaining accurate results in miRNA profiling. The procedure adjusts data in order to reduce or remove technical biases introduced either by variations in input and quality of template, or during profiling. Optimized normalization allows for better identification of true biological differences across samples. However, some of the discrepancies noted both between, but also within, miRNA-profiling studies are probably due to the different choices in normalization methods [35]. Here, there is an urgent need for more comparison studies to unravel the effect of applying different normalizations methods across different technologies.

Today there are several methods to choose from for normalization, and the methods can be divided into two main categories, either as reference normalization or as global expression normalization. Often a broad knowledge of statistics is needed in order to understand the methods and how they transform the data generated by profiling. Here, we give a brief overview of the most common normalization methods, and also emphasizes some of the pitfalls that may accompany them.

3.2 Normalization Methods

3.2.1 Reference Normalization

The reference normalization strategy utilizes RNA species that are universally stable and do not vary in expression among the stages or tissues under investigation. A good reference RNA should be consistently stable and highly abundant

independent of tissue types or treatments. In addition, it should have characteristics similar to miRNAs in respect to size, biogenesis, and stability.

If possible, authentic miRNA references should be used for normalization. MiRNA expression is highly cancer specific, and therefor miRNAs used for normalization has to be established for each cancer type. For breast cancer, a few miRNA species have been proposed, including miR-16, miR-425, and let-7a [74, 75]. However, it has proven difficult to find suitable miRNA candidates that are stably expressed, and the method has to be used with caution.

Other alternatives include housekeeping gene transcripts as well as large stable RNAs such as ribosomal RNAs (rRNAs). In mRNA profiling by RT-qPCR it has become a standard to use the geometric averaging of multiple internal control gene transcripts as a normalization factor [76]. This method could also be adapted for miRNA profiling. Another option is to use the small nuclear RNAs (snRNAs) or small nucleolar RNAs (snoRNAs) for normalization of miRNA data sets [77]. The small RNAs RNU6B, RNU44, and RNU48 are some of the mostly described normalization references in miRNA profiling [78], and the snoRNAs are often included in pre-fabricated RT-qPCR panels. However, using large reference RNAs, including the snRNAs and snoRNAs, are not compatible when analyzing miRNA-enriched samples where larger RNA transcripts are absent, as for the NGS platforms.

Synthetic spike transcripts are good alternatives to endogenous expressed RNA references. Here, a spike RNA is pre-aliquoted in experimental samples according to for example total RNA concentration, cell number, or serum amount. Synthetic spikes that mimic miRNAs have important advantages in neutralizing technical biases introduced in library preparation. An option is to use synthetic miRNA from a heterologous species, like the *C. elegans* miRNAs *cel-miR-39*, *cel-miR-54*, and *cel-miR-238* [79]. However, estimating the correct amount of spike to use can be a challenge, especially for tissue samples where cell numbers are difficult to estimate.

3.2.2 Global Expression Normalization

The global expression normalization strategy has become the preferred methods of choice for high-throughput profiling based on microarray and NGS [80]. Global normalization is based on the assumption that the total expression of RNA species investigated does not change significantly between samples. This assumption will only be valid if a large amount of genes are being investigates, if an equal number of genes in a sample are being upregulated as downregulated, and if only a small fraction of genes in a sample are regulated [81]. In human miRNA profiling, however, several of these criteria are not well accommodated. The total number of miRNAs under investigation is less than 3000, and often the majority is weakly or even not expressed [82]. In addition, the total expression level of miRNAs has been shown to be significantly reduced in cell lines and in cancer cells compared to normal tissue [13].

The global expression normalization strategy includes both straightforward approaches based on the total, mean, or median expression value [81, 83], and more complex algorithms such as Lowess normalization, quantile normalization, rank invariant normalization, or TMM [84–87]. Global expression normalization is the most common strategy in NGS (total linear count scaling). In total linear count scaling each expression value is divided by the total expression value for the sample that is multiplied with a common factor (e.g., 1,000,000 for reads per million). For RT-qPCR, global mean normalization has been proposed as the method of choice [83]. In global mean normalization, the arithmetic average Cq value is calculated for each individual sample and subsequently subtracted from each individual Cq value for that sample. This procedure results in normalized expression values in the log2 scale.

Quantile normalization and Lowess normalization are popular methods in miRNA microarray-based profiling. In quantile normalization, the data sets are first sorted according to expression before normalized values are set accordingly to the arithmetical mean of the distributions. The highest value in all datasets is transformed to the mean of the highest values, the second highest value is transformed to the mean of the second highest values, and so on. The Lowess (locally weighted scatter plot smooth) normalization transforms expression values based on a regression weight function for all neighboring miRNA data points within a predefined span in absolute values. Thus, Lowess does not include the entire dataset compared to most other global normalization methods.

4 Conclusion

While miRNA profiling studies are rapidly increasing in the scientific literature, miRNA signatures for specific diseases are lagging behind. It is therefore an urgent need for a better understanding in handling large dataset and how to interpret results obtained from different technologies. A major concern is the use of absolute values in miRNA profiling studies, as apparently none of the existing platforms can correctly identify exact expression values. In addition, the complexity of NGS-based miRNA profiles is expanding due to the inclusion of isomiR variants. The biological role of isomiRs in cells and tissue is still obscure and more basic research is needed to establish practical solutions on how to interpret the massive number of unique miRNAs sequences profiled. An important challenge in miRNA profiling is to develop affordable high-throughput platforms, which are both highly sensitive, accurate, and with short handling time. The inclusion of the new benchtop sequencing platforms in miRNA performance comparisons studies will be an important step towards increasing their performance in regard to the more well-established methods.

References

1. Lee RC, Feinbaum RL, Ambros V. The C. elegans heterochronic gene lin-4 encodes small RNAs with antisense complementarity to lin-14. Cell. 1993;75(5):843–54.
2. Kim VN, Han J, Siomi MC. Biogenesis of small RNAs in animals. Nat Rev Mol Cell Biol. 2009;10(2):126–39.
3. Pasquinelli AE. MicroRNAs and their targets: recognition, regulation and an emerging reciprocal relationship. Nat Rev Genet. 2012;13(4):271–82.
4. Griffiths-Jones S. The microRNA registry. Nucleic Acids Res. 2004;32(Database issue): D109–11.
5. Griffiths-Jones S, Grocock RJ, van Dongen S, Bateman A, Enright AJ. miRBase: microRNA sequences, targets and gene nomenclature. Nucleic Acids Res. 2006;34(Database issue): D140–4.
6. Griffiths-Jones S, Saini HK, van Dongen S, Enright AJ. miRBase: tools for microRNA genomics. Nucleic Acids Res. 2008;36(Database issue):D154–8.
7. Kozomara A, Griffiths-Jones S. miRBase: integrating microRNA annotation and deep-sequencing data. Nucleic Acids Res. 2011;39(Database issue):D152–7.
8. Kozomara A, Griffiths-Jones S. miRBase: annotating high confidence microRNAs using deep sequencing data. Nucleic Acids Res. 2014;42(Database issue):D68–73.
9. Di Leva G, Garofalo M, Croce CM. MicroRNAs in cancer. Annu Rev Pathol. 2014;9: 287–314.
10. Jansson MD, Lund AH. MicroRNA and cancer. Mol Oncol. 2012;6(6):590–610.
11. Iorio MV, Croce CM. MicroRNA dysregulation in cancer: diagnostics, monitoring and therapeutics. A comprehensive review. EMBO Mol Med. 2012;4(3):143–59.
12. Calin GA, Dumitru CD, Shimizu M, Bichi R, Zupo S, Noch E, Aldler H, Rattan S, Keating M, Rai K, Rassenti L, Kipps T, Negrini M, Bullrich F, Croce CM. Frequent deletions and down-regulation of micro-RNA genes miR15 and miR16 at 13q14 in chronic lymphocytic leukemia. Proc Natl Acad Sci U S A. 2002;99(24):15524–9.
13. Lu J, Getz G, Miska EA, Alvarez-Saavedra E, Lamb J, Peck D, Sweet-Cordero A, Ebert BL, Mak RH, Ferrando AA, Downing JR, Jacks T, Horvitz HR, Golub TR. MicroRNA expression profiles classify human cancers. Nature. 2005;435(7043):834–8.
14. Volinia S, Calin GA, Liu CG, Ambs S, Cimmino A, Petrocca F, Visone R, Iorio M, Roldo C, Ferracin M, Prueitt RL, Yanaihara N, Lanza G, Scarpa A, Vecchione A, Negrini M, Harris CC, Croce CM. A microRNA expression signature of human solid tumors defines cancer gene targets. Proc Natl Acad Sci U S A. 2006;103(7):2257–61.
15. Martello G, Rosato A, Ferrari F, Manfrin A, Cordenonsi M, Dupont S, Enzo E, Guzzardo V, Rondina M, Spruce T, Parenti AR, Daidone MG, Bicciato S, Piccolo S. A MicroRNA targeting dicer for metastasis control. Cell. 2010;141(7):1195–207.
16. Croce CM. Causes and consequences of microRNA dysregulation in cancer. Nat Rev Genet. 2009;10(10):704–14.
17. Blenkiron C, Goldstein LD, Thorne NP, Spiteri I, Chin SF, Dunning MJ, Barbosa-Morais NL, Teschendorff AE, Green AR, Ellis IO, Tavaré S, Caldas C, Miska EA. MicroRNA expression profiling of human breast cancer identifies new markers of tumor subtype. Genome Biol. 2007;8(10):R214.
18. Qian B, Katsaros D, Lu L, Preti M, Durando A, Arisio R, Mu L, Yu H. High miR-21 expression in breast cancer associated with poor disease-free survival in early stage disease and high TGF-beta1. Breast Cancer Res Treat. 2009;117(1):131–40.
19. Zhu S, Wu H, Wu F, Nie D, Sheng S, Mo YY. MicroRNA-21 targets tumor suppressor genes in invasion and metastasis. Cell Res. 2008;18(3):350–9.
20. Mattie MD, Benz CC, Bowers J, Sensinger K, Wong L, Scott GK, Fedele V, Ginzinger D, Getts R, Haqq C. Optimized high-throughput microRNA expression profiling provides novel biomarker assessment of clinical prostate and breast cancer biopsies. Mol Cancer. 2006;5:24.

21. Jiang Q, Wang Y, Hao Y, Juan L, Teng M, Zhang X, Li M, Wang G, Liu Y. miR2Disease: a manually curated database for microRNA deregulation in human disease. Nucleic Acids Res. 2009;37(Database issue):D98–104.
22. Cortez MA, Bueso-Ramos C, Ferdin J, Lopez-Berestein G, Sood AK, Calin GA. MicroRNAs in body fluids–the mix of hormones and biomarkers. Nat Rev Clin Oncol. 2011;8(8):467–77.
23. Chen Y, Zhang L, Hao Q. Candidate microRNA biomarkers in human epithelial ovarian cancer: systematic review profiling studies and experimental validation. Cancer Cell Int. 2013;13(1):86.
24. Pritchard CC, Cheng HH, Tewari M. MicroRNA profiling: approaches and considerations. Nat Rev Genet. 2012;13(5):358–69.
25. Geiss GK, Bumgarner RE, Birditt B, Dahl T, Dowidar N, Dunaway DL, Fell HP, Ferree S, George RD, Grogan T, James JJ, Maysuria M, Mitton JD, Oliveri P, Osborn JL, Peng T, Ratcliffe AL, Webster PJ, Davidson EH, Hood L, Dimitrov K. Direct multiplexed measurement of gene expression with color-coded probe pairs. Nat Biotechnol. 2008;26(3):317–25.
26. Git A, Dvinge H, Salmon-Divon M, Osborne M, Kutter C, Hadfield J, Bertone P, Caldas C. Systematic comparison of microarray profiling, real-time PCR, and next-generation sequencing technologies for measuring differential microRNA expression. RNA. 2010; 16(5):991–1006.
27. Chugh P, Dittmer DP. Potential pitfalls in microRNA profiling. Wiley Interdiscip Rev RNA. 2012;3(5):601–16.
28. Ståhlberg A, Kubista M, Pfaffl M. Comparison of reverse transcriptases in gene expression analysis. Clin Chem. 2004;50(9):1678–80.
29. Kitchen RR, Kubista M, Tichopad A. Statistical aspects of quantitative real-time PCR experiment design. Methods. 2010;50(4):231–6.
30. Creighton CJ, Reid JG, Gunaratne PH. Expression profiling of microRNAs by deep sequencing. Brief Bioinform. 2009;10(5):490–7.
31. Hafner M, Renwick N, Brown M, Mihailović A, Holoch D, Lin C, Pena JT, Nusbaum JD, Morozov P, Ludwig J, Ojo T, Luo S, Schroth G, Tuschl T. RNA-ligase-dependent biases in miRNA representation in deep-sequenced small RNA cDNA libraries. RNA. 2011; 17(9):1697–712.
32. Bissels U, Wild S, Tomiuk S, Holste A, Hafner M, Tuschl T, Bosio A. Absolute quantification of microRNAs by using a universal reference. RNA. 2009;15(12):2375–84.
33. Stokowy T, Eszlinger M, Świerniak M, Fujarewicz K, Jarząb B, Paschke R, Krohn K. Analysis options for high-throughput sequencing in miRNA expression profiling. BMC Res Notes. 2014;7:144.
34. Tam S, de Borja R, Tsao MS, McPherson JD. Robust global microRNA expression profiling using next-generation sequencing technologies. Lab Invest. 2014;94(3):350–8.
35. Knutsen E, Fiskaa T, Ursvik A, Jørgensen TE, Perander M, Lund E, Seternes OM, Johansen SD, Andreassen M. Performance comparison of digital microRNA profiling technologies applied on human breast cancer cell lines. PLoS One. 2013;8(10):e75813.
36. Kolbert CP, Feddersen RM, Rakhshan F, Grill DE, Simon G, Middha S, Jang JS, Simon V, Schultz DA, Zschunke M, Lingle W, Carr JM, Thompson EA, Oberg AL, Eckloff BW, Wieben ED, Li P, Yang P, Jen J. Multi-platform analysis of microRNA expression measurements in RNA from fresh frozen and FFPE tissues. PLoS One. 2013;8(1):e52517.
37. Leshkowitz D, Horn-Saban S, Parmet Y, Feldmesser E. Differences in microRNA detection levels are technology and sequence dependent. RNA. 2013;19(4):527–38.
38. Pradervand S, Weber J, Lemoine F, Consales F, Paillusson A, Dupasquier M, Thomas J, Richter H, Kaessmann H, Beaudoing E, Hagenbüchle O, Harshman K. Concordance among digital gene expression, microarrays, and qPCR when measuring differential expression of microRNAs. Biotechniques. 2010;48(3):219–22.
39. Fehniger TA, Wylie T, Germino E, Leong JW, Magrini VJ, Koul S, Keppel CR, Schneider SE, Koboldt DC, Sullivan RP, Heinz ME, Crosby SD, Nagarajan R, Ramsingh G, Link DC, Ley TJ. Mardis ER (2010) Next-generation sequencing identifies the natural killer cell microRNA transcriptome. Genome Res. 2010;20(11):1590–604.

40. Willenbrock H, Salomon J, Søkilde R, Barken KB, Hansen TN, Nielsen FC, Møller S, Litman T. Quantitative miRNA expression analysis: comparing microarrays with next-generation sequencing. RNA. 2009;15(11):2028–34.
41. Linsen SE, de Wit E, Janssens G, Heater S, Chapman L, Parkin RK, Fritz B, Wyman SK, de Bruijn E, Voest EE, Kuersten S, Tewari M, Cuppen E. Limitations and possibilities of small RNA digital gene expression profiling. Nat Methods. 2009;6(7):474–6.
42. Toedling J, Servant N, Ciaudo C, Farinelli L, Voinnet O, Heard E, Barillot E. Deep-sequencing protocols influence the results obtained in small-RNA sequencing. PLoS One. 2012;7(2):e32724.
43. Tian G, Yin X, Luo H, Xu X, Bolund L, Zhang X, Gan SQ, Li N. Sequencing bias: comparison of different protocols of microRNA library construction. BMC Biotechnol. 2010;10:64.
44. Etheridge A, Lee I, Hood L, Galas D, Wang K. Extracellular microRNA: a new source of biomarkers. Mutat Res. 2011;717(1–2):85–90.
45. Kosaka N, Iguchi H, Ochiya T. Circulating microRNA in body fluid: a new potential biomarker for cancer diagnosis and prognosis. Cancer Sci. 2010;101(10):2087–92.
46. Chen X, Ba Y, Ma L, Cai X, Yin Y, Wang K, Guo J, Zhang Y, Chen J, Guo X, Li Q, Li X, Wang W, Zhang Y, Wang J, Jiang X, Xiang Y, Xu C, Zheng P, Zhang J, Li R, Zhang H, Shang X, Gong T, Ning G, Wang J, Zen K, Zhang J, Zhang CY. Characterization of microRNAs in serum: a novel class of biomarkers for diagnosis of cancer and other diseases. Cell Res. 2008;18(10):997–1006.
47. Etheridge A, Gomes CP, Pereira RW, Galas D, Wang K. The complexity, function and applications of RNA in circulation. Front Genet. 2013;4:115.
48. Turchinovich A, Weiz L, Langheinz A, Burwinkel B. Characterization of extracellular circulating microRNA. Nucleic Acids Res. 2011;39(16):7223–33.
49. Turchinovich A, Weiz L, Burwinkel B. Extracellular miRNAs: the mystery of their origin and function. Trends Biochem Sci. 2012;37(11):460–5.
50. von Ahlfen S, Missel A, Bendrat K, Schlumpberger M. Determinants of RNA quality from FFPE samples. PLoS One. 2007;2(12):e1261.
51. Doleshal M, Magotra AA, Choudhury B, Cannon BD, Labourier E, Szafranska AEJ. Evaluation and validation of total RNA extraction methods for microRNA expression analyses in formalin-fixed, paraffin-embedded tissues. Mol Diagn. 2008;10(3):203–11.
52. Xi Y, Nakajima G, Gavin E, Morris CG, Kudo K, Hayashi K, Ju J. Systematic analysis of microRNA expression of RNA extracted from fresh frozen and formalin-fixed paraffin-embedded samples. RNA. 2007;13(10):1668–74.
53. Hall JS, Taylor J, Valentine HR, Irlam JJ, Eustace A, Hoskin PJ, Miller CJ, West CM. Enhanced stability of microRNA expression facilitates classification of FFPE tumour samples exhibiting near total mRNA degradation. Br J Cancer. 2012;107(4):684–94.
54. Torres L, Ribeiro FR, Pandis N, Andersen JA, Heim S, Teixeira MR. Intratumor genomic heterogeneity in breast cancer with clonal divergence between primary carcinomas and lymph node metastases. Breast Cancer Res Treat. 2007;102(2):143–55.
55. Navin N, Kendall J, Troge J, Andrews P, Rodgers L, McIndoo J, Cook K, Stepansky A, Levy D, Esposito D, Muthuswamy L, Krasnitz A, McCombie WR, Hicks J, Wigler M. Tumour evolution inferred by single-cell sequencing. Nature. 2011;472(7341):90–4.
56. Russnes HG, Navin N, Hicks J, Borresen-Dale AL. Insight into the heterogeneity of breast cancer through next-generation sequencing. J Clin Invest. 2011;121(10):3810–8.
57. Marusyk A, Almendro V, Polyak K. Intra-tumour heterogeneity: a looking glass for cancer? Nat Rev Cancer. 2012;12(5):323–34.
58. Wang Y, Waters J, Leung ML, Unruh A, Roh W, Shi X, Chen K, Scheet P, Vattathil S, Liang H, Multani A, Zhang H, Zhao R, Michor F, Meric-Bernstam F, Navin NE. Clonal evolution in breast cancer revealed by single nucleus genome sequencing. Nature. 2014;512(7513):155–60.
59. Jensen SG, Lamy P, Rasmussen MH, Ostenfeld MS, Dyrskjøt L, Orntoft TF, Andersen CL. Evaluation of two commercial global miRNA expression profiling platforms for detection of less abundant miRNAs. BMC Genomics. 2011;12:435.

60. Koshimizu E, Miyatake S, Okamoto N, Nakashima M, Tsurusaki Y, Miyake N, Saitsu H, Matsumoto N. Performance comparison of bench-top next generation sequencers using microdroplet PCR-based enrichment for targeted sequencing in patients with autism spectrum disorder. PLoS One. 2013;8(9):e74167.
61. van Dijk EL, Auger H, Jaszczyszyn Y, Thermes C (2014)Ten years of next-generation sequencing technology. Trends Genet. 2014 Aug 6. pii: S0168-9525(14)00112-7.
62. Guo L, Chen F. A challenge for miRNA: multiple isomiRs in miRNAomics. Gene. 2014;544(1):1–7.
63. Ameres SL, Zamore PD. Diversifying microRNA sequence and function. Nat Rev Mol Cell Biol. 2013;14(8):475–88.
64. Neilsen CT, Goodall GJ, Bracken CP. IsomiRs–the overlooked repertoire in the dynamic microRNAome. Trends Genet. 2012;28(11):544–9.
65. Lee LW, Zhang S, Etheridge A, Ma L, Martin D, Galas D, Wang K. Complexity of the microRNA repertoire revealed by next-generation sequencing. RNA. 2010;16(11):2170–80.
66. Cloonan N, Wani S, Xu Q, Gu J, Lea K, Heater S, Barbacioru C, Steptoe AL, Martin HC, Nourbakhsh E, Krishnan K, Gardiner B, Wang X, Nones K, Steen JA, Matigian NA, Wood DL, Kassahn KS, Waddell N, Shepherd J, Lee C, Ichikawa J, McKernan K, Bramlett K, Kuersten S, Grimmond SM. MicroRNAs and their isomiRs function cooperatively to target common biological pathways. Genome Biol. 2011;12(12):R126.
67. Fernandez-Valverde SL, Taft RJ, Mattick JS. Dynamic isomiR regulation in Drosophila development. RNA. 2010;16(10):1881–8.
68. Ruby JG, Stark A, Johnston WK, Kellis M, Bartel DP, Lai EC. Evolution, biogenesis, expression, and target predictions of a substantially expanded set of Drosophila microRNAs. Genome Res. 2007;17(12):1850–64.
69. Hinton A, Hunter SE, Afrikanova I, Jones GA, Lopez AD, Fogel GB, Hayek A, King CC. sRNA-seq Analysis of Human Embryonic Stem Cells and Definitive Endoderm Reveals Differentially Expressed MicroRNAs and Novel IsomiRs with Distinct Targets. Stem Cells. 2014;32(9):2360–72.
70. Swierniak M, Wojcicka A, Czetwertynska M, Stachlewska E, Maciag M, Wiechno W, Gornicka B, Bogdanska M, Koperski L, de la Chapelle A, Jazdzewski K. In-depth characterization of the microRNA transcriptome in normal thyroid and papillary thyroid carcinoma. J Clin Endocrinol Metab. 2013;98(8):E1401–9.
71. Tan GC, Chan E, Molnar A, Sarkar R, Alexieva D, Isa IM, Robinson S, Zhang S, Ellis P, Langford CF, Guillot PV, Chandrashekran A, Fisk NM, Castellano L, Meister G, Winston RM, Cui W, Baulcombe D, Dibb NJ. 5′ isomiR variation is of functional and evolutionary importance. Nucleic Acids Res. 2014;42(14):9424–35.
72. Vaz C, Ahmad HM, Bharti R, Pandey P, Kumar L, Kulshreshtha R, Bhattacharya A. Analysis of the microRNA transcriptome and expression of different isomiRs in human peripheral blood mononuclear cells. BMC Res Notes. 2013;6:390.
73. Bizuayehu TT, Lanes CF, Furmanek T, Karlsen BO, Fernandes JM, Johansen SD, Babiak I. Differential expression patterns of conserved miRNAs and isomiRs during Atlantic halibut development. BMC Genomics. 2012;13:11.
74. Davoren PA, McNeill RE, Lowery AJ, Kerin MJ, Miller N. Identification of suitable endogenous control genes for microRNA gene expression analysis in human breast cancer. BMC Mol Biol. 2008;9:76.
75. McDermott AM, Kerin MJ, Miller N. Identification and validation of miRNAs as endogenous controls for RQ-PCR in blood specimens for breast cancer studies. PLoS One. 2013; 8(12):e83718.
76. Vandesompele J, De Preter K, Pattyn F, Poppe B, Van Roy N, De Paepe A, Speleman F (2002) Accurate normalization of real-time quantitative RT-PCR data by geometric averaging of multiple internal control genes. Genome Biol. 3(7):RESEARCH0034.
77. Benes V, Castoldi M. Expression profiling of microRNA using real-time quantitative PCR, how to use it and what is available. Methods. 2010;50(4):244–9.

78. Wotschofsky Z, Meyer HA, Jung M, Fendler A, Wagner I, Stephan C, Busch J, Erbersdobler A, Disch AC, Mollenkopf HJ, Jung K. Reference genes for the relative quantification of microRNAs in renal cell carcinomas and their metastases. Anal Biochem. 2011;417(2): 233–41.
79. Brase JC, Johannes M, Schlomm T, Fälth M, Haese A, Steuber T, Beissbarth T, Kuner R, Sültmann H. Circulating miRNAs are correlated with tumor progression in prostate cancer. Int J Cancer. 2011;128(3):608–16.
80. Garmire LX, Subramaniam S. Evaluation of normalization methods in mammalian microRNA-Seq data. RNA. 2012;18(6):1279–88.
81. D'haene B, Mestdagh P, Hellemans J, Vandesompele J. miRNA expression profiling: from reference genes to global mean normalization. Methods Mol Biol. 2012;822:261–72.
82. Pradervand S, Weber J, Thomas J, Bueno M, Wirapati P, Lefort K, Dotto GP, Harshman K. Impact of normalization on miRNA microarray expression profiling. RNA. 2009; 15(3):493–501.
83. Mestdagh P, Van Vlierberghe P, De Weer A, Muth D, Westermann F, Speleman F, Vandesompele J. A novel and universal method for microRNA RT-qPCR data normalization. Genome Biol. 2009;10(6):R64.
84. Smyth GK, Yang YH, Speed T. Statistical issues in cDNA microarray data analysis. Methods Mol Biol. 2003;224:111–36.
85. Bolstad BM, Irizarry RA, Astrand M, Speed TP. A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. Bioinformatics. 2003; 19(2):185–93.
86. Li C, Hung Wong W (2001) Model-based analysis of oligonucleotide arrays: model validation, design issues and standard error application. Genome Biol. 2(8):RESEARCH0032.
87. Robinson MD, Oshlack A. A scaling normalization method for differential expression analysis of RNA-seq data. Genome Biol. 2010;11(3):R25.

Small RNA Sequencing for Squamous Cell Carcinoma Research

Patricia Severino, Liliane Santana Oliveira, and Alan Mitchell Durham

Abstract Small noncoding RNAs are important transcriptional regulators implicated in several aspects of cell biology. Recent studies have demonstrated that qualitative and quantitative information on these molecules may be clinically useful for the cancer community. Next-Generation Sequencing is quickly becoming the technology of choice to study their role since a single equipment run may provide thorough information on small RNA populations. Expression levels, mutational status, and the opportunity to identify novel molecules are among the resources of most sequencing platforms. However, challenges include sample processing and an appropriate data analysis pipeline. This chapter describes the workflow for small RNA sequencing analysis, discussing sample preparation, small RNA alignment for quantification, differential expression analysis, and novel small RNA molecule identification. We are currently studying small noncoding RNAs in head and neck squamous cell carcinoma (HNSCC). Arising from epithelial cells in the lining of the upper aerodigestive tract and strongly associated with tobacco and alcohol consumption, HNSCC is among the leading cancers by incidence worldwide. Information on the relevance of small RNA molecules for this cancer type is still scarce but may become useful for diagnostic and prognostic purposes when comprehensive datasets become available.

P. Severino (✉) • L.S. Oliveira
Albert Einstein Research and Education Institute, Hospital Israelita Albert Einstein, Sao Paulo, SP, Brazil
e-mail: patricia.severino@einstein.br; liliane.oliveira@einstein.br

A.M. Durham
Instituto de Matemática e Estatística, Universidade de Sao Paulo, Sao Paulo, SP, Brazil
e-mail: aland@usp.br

W. Wu, H. Choudhry (eds.), *Next Generation Sequencing in Cancer Research, Volume 2*, DOI 10.1007/978-3-319-15811-2_16

1 Small Noncoding RNAs and the Contribution of Next-Generation Sequencing for Studying Their Role in Cancer

Noncoding RNA molecules vary greatly in size, ranging from a diverse range of long noncoding RNAs, to smaller molecules currently called *small noncoding RNAs*. Small noncoding RNAs are regulatory molecules of ~18–30 nucleotides in length. They have emerged as important players in several aspects of cellular biology, mostly acting through the inactivation of complementary sequences. MicroRNAs (miRNAs), small interfering RNAs (siRNAs), and PIWI interacting RNAs (piRNA), for instance, are all involved in sequence-specific posttranscriptional gene silencing as well as chromatin-dependent gene silencing, although they may vary in the subcellular location where they act (for a review see ref. [1]).

The advent of Next-Generation Sequencing (NGS) technologies has greatly contributed to speed up this research area [2]. Such efforts have allowed the identification of several other classes of small RNAs, some already shown to participate in biological processes, while others are still not fully understood. Among these, tncRNA (tiny noncoding RNAs), rasiRNA (repeat-associated siRNA), hcRNA (heterochromatic small RNA), scnRNA (scan RNA), tiRNA (transcription initiation RNA), PASR/TASR (promoter/termini-associated sRNA), easRNA (exon-associated small RNA), rasRNA (repeat-associated small RNA), and tRFs (tRNA-derived RNA fragments) can be mentioned [3–5].

The depth of the biological significance of small RNA functional classes in cancer is also still unclear, but gene expression in cancer is known to be controlled by numerous regulatory molecules, including small regulatory RNAs. MiRNAs are currently the most studied class of small RNAs implicated in the pathogenesis of this disease [6, 7]. Experimental evidence points to the deregulation of miRNAs as consequence of cancer progression as well to their involvement in mechanisms leading to tumor initiation and progression. Additionally, miRNA profiles of human malignancies have been addressed as potential biomarkers in cancer patient management and as drug targets or therapeutics [8].

Since 2006, small RNAs bound to PIWI proteins have been purified and identified, the piRNAs [9]. PiRNAs are interacting RNAs ranging from 25 to 31 nt in length, thus longer than miRNAs and siRNAs. They interact with PIWI proteins, a subclass of the Argonaute family of proteins, originally found exclusively in germline cells. Considering that cancer cells and germ cells share characteristics (e.g., rapid proliferation and self-renewal) it is expected that germ-line factors would also be implicated in oncogenesis. In spite of the study of PIWI in cancer being relatively new, the expression of PIWI has already been detected in a variety of cancers and even associated with cancer prognosis [10, 11]. On the other hand, studies on the contribution of piRNAs to tumorigenesis or their function in cancer cells are still rare, some of which addressing the epigenetic functions of PIWI/piRNA complexes [12].

Recently, the human genome-wide analysis carried out by The Cancer Genome Atlas (http://cancergenome.nih.gov/) showed that changes in the expression levels of small noncoding RNAs near the transcription start site of genes is associated with disease and could be considered for diagnostics purposes [13].

Head and neck squamous cell carcinomas (HNSCC) arise from epithelial cells in the lining of the upper aerodigestive tract [14]. The most important risk factors are tobacco and alcohol consumption [15]. HNSCC is the sixth leading cancer by incidence worldwide, with a 5-year survival rate of about 50 % and no prognostic biomarkers or molecular markers for early diagnosis [16]. Among small RNA molecules, miRNAs have been evaluated as potential biomarkers with clinical application in HNSCC, with expression levels associated to survival rates or metastatic potential, as well as to tumorigenesis and tumor progression [17, 18]. To our knowledge there is currently no data on a possible role for other small RNAs for this cancer type.

We recently used NGS of small RNAs to study their potential role in HNSCC of the oral cavity and biological features targeted by miRNAs in cell models used for HNSCC research [19]. In this chapter we share our views on this research approach and describe a workflow for small RNA sequencing analysis, including sample preparation, small RNA quantification, differential expression analysis, and novel small RNA molecule identification.

2 Sample Processing for Small RNA Sequencing

The first step when looking to perform next-generation sequencing of small RNA libraries is choosing the appropriate method for RNA isolation. MicroRNAs range from 16 to 27 nucleotides, while PIWI interacting RNAs (piRNA) are about 30 nucleotides long, thus a method for RNA isolation that allows for the sequencing of the two RNA classes must preserve molecules ranging from 16 to 30 nucleotides.

Commercially available column-based kits for total RNA isolation from tissues or cells claim to recover all RNA types, but recent research addressing small RNA molecules have demonstrated that this assumption is untrue. The majority of these methods are based on solutions containing guanidinium thiocyanate for cell lysis and protein denaturation, followed by phenol/chloroform extraction to isolate RNA molecules. In the particular case of HNSCC, we obtained satisfactory results with commercially available kits, including miRNeasy Kit (Qiagen, USA), mirVana Kit (Life Technologies, USA), and mirVana PARIS Kit (Life Technologies, USA). However, in the general case, researchers should be aware of potential problems with other available solutions. First, in several of the column-based kits for RNA extraction the small RNA population is washed off the column during the washing steps. Second, the proper salt/alcohol ratio is an issue for small RNA precipitation. Third, the intrinsic characteristics of the tissue type, which can be more or less fibrous or lipid-rich, may demand adaptations throughout the procedures. Concentration and quality of the obtained RNA as well as the percentage of the small RNA population in the sample to be sequenced should always be assessed. Noteworthy is the fact that, despite similarities in the protocols and claims, the performance will vary between each kit and, for the purpose of a research project, all samples should be treated with the same procedure for the sake of minimizing variability in the results.

3 Small RNA Library Construction and Sequencing

Once RNA samples are obtained, library construction usually follows manufacturer's protocols strictly for better standardization and reproducibility of sequencing results. In brief, for each small RNA sample it involves 3′ adapter ligation, 5′ RT primer annealing, 5′ adapter ligation, reverse transcription, and PCR amplification. In one of our works [19] we aimed to access differences and similarities between cancer cell types used for functional studies in HNSCC and clinical samples. Three libraries were constructed for each cell type (i.e., technical replicates), and a single library was constructed for each clinical sample, constituting the biological replicates of tumor and tumor-free samples.

The sequencing throughput of current NGS platforms greatly exceeds what is necessary to quantify the small RNA population of a single sample. A way of reducing costs is increasing scale by multiplexing multiple clinical samples in a single run. Multiplexing is achieved by attaching a specific sequence tag, a *barcode*, to each sample before combining them for sequencing. Considering the library construction protocol briefly described above, the forward PCR primer is the same for every sample but a different, sample-specific, reverse PCR primer, containing a unique barcode, is used for each small RNA sample. Due to the small sizes of the RNA molecules and platform throughput, we chose the SOLiD platform (*Sequencing by Oligonucleotide Ligation and Detection*, Applied Biosystems), generating 35 nucleotides-long reads of for small RNA sequencing and an output of up to 30 Gb per run, and the SOLiD Total RNA-Seq Kit for Small RNA Libraries protocol (Ambion, Life Technologies, USA). For sample multiplexing we used the SOLiD RNA Barcoding System (Ambion, Life Technologies, USA), in which predefined sequence-tags present uniform melting temperature and are unique in *color-space.*[1] The barcodes are added to the 3′ end of the target sequence using a modified adapter, assigning a unique identifier to templates made from a single library. Multiple batches of templates are then pooled together for the PCR step and then sequenced.

Other sequencing platforms use similar workflows: Small RNA Sample Preparation Kit for Illumina; Ion Total RNA-Seq Kit for Ion PGM System or Ion Proton System, Life Technologies. However, library construction and/or multiplexing can also be accomplished using reagents from independent companies. At the time of this publication, for example, KappaBiosystems produced Illumina and Ion compatible reagents, and BiooScientific was commercializing Illumina, SOLiD, and Ion compatible reagents, constituting important alternatives in terms of sample input, costs, timing, and optimization procedures.

[1] Color-space is a characteristic output format of the Applied Biosystems' SOLiD sequencing platform that is better characterized later in this chapter.

4 Investigation of miRNA Role in HNSCC: Identification, Quantification and Differential Expression

Sequence data analysis in our research study followed the workflow described in Fig. 1. The initial material for data analysis is the *raw sequences* or *raw reads*. The format for describing raw reads depends on the sequencing platform. Illumina and Ion Torrent or Proton sequencers, for example, generate traditional FASTA files, while SOLiD sequencers generate data in XSQ or CSFASTA formats. In particular, the SOLiD technology we used generates read data in *color-space*. Color-space is a raw data type described as a sequence of colors where each color represents two possible nucleotides (i.e., a dinucleotide) and each nucleotide must be represented by two consecutive colors. The fact that each nucleotide is read twice leads to one of the advantages of color-space data: the increased ability to distinguish polymorphisms from sequencing errors. Despite this advantage, this technology requires the development and implementation of specific algorithms to map the color-space reads to genomes.

For the detection of miRNAs in color-space sequences we used the Small RNA Analysis Tool (RNA2MAP; implemented within the LifeScope Genomic Analysis Software) [20]. RNA2MAP maps the reads from 5′ to 3′ end by extending an initial

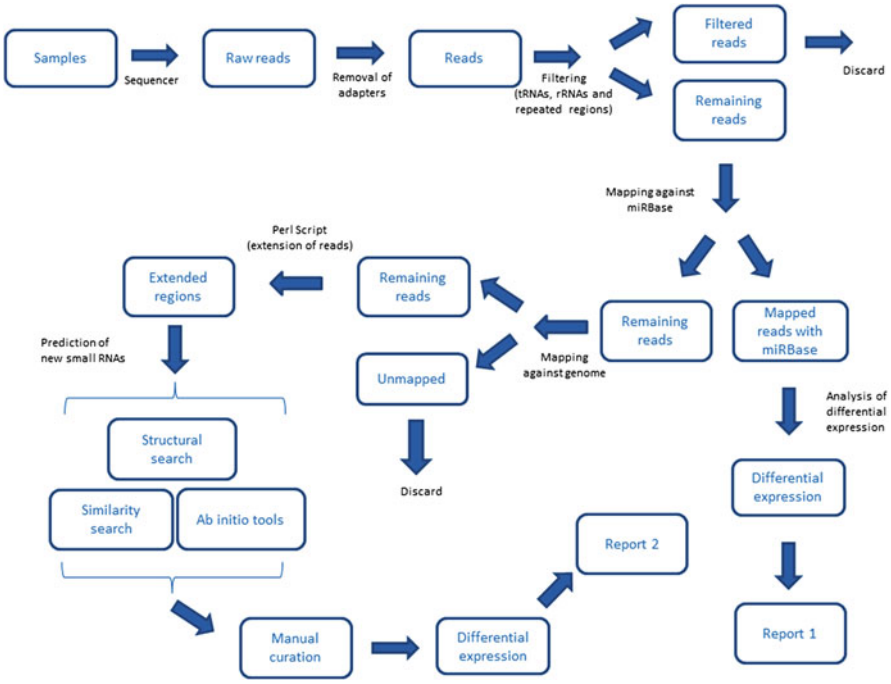


Fig. 1 Data analysis workflow for small RNA sequencing results. MiRNAs were annotated using the most recent miRBAse database. *Report 1* provides data on differentially expressed miRNAs. *Report 2* reports data on putative novel miRNA molecules based on in silico prediction

aligned "seed sequence" using hypothetical reads made of concatenating genome fragments and adapter sequences. RNA2MAP does not prefilter sequences based on quality values since color-space allows the identification of errors in color calls. We used RNA2MAP default parameters, which worked well for our purposes: three color-space mismatches within the "seed sequence" (first 18 bases of the reads), and six color-space mismatches on the following positions of the read as mapping parameters.

Adapters were trimmed out using a proprietary algorithm within LifeScope, but for other sequencing technologies algorithms such as AdapterRemoval [21], cutadapt [22], btrim [23] and ConDeTri [24] may be useful. Then, we filtered out rRNA and tRNA and repeated sequences using BLAST similarity search. To perform miRNA identification we matched the reads against a .gff annotation file of the mature miRNA sequences deposited in the miRBase database (http://www.mirbase.org). It is important to note that the LifeScope analysis tools allow the use of other miRNA annotation files, provided they are described in the .gff format.

During miRNA identification, reads may map uniquely to a miRNA sequence within miRBase or show multiple hits. We restricted multiple hits to hits within variations of a single miRNA family. For example, reads mapping identically to hsa-mir-103a-1 and hsa-mir-103a-2 were accepted and they were counted as "hsa-mir-103a" for the sake of facilitating quantification and downstream functional analysis issues.

To compare the expression levels of miRNAs between our datasets we used the output *Report 1* depicted in Fig. 1. It contains a list of miRNAs and the correspondent read counts. The first step for this analysis is normalization between samples. Several normalization procedures have been proposed in the literature so far, but no consensus has been reached [25, 26]. The normalized expression of each mature miRNA is then considered for the selection of differentially expressed miRNAs between datasets using statistical packages such as edgeR [27] and DESeq [28] implemented within the R-Project.

In our HNSCC study, we used three datasets: HNSCC-derived cell line, normal oral keratinocytes, and clinical samples. This analysis resulted in a set of miRNAs expressed in similar levels in the three datasets and in a set of miRNAs expressed in different levels. In order to evaluate the possible role of these miRNAs, functional analysis using Gene Ontology (GO) term enrichment analysis of miRNA targets is a valuable approach. MiRNA targets may be identified using databases that report predicted targets such as TargetScan (http://www.targetscan.org/) and PicTar (http://pictar.mdc-berlin.de/). Results for miRNA target prediction vary greatly in these databases due to the algorithms used for prediction. TarBase, for instance, is based mostly on target/miRNA complementarity and seed region conservation, while PicTar uses sequence thermodynamics for target/miRNA matching. There are databases that focus on experimentally validated miRNA targets, such as TarBase (http://mirtarbase.mbc.nctu.edu.tw/) or the commercially available databases such as Ingenuity (http://www.ingenuity.com/) and MetaCore (http://thomsonreuters.com/metacore/).

With the list of miRNA targets for the differentially expressed miRNAs in each dataset, and a GO enrichment analysis we can find which GO terms are overrepresented (or underrepresented) and discuss this functional result in the light of the disease under study. The annotations we used are part of structured vocabularies put together by the GO project (http://geneontology.org) describing gene products in terms of their association with biological processes, cellular components, and molecular functions, in a species-independent manner. There are several tools freely available on the Internet to perform GO term enrichment analysis, such as: PANTHER (http://pantherdb.org/), BiNGO (http://www.psb.ugent.be/cbd/papers/BiNGO/Home.html), gPROFILER (http://biit.cs.ut.ee/gprofiler/), and DAVID (http://david.abcc.ncifcrf.gov/).

As a result of applying this workflow, we described the possible activity of miRNAs in terms of processes associated with the cancer phenotype. Several miRNAs previously described in cancer samples were shown to be ubiquitously expressed in squamous cells and mostly targeting processes associated with the biology of this cell type. Others seemed to be more associated with the cancer phenotype, with possible targets grouped under GO terms related to deregulated cell processes in cancer disease.

5 Identification of Putative New miRNA Molecules

The preparation of the sample submitted to sequencing ensures that most of the RNA molecules present are small RNAs. This means that those sequences that did not match known miRNAs could either be novel miRNA molecules or molecules that belong to other classes of small RNAs. So the next phase in our workflow was the discovery of new miRNAs, summarized in the lower left part of Fig. 1. We based this discovery process on ab initio classification, similarity search and structural search.

If our reads corresponded to mature miRNAs they should belong to precursor miRNAs that could be predicted by ab initio classification and structural search both approaches highly influenced by the nucleotide sequence and its length. Since at the time of the publication of the study approximately 98 % of mature miRNA sequences deposited in miRBAse were less than 135 nucleotides in length [19], and since our reads could in fact be from any part of the precursor sequences, we mapped our reads in the genome and extracted three candidate sequences for each read: two 135 nucleotide sequences that extended the original read an additional 100 nucleotides either at the 3′ or 5′ end, and one that extended the original read 50 nucleotides at each side. These sequences were called *miRNA-candidates.*

Prior to the miRNA discovery using ab initio classification and structural search we used a conservative false positive detection approach by excluding all miRNA-candidates with good scores with snoRNA ab initio prediction from SnoScan [29] and SnoReport [30], and candidates mapping on known exons or matching other ncRNA in additional databases such as ASRP [31], CSRDB [32], fRNAdb [33], ncRNAdb [34], and NONCODE [35].

After false positive filtering, we performed structural search against the RFAM database [36] using the Infernal for RNA alignment (INFERence of RNA Alignment, http://infernal.janelia.org/) and ab initio classification using HHMMIR, an algorithm for miRNA *de novo* prediction [37] based on the structural folding provided by RNAfold (http://rna.tbi.univie.ac.at/cgi-bin/RNAfold.cgi). Structural and ab initio classification searches are highly influenced by the nucleotide sequence and its length. Figure 2 illustrates this influence, presenting foldings of two candidate sequences resulting from different 5′ and 3′extensions of the same original read.

A total of 400 *miRNA-candidates*, constituting sequences that had not matched the miRBase, were submitted to this analysis approach. Of these, only 13 remained as putative candidates and, after careful manual curation, we were able to propose a single molecule as a possible novel miRNA molecule. The RNAfold structure of this molecule is depicted in Fig. 3, it contains 2 of the 13 candidates initially selected, each one located on a side of the miRNA precursor stem, possibly indicating leading and star strands of the miRNA. An indication of a biological role for this molecule in HNSCC comes from the fact that both reads were more expressed in the HNSCC line and in HNSCC samples when compared to normal oral keratinocytes or cancer-free patient tissues, respectively.

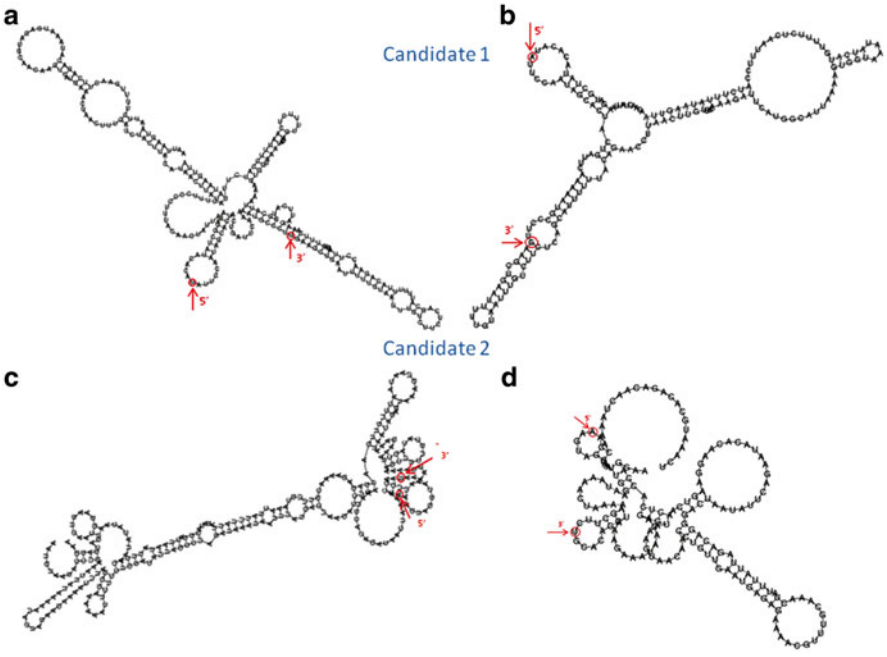


Fig. 2 RNAfold structure prediction based on nucleotide sequences of two miRNA-candidate sequences. The 35 nt sequenced reads are located between arrows in the depicted structures and the 5′ and 3′ ends of the sequenced reads are indicated by a *red circle*. **a** and **b** structures include Candidate Sequence 1, either in a central position within the structure (**a**) or at the 5′ end (**b**). **c** and **d** contain Candidate Sequence 2 either in a central position (**c**) or at the 5′ end (**d**)

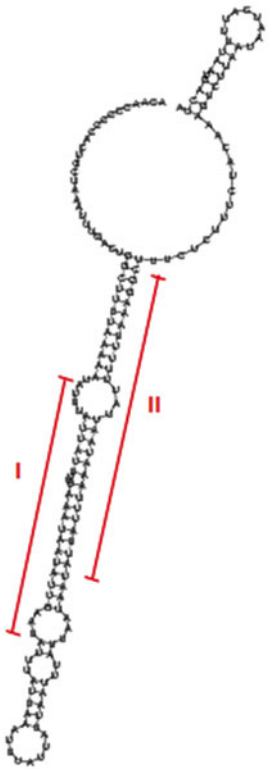


Fig. 3 RNAfold structure depicting the position of two sequenced reads (I and II selected regions in the figure) selected for miRNA discovery. Both reads mapped to neighboring regions of the genome and seem to make up the stem region of a miRNA precursor molecule

6 Other Small Noncoding RNAs Annotation and Prediction

In the work described in this chapter [19] we did not include annotation of other small RNAs, since it was not the focus of the manuscript. However, we are currently using this approach for new studies. In order to search for small RNAs other than miRNAs we take the trimmed and filtered sequences that did not match miRBase (i.e., 35 nt long sequences) and match them to noncoding RNA databases such as the ones previously mentioned in this chapter. We look for 100 % similarity between our query and the deposited sequence, as well as identical coordinates in the genome in order to consider the annotation.

7 Conclusions

In this chapter we describe a workflow for studying small RNAs using NGS. The approach has been recently used in a publication focusing on HNSCC, a prevalent cancer type for which little is known in terms of the contribution of these small

molecules for the disease phenotype. The workflow holds particularities associated with the sequencing platform of choice but it can be applied in other contexts when the research purpose is similar. The databases and bioinformatics tools mentioned in the text were useful and available at the time of this publication, but one must keep in mind that this is a new and rapidly growing research field, so adaptations will be necessary at each new research initiative.

References

1. Moazed D. Small RNAs in transcriptional gene silencing and genome defense. Nature. 2009;457(7228):412–20.
2. Morozova O, Marra MA. Applications of next-generation sequencing technologies in functional genomics. Genomics. 2008;92:255–64.
3. Taft RJ, Glazov EA, Cloonan N, Simons C, Stephen S, Faulkner GJ, Lassmann T, Forrest AR, Grimmond SM, Schroder K, et al. Tiny RNAs associated with transcription start sites in animals. Nat Genet. 2009;41:572–8.
4. Lee YS, Shibata YI, Malhotra A, Dutta A. A novel class of small RNAs: tRNA-derived RNA fragments (tRFs). Genes Dev. 2009;23(22):2639–49.
5. Kapranov P, Cheng J, Dike S, Nix DA, Duttagupta R, Willingham AT, Stadler PF, Hertel J, Hackermuller J, Hofacker IL, et al. RNA maps reveal new RNA classes and a possible function for pervasive transcription. Science. 2007;316:1484–8.
6. Calin GA, Croce CM. MicroRNA signatures in human cancers. Nat Rev Cancer. 2006;6(11):857–66.
7. Lee YS, Dutta A. MicroRNAs in cancer. Annu Rev Pathol. 2009;4:199–227.
8. Negrini MI, Nicoloso MS, Calin GA. MicroRNAs and cancer—new paradigms in molecular oncology. Curr Opin Cell Biol. 2009;3:470–9.
9. Aravin AA, Sachidanandam R, Bourc'his D, Schaefer C, Pezic D, Toth KF, et al. A piRNA pathway primed by individual transposons is linked to de novo DNA methylation in mice. Mol Cell. 2008;31:785–99.
10. Sun G, Wang Y, Sun L, Luo H, Liu N, Fu Z, et al. Clinical significance of Hiwi gene expression in gliomas. Brain Res. 2011;2011(1373):183–8.
11. Wang Y, Liu Y, Shen X, Zhang X, Chen X, Yang C, et al. The PIWI protein acts as a predictive marker for human gastric cancer. Int J Clin Exp Pathol. 2012;5:315–25.
12. Siddiqi S, Terry M, Matushansky I. Hiwi mediated tumorigenesis is associated with DNA hypermethylation. PLoS One. 2012;7:e33711. 10.1371/journal.pone.0033711.
13. Zovoilis A, Mungall AJ, Moore R, Varhol R, Chu A, Wong T, Marra M, Steven Jones SJM. The expression level of small non-coding RNAs derived from the first exon of protein coding genes is predictive of cancer status. EMBO Rep. 2014;15(4):402–10.
14. Leemans CR, Braakhuis BJ, Brakenhoff RH. The molecular biology of head and neck cancer. Nat Rev Cancer. 2011;11(1):9–22.
15. Zhang ZF, Morgenstern H, Spitz MR, Tashkin DP, Yu GP, Hsu TC, Schantz SP. Environmental tobacco smoking, mutagen sensitivity, and head and neck squamous cell carcinoma. Cancer Epidemiol Biomarkers Prev. 2000;9(10):1043–9.
16. Jemal A, Bray F, Center MM, Ferlay J, Ward E, Forman D. Global cancer statistics. CA Cancer J Clin. 2011;61(2):69–90. Erratum in: CA Cancer J Clin. 2011 61(2):134.
17. Childs G, Fazzari M, Kung G, Kawachi N, Brandwein-Gensler M, McLemore M, Chen Q, Burk RD, Smith RV, Prystowsky MB, et al. Low-level expression of microRNAs let-7d and miR-205 are prognostic markers of head and neck squamous cell carcinoma. Am J Pathol. 2009;174(3):736–45.

18. Babu JM, Prathibha R, Jijith VS, Hariharan R, Pillai MR. A miR-centric view of head and neck cancers. Biochim Biophys Acta. 2011;1816(1):67–72.
19. Severino P, Oliveira LS, Torres N, Andreghetto FM, Klingbeil Mde F, Moyses R, Wünsch-Filho V, Nunes FD, Mathor MB, Paschoal AR, Durham AM. High-throughput sequencing of small RNA transcriptomes reveals critical biological features targeted by microRNAs in cell models used for squamous cell cancer research. BMC Genomics. 2013;14:735.
20. Applied Biosystems. SOLiD3 System Application Documentation Small RNA Analysis Tool. 2009. Available: http://solidsoftwaretools.com/gf/.
21. Lindgreen S. AdapterRemoval: easy cleaning of next-generation sequencing reads. BMC Res Notes. 2012;5:337.
22. Martin M. Cutadapt removes adapter sequences from high-throughput sequencing reads. EMBnet J. 2011;17(1):10–2.
23. Kong Y. Btrim: A fast, lightweight adapter and quality trimming program for next-generation sequencing technologies. Genomics. 2011;98(2):152–3.
24. Smeds L, Künstner A. ConDeTri—a content dependent read trimmer for Illumina data. PLoS One. 2011;6(10):e26314.
25. Dillies MA, Rau A, Aubert J, Hennequet-Antier C, Jeanmougin M, Servant N, Keime C, Marot G, Castel D, Estelle J, Guernec G, Jagla B, Jouneau L, Laloe D, Gall C, Schaeffer B, Crom S, Guedj M, Jaffrezic F, behalf of The French StatOmique Consortium. A comprehensive evaluation of normalization methods for Illumina high-throughput RNA sequencing data analysis. Brief Bioinform. 2012;14(6):671–83.
26. Garmire LX, Subramaniam S. Evaluation of normalization methods in mammalian microRNA-Seq data. RNA. 2012;18(6):1279–88.
27. Robinson MD, McCarthy DJ, Smyth GK. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics. 2010;26(1):139–40.
28. Anders S, Huber W. Differential expression of RNA-Seq data at the gene level—the DESeq package. 2013.
29. Lowe TM, Eddy SE. A computational screen for methylation guide snoRNAs in yeast. Science. 1999;283:1168–71.
30. Hertel J, Hofacker IL, Stadler PF. SnoReport: computational identification of snoRNAs with unknown targets. Bioinformatics. 2008;24(2):158–64.
31. Backman TW, Sullivan CM, Cumbie JS, Miller ZA, Chapman EJ, Fahlgren N, Givan SA, Carrington JC, Kasschau KD. Update of ASRP: the Arabidopsis small RNA project database. Nucleic Acids Res. 2008;36(Database issue):D982–5.
32. Johnson C, Bowman L, Adai AT, Vance V, Sundaresan V. CSRDB: a small RNA integrated database and browser resource for cereals. Nucleic Acids Res. 2007;35(Database issue): D829–33.
33. Kin T, Yamada K, Terai G, Okida H, Yoshinari Y, Ono Y, Kojima A, Kimura Y, Komori T, Asai K. fRNAdb: a platform for mining/annotating functional RNA candidates from non-coding RNA sequences. Nucleic Acids Res. 2007;35(Database issue):D145–8.
34. Szymanski M, Erdmann VA, Barciszewski J. Noncoding RNAs database (ncRNAdb). Nucleic Acids Res. 2007;35(Database issue):D162–4.
35. Bu D, Yu K, Sun S, Xie C, Skogerbo G, Miao R, Xiao H, Liao Q, Luo H, Zhao G, et al. NONCODE v3.0: integrative annotation of long noncoding RNAs. Nucleic Acids Res. 2012;40(Database issue):D210–5.
36. Griffiths-Jones S. Annotating non-coding RNAs with Rfam. In: Current protocols in bioinformatics Edited by Baxevanis AD. 2005. Chapter 12:Unit 12.5.
37. Kadri S, Hinman V, Benos PV. HHMMiR: efficient de novo prediction of microRNAs using hierarchical hidden Markov models. BMC Bioinform. 2009;10 Suppl 1:S35.

Exome Capture and Capturing Technologies in Cancer Research

Chandra Sekhar Reddy Chilamakuri and Leonardo A. Meza-Zepeda

Abstract Next-generation sequencing technologies are revolutionizing the study of genomic variation. Although whole-genome sequencing is the most comprehensive strategy for genome-wide variant detection, this approach is still expensive for routine clinical use. The targeted sequencing of all coding regions, approximately 2 % of the human genome, is termed whole-exome sequencing, and it has emerged as an indispensable tool for clinical research, particularly cancer research. Indeed, the application of whole-exome sequencing in cancer, a disease of the genome, has transformed our knowledge about this disease. In this chapter, we review exome-capture methods and technologies and their application in cancer research.

1 Introduction

Theodor Boveri was the first to propose the connection between cancer and chromosomal anomalies [1], and Stehelin et al. first demonstrated the relationship between genes and cancer in 1976 [2]. By the mid-1980s, researchers had established two main types of cancer-causing genes, oncogenes and tumor suppressor genes. For example, in 1982, the MYC gene was shown to be affected by a chromosomal translocation observed in Burkitt's lymphoma [3]. While studying retinoblastoma, Knudson et al. discovered the first tumor suppressor gene, RB1 [4]. These studies and others began to unravel the mutational complexity of cancer, i.e., the variability of cancer-causing genes across and within cancer types. Intensive research has established that cancer often originates as the result of somatic alterations to a cell's genome. These changes include single-nucleotide variations, small insertions and deletions, and large and complex structural changes that can affect

C.S.R. Chilamakuri (✉) • L.A. Meza-Zepeda
Department of Tumor Biology, Norwegian Radium Hospital, Oslo University Hospital, Oslo, Norway

Norwegian Cancer Genomics Consortium (cancergenomics.no), Oslo, Norway
e-mail: chichi@rr-research.no; Leonardo.A.Meza-Zepeda@rr-research.no

W. Wu, H. Choudhry (eds.), *Next Generation Sequencing in Cancer Research, Volume 2*, DOI 10.1007/978-3-319-15811-2_17

entire chromosomes. These somatic alterations eventually give rise to uncontrolled cell growth and division. In the late 1980s, there were several calls to sequence the human genome, a step that is essential for systematically discovering all of the genes responsible for cancer development [5]. The Human Genome Project was launched in the year 1990 and was completed in 2004 [1, 6]. Furthermore, rapid advances in sequencing technologies have dramatically decreased the cost of sequencing, and next-generation sequencing instrumentation, available since 2005, has altered our approach to sequencing genomes.

The identification of the genetic basis of a human disease is an important area of research, particularly with regard to cancer research, and a good approach is to sequence the entire genome at high resolution. Although rapid advances in next-generation sequencing technologies have dramatically reduced the cost of DNA sequencing, the cost of whole-genome sequencing remains significantly high when applied to large numbers of individual samples. Alternatively, targeted sequencing strategies have been developed to reduce the associated costs. The protein-coding portion of the human genome, the "exome," which is approximately 1–2 % of the human genome, is an attractive target for targeted resequencing. In fact, exome sequencing has emerged as an indispensable tool in the era of predictive and precision medicine, and much research has been performed on the clinical benefits and risks of sequencing to screen healthy persons [2, 7]. Exome sequencing (Exome-seq) has certain advantages over whole-genome sequencing: Exome-seq is significantly less expensive, and the protein-coding regions of the genome are well studied. Therefore, the functional interpretation of exome variants is relatively easy, and the time required for the analysis of Exome-seq data is significantly lower. One important objective of cancer genomics is to identify driver genes, which requires the sequencing of large numbers of samples at very high coverage; whole-genome sequencing is currently still expensive, whereas exome sequencing is a very attractive tool for cancer researchers.

Although exome sequencing has certain advantages with regard to cost and ease of data handling and analysis, there are certain limitations to this approach, especially when investigating complex genomes such as cancer genomes. Specifically, even at a very high sequencing depth, a small proportion of the exome will not be covered, and the non-covered exome sequences often contain a very high GC content [3, 8]. Regardless, exome sequencing is very useful for the detection of single-nucleotide variations and small insertions and deletions. However, it is less accurate for other types of genetic variations: exome sequencing cannot detect large structural alterations and has limited ability to detect copy-number changes. Exome sequencing also requires complex library-preparation procedures that normally demand several days to complete, including a hybridization step. In general, exome sequencing cannot cover the functional elements outside of exons, such as promoters, transcription factor binding sites, and enhancers. Nonetheless, Exome-seq is of great interest to clinical and translational researchers; in particular, significant efforts have been applied to cancer research. The relationship between cancer and genomic mutations, especially mutations in protein-coding regions, is well established, yet studying the complexity and heterogeneity of cancer genomes often requires sequencing at very high coverage rates. Therefore, the Exome-seq strategy is a natural choice for cancer researchers.

In this chapter, we first provide a brief overview of exome-capture strategies and a brief comparison of popular commercial exome-capture technologies. Then, we provide an overview of the application of exome sequencing in cancer research.

2 The Exome

The exome is defined as the protein-coding and RNA-coding regions of known human genes. There are many different databases available for human coding genes, such as RefSeq [4, 9] and Ensembl [5, 10], which differ in the total number of non-coding RNAs and the total number of exons present as well as the start and end coordinates of the exons. However, the majority of sequences are in common among the different databases. The consensus coding DNA sequence (CCDS) database contains protein-coding sequences with high-quality annotations [11]. RefSeq and CCDS share a greater proportion of sequences in common, whereas Ensembl has more unique bases compared with the other two databases (Fig. 1).

3 Capture Strategies

During the next-generation sequencing process, genomic DNA is first fragmented, and these DNA segments are then sequenced simultaneously. Without selectively targeting genomic regions of interest, any of the genomic fragments has an equal chance of being sequenced. To sequence a subset of regions from the human genome, one needs to first extract the regions of interest from a genomic DNA library, and there are different ways that this outcome can be achieved. The principles of target-capture strategies have been comprehensively discussed in other reviews [12–15]; various target-capture methods are depicted in Fig. 2.

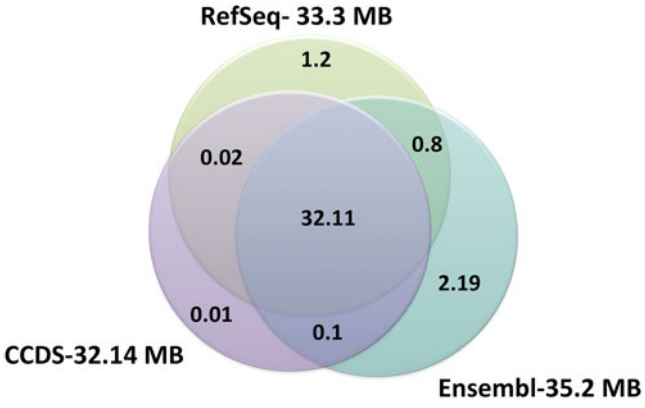


Fig. 1 Comparison of coding databases. Overlap of exon bases among the RefSeq, CCDS, and Ensembl exon databases. 32.11 MB (megabase pairs) shared by all three databases. 2.19 MB unique to Ensembl database

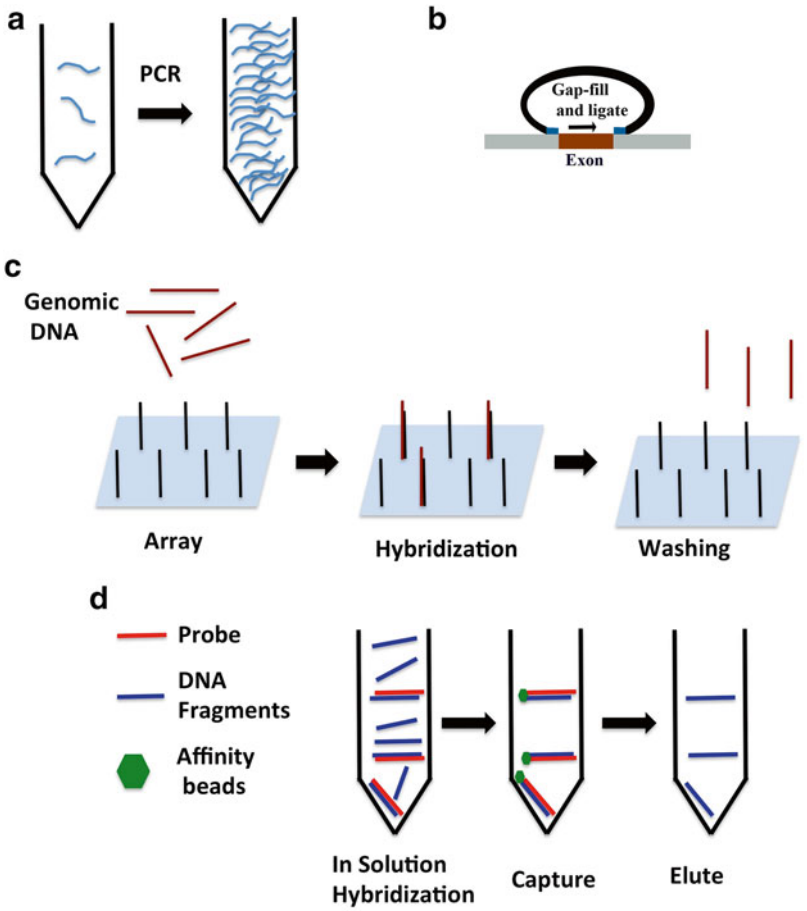


Fig. 2 Different exome capture technologies. (**a**) PCR-based exome enrichment. (**b**) Molecular inversion probes (MIPs) in exome capture: probes (*blue*) flanking a region of interest hybridize to their target (*brown*). Polymerase and ligase reactions allow the surrounded region to be filled in and effectively circularize the probe. (**c**) Microarray-based exome capture: a microarray with probes (*black*) corresponding to regions of interest is exposed to a pool of fragments (*brown*). Target fragments hybridize to the probes on the microarray, and the hybridized fragments are then eluted for downstream sequencing. (**d**) Solution-based exome capture: oligonucleotide probes (*red*) in a single test tube capture DNA fragments (*blue*). The oligonucleotide probes can be selectively extracted by the use of affinity beads (*green*)

3.1 Polymerase Chain Reaction (PCR)-Based Enrichment

PCR has been extensively used as a pre-sequencing sample preparation method for the last few decades [16], and this method is highly suitable for a Sanger sequencing-based approach. PCR-based capture is also potentially suitable for any next-generation sequencing approach to make use of its throughput capacity when a large number of amplicons must be sequenced. Nevertheless, as the target size and number of amplicons increase, the efficiency, as measured in terms of cost, feasibility, and input requirements,

rapidly decreases. Despite its limitations, some of the initial large-scale sequencing projects used PCR for selective amplification [17, 18]. The sequencing of targeted regions of the genome by next-generation sequencing instruments necessitates a massively parallel enrichment method for the targets to be sequenced. A number of technologies are currently available that allow the selective amplification of tens to hundreds of targeted regions. These methods are widely used for focused gene or hot spot cancer panels and allow the detection of low allele-frequency mutations by sequencing at very high coverage. PCR-based methods are easy to implement in a clinical diagnostic setting yet are limited by the number of amplicons that can simultaneously be amplified. To overcome this limitation, other methods that allow a higher number of parallel reactions have been developed to enrich for larger segments of the genome.

3.2 Molecular Inversion Probes (MIPs)

MIPs are single-stranded DNA molecules that contain two regions complementary to regions in the target genomic DNA. Successful binding to a target sequence results in a conformational change that allows the molecules to be directly selected by amplification methods. After successful hybridization to a target, MIPs are elongated along the region of interest and closed by ligation to generate a circular molecule, which is protected against exonucleases. Treatment with restriction enzymes linearizes the circular molecules, which can then be directly sequenced. Unlike shotgun-based library preparations, MIP-based capturing procedures do not require a library-preparation step; because of the two enzymatic steps, the specificity is very high, and the DNA input requirements are low. The disadvantage of MIPs for target enrichment is that capture uniformity is poor compared with capture by hybridization. Moreover, MIP oligonucleotides can be expensive, and a large number of oligonucleotides are required to cover a large number of targets.

3.3 Hybridization-Based Capture

Enrichment by hybridization entails the incubation of oligonucleotide probes that are complementary to target regions with a genomic DNA library; nonbinding fragments are removed, and the bound, enriched DNA is then eluted for sequencing. However, due to nonspecific binding, hybridization-based capture methods generally have a lower capture specificity compared with other approaches.

3.3.1 Array-Based Capture

Array-based hybridization methods typically use probes that are complimentary to the sequence of a target fixed to a solid support such as a microarray. A genomic library of interest is hybridized; after incubation, nonbinding fragments are removed, and the hybridized fragments are eluted for sequencing. This method can enrich regions of interest by approximately 1,000–2,000-fold in one round of hybridization.

Utilizing multiple enrichment cycles can further enhance the enrichment efficiency, and microarray-based hybridization platforms are essentially reusable. As a mature technology, oligonucleotide microarrays are also relatively inexpensive compared with other targeted sequencing strategies. Array-based capture offers greater flexibility and can be used to capture either a contiguous region or many short, dis-contiguous regions, and by varying the probe spacing, the target size of an array can be changed from hundreds of kilobases to tens of megabases using the same protocol. The throughput of array-based target capture, nevertheless, is dependent on the spatial resolution of the oligonucleotide probe array; therefore, the number of targets that can be captured is limited. Another disadvantage is that scaling array-based capture for hundreds of samples can be more difficult compared with solution capture. Although array capture is relatively inexpensive, the cost of the arrays is still high when considering a large number of samples.

3.3.2 Solution-Based Capture

Bashiardes et al. described a modified DNA-based genomic DNA selection protocol for performing hybridization-based targeted capture of shotgun fragments corresponding to bacterial artificial chromosome (BAC)-sized genomic regions [19]. In this method, a shotgun library is generated from the genomic DNA of interest, and adaptors are ligated. The library is then hybridized in solution to biotinylated DNA that is derived from the regions of interest. The target-probe hybrids are pulled down by streptavidin beads followed by washing to reduce nonspecific hybridization. The captured target regions are eluted, PCR amplified, and sequenced. In solution hybridization, the probe molecules can be either DNA or RNA; for instance, Gnirke et al. implemented this procedure using RNA as the probe molecule [20], whereas others have used DNA as the probe [21]. An important advantage of using RNA as the probe is that RNA molecules are single-stranded and present only one orientation, and a high concentration of RNA can drive the hybridization kinetics.

The following are some of the notable advantages of solution-based hybridization over array-based capture. First, a relatively low input of genomic DNA is required. Second, solution-based capture is more automatable than array-based hybridization capture. Third, systematic bias can be reduced by using longer capture probes compared with the short probes used in array-based capture. Finally, solution-based capture has higher specificity compared with array-based capture.

4 Overview of Commercial Exome Capture Technologies

The majority of commercially available exome-capture kits employ a solution hybridization strategy to capture exomes. Although the sample preparation methods are similar across different platforms, there are differences in the choice of their target regions, probe lengths, bait density, molecule used for target region capture, and genome fragmentation method (Table 1). There are currently four major solution-based exome-capture kits available: Agilent SureSelect Human All Exon,

Table 1 Comparison of commercial exome-capturing platforms

	NimbleGen	Agilent	Illumina TruSeq	Illumina Nextera
Probe type	DNA	RNA	DNA	DNA
Probe length range (bp)	55–105	114–126	95	95
Number of probes	2,100,000[a]	554,079	347,517	347,517
Total probe length (Mb)	NA	66.48	33.01	33.01
Target length range (bp)	59–742	114–21,747	2–37,917	2–37,917
Median target length (bp)	171	200	135	135
Number of targets	368,146	185,636	201,071	201,071
Total target length (Mb)	64.19	51.18	62.08	62.08
Fragmentation method	Ultrasonication	Ultrasonication	Ultrasonication	Transposomes

NA, not available

The comparisons are based on Agilent SureSelect Human All Exon v4.0, NimbleGen SeqCap EZ Exome Library V3.0, Illumina TruSeq Exome Enrichment Kit, and Illumina Nextera Exome Enrichment Kit

[a]The number of probes for the NimbleGen platform was obtained from a previous report [22]

NimbleGen SeqCap EZ Exome Library, Illumina TruSeq Exome Enrichment Kit, and Illumina Nextera Exome Enrichment Kit. Prior studies have reported comparisons of the different exome-capturing systems. For instance, Clark et al. compared three capture technologies and showed that the NimbleGen technology required the least number of reads to sensitively detect small variants; in contrast, the Agilent and Illumina technologies were able to detect a higher total number of variants with additional reads [22]. In another study, Sulonen et al. compared the NimbleGen and Agilent technologies and demonstrated that there were no significant differences between the two except that NimbleGen platform showed a greater efficiency in covering the exome, with a minimum of 20X coverage [23]. Asan et al. compared NimbleGen Sequence Capture Array, NimbleGen SeqCap EZ, and Agilent SureSelect and showed that all three platforms achieved a similar accuracy of genotype assignment and single-nucleotide polymorphism (SNP) detection and similar levels of reproducibility and GC bias [8]. In another exome-capture comparison study, Parla et al. demonstrated that both NimbleGen SeqCap EZ Exome Library SR and Agilent SureSelect All Exon were similar to each other and captured most of the human exons that were targeted by their probe sets; nevertheless, these kits failed to cover a noteworthy percentage of the exons in the CCDS annotations compared with high-coverage, whole-genome sequencing [24]. In a very recent comparative study looking at four exome-capture kits, Chilamakuri et al. demonstrated that the Agilent kit provided higher coverage for a given number of reads and that the Illumina Nextera kit showed bias toward targeted regions with a high GC content [25].

5 Strategies for Cancer Exome Sequencing Data Analysis

Exome sequencing data are processed in different steps (Fig. 3). The first step in identifying variants from exome sequencing is to align short reads to a reference genome. There are many different short-read aligners, and Ruffalo et al. comprehensively

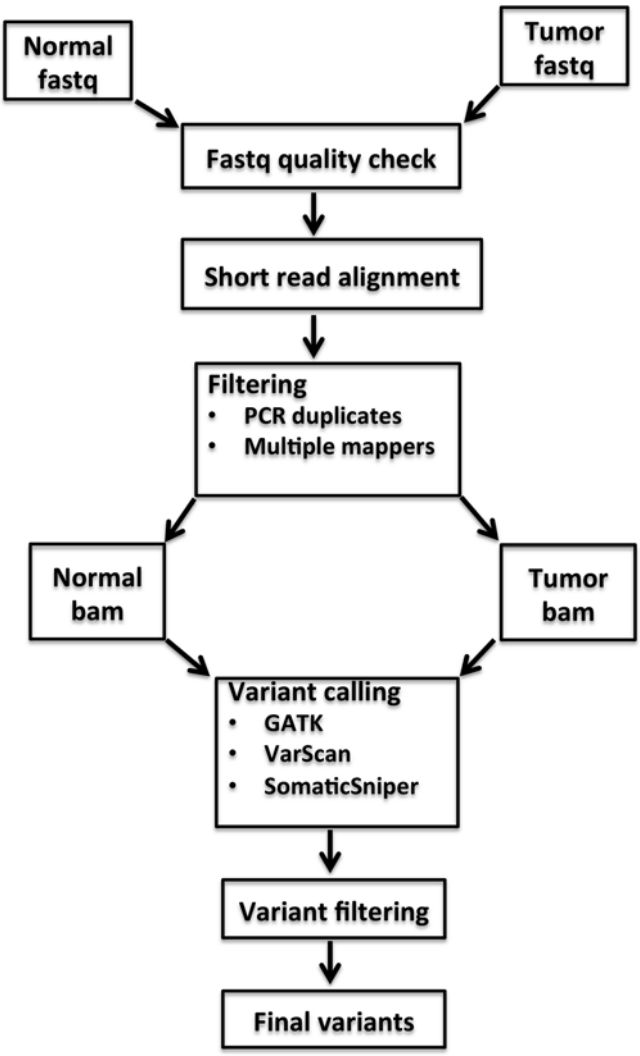


Fig. 3 Overview of somatic variant calling

compared the pros and cons of different aligners [26]. PCR is generally used to add adapters and to amplify the genomic library for sequencing. If the starting DNA sample is small, an increase in the number of PCR cycles is used to overcome this limitation, but additional PCR cycles increase the number of duplicates, significantly affecting the randomness of the sequencing process. Some parts of the genome result in very high coverage compared with others. Amplification errors in early PCR cycles could then be present in multiple reads, and these errors are difficult to distinguish from real genomic variations; therefore, it is essential to remove duplicate reads. Filtered normal and tumor BAM files are used for somatic variant calling, and some of the popular somatic variant callers include MuTect [27], VarScan2 [28], and SomaticSniper [29]. Xu et al. systematically compared somatic mutation-calling methods [30].

Variant callers generally identify tens of thousands of variants, and different computational algorithms, such as Sorting Intolerant from Tolerant (SIFT) [31] and PolyPhen [32], are used to predict the potential effect of a variant on protein function.

6 Exome Sequencing in Cancer Research

The gradual accumulation of mutations during the lifespan can lead to cancer. Because human cancers are very heterogeneous entities, multiple gene tests should be performed to determine the underlying mutations, and the application of deep-sequencing technologies has allowed significant advances in cancer genomics research. Indeed, advances in deep-sequencing technologies, coupled with advances in capturing technologies, have made the systematic identification of somatic mutations in a particular cancer feasible. Because whole-exome sequencing targets all of the coding regions of the genome, the use of this approach avoids the requirement for multiple genetic tests. Additionally, the total cost of exome sequencing has fallen sharply such that it is now within the budgets of many cancer-related research projects. The applications of whole-exome sequencing have been comprehensively reviewed [33–35]. One limitation of targeted exome sequencing is that researchers should know the targets prior to the experiments; however, whole-exome sequencing removes this bias by targeting all exonic regions. Another limitation of exome sequencing is that it cannot be used to detect structural variations. Despite these technical limitations, exome sequencing has emerged as a popular tool for discovering clinically relevant mutations, such as those involved in cancer.

The human genome project was launched in 1990, and the initial draft sequence was finished by 2001 [36, 37]. With the availability of the public human genome sequence, rapid progress has been achieved in cancer genomics, and genome sequencing, especially targeted resequencing, has attained widespread use in cancer research. Initial attempts at the large-scale DNA sequencing of cancer samples faced limitations such as the high cost and extensive infrastructure required, which placed tight constraints on the amount of data that could be collected. By analyzing exome sequencing from 13,023 genes in 11 breast and 11 colorectal cancers, Sjöblom et al. showed in 2007 that 189 were mutated at a significant frequency [18]. In the same year, based on an analysis of exons representing 20,857 transcripts from 18,191 genes in 11 breast and 11 colorectal tumors, Wood et al. concluded that the genomic landscapes of breast and colorectal cancers are composed of a handful of commonly mutated gene “mountains” and a much larger number of gene “hills” mutated at lower frequencies [38]. By analyzing exome sequencing from 623 candidate cancer genes in 188 lung adenocarcinomas, in 2008, Ding et al. discovered somatic mutations in the coding exons of those genes [39]. In 2007, Greenman et al. reported more than 1,000 somatic mutations in the coding exons of 518 protein kinase genes in 210 diverse human cancers [40]. As the cost of sequencing started to fall sharply beginning in 2008, the application of exome sequencing largely increased in cancer research, resulting in the discovery of important factors in various cancers (Table 2). Whole-exome sequencing, despite its technical challenges, has emerged as a popular tool for identifying clinically relevant somatic mutations in cancer.

Table 2 Application of exome sequencing in various cancers

Cancer type	Subtype	No. of samples	Sequencing platform	Capture technology	Important findings	Reference
Carcinoma	Prostate cancer (PC)	33	HiSeq2000	SureSelectXT Human All Exon 50mB Kit	Four missense variants in *BLM* gene showed complete co-segregation with PC status	[41]
		7	HiSeq2000	NimbleGen SeqCap EZ Exome v2.0	Mutations in cancer-associated genes and the p53-signaling pathway were found exclusively in high-grade foci and metastases	[42]
		3	HiSeq2000	Agilent SureSelect Human All Exome Kit	Substantial genetic intratumoral heterogeneity in unifocal PCAs at the mutation, copy number, and expression levels	[43]
		1	HiSeq2000	Agilent SureSelect All Exon v3 Kit	Damaging somatic mutations were observed in the retained *TP53* and *RB1* alleles	[44]
		1	Illumina HiSeq	Agilent v2 Human Exon Kit	Mapping of >99.995 % of the standard exome is possible in circulating tumor cells	[45]
		2 cell lines	Illumina GAIIx sequencer	Agilent SureSelect Human All Exon	Exome sequencing detected 2188 and 3840 mutations in LNCaP and C4-2B cells, respectively, of which 1784 were found in both cell lines	[46]
		1	HiSeq2000	Agilent SureSelect Human All Exon Kit v2	Identified plausibly actionable somatic genomic alterations that dysregulate the phosphoinositide 3-kinase pathway, as well as a theoretically actionable germ line variant in the *BRCA2* gene	[47]
		5 metastatic samples	SOLiD	Agilent SureSelect All Exon	Inhibition of *YWHAZ* and *PTK2* could delay the progression of the disease	[48]
		91	HiSeq2000	SureSelect Human All Exon 50 Mb	Rare *BTNL2* variants play a role in susceptibility to both familial and sporadic prostate cancer	[49]
		64	Illumina	NimbleGen SeqCap EZ Human Exome Library v1	The mitochondrial genome displays an elevated mutation rate compared to the autosomal chromosomes	[50]
		112	Illumina HiSeq	Agilent SureSelect v2 Exome Kit	Recurrent *SPOP*, *FOXA1*, and *MED12* mutations in prostate cancer	[51]

		1 cell line	Illumina GAIIx	Agilent SureSelect Human All Exon	Detected 1,802 nonsynonymous SNVs and 218 small insertions and deletions in the LNCaP cell line exome	[52]
	Colorectal carcinoma (CRC)	22	HiSeq2000	Agilent SureSelect Human All Exon Kit	Seven significantly mutated genes, *APC*, *TP53*, *KRAS*, *SMAD4*, *CDH10*, *FAT4*, and *DOCK2*, exhibited high mutation prevalence	[53]
		26	HiSeq2000	NimbleGen SeqCap EZ Exome 2.0	Germ line mutation of *RPS20* causes predisposition to hereditary nonpolyposis colorectal carcinoma	[54]
		70 cell lines	HiSeq2000	TruSeq Exome Enrichment Kit	Mutations enriched in genes involved in chromatin remodeling *ARID1A*, *CHD6*, and *SRCAP* and histone methylation or acetylation *ASH1L*, *EP300*, *EP400*, *MLL2*, *MLL3*, *PRDM2*, and *TRRAP*	[55]
		53	Illumina HiSeq	Agilent SureSelect All Exon Kit 38 Mb	Half of the metastatic colorectal carcinoma had the same clonal origin with their primary colorectal carcinomas, and the remaining cases were genetically distinct from their primary carcinomas	[56]
		20	HiSeq2000	NimbleGen SeqCap EZ Human Exome Library v2.0 or Agilent SureSelectXT Human All Exon v4	Identified *RNF43* as a regulator of the DNA damage response and associated nonsense variants in this gene with a high risk of developing sessile serrated adenomas	[57]
		11	HiSeq2000	NimbleGen SeqCap EZ Human Exome Library v2.0	*Fusobacterium* enrichment is associated with specific molecular subsets of colorectal cancers	[58]
		25	Illumina GA II	Agilent SureSelect All Exon Kit v1	Inactivating mutations in *ARID1A*, *ARID1B*, *ARID2*, and *ARID4A* in microsatellite unstable colorectal cancer	[59]
		16	HiSeq2000	NimbleGen SeqCap EZ Human Exome v2.0	Elevated rate of mutations in *CHD7* and *CHD8* genes	[60]

(continued)

Table 2 (continued)

Cancer type	Subtype	No. of samples	Sequencing platform	Capture technology	Important findings	Reference
		1514	Illumina GAII or HiSeq	Agilent SureSelect Human All Exon Kit v1	Identified rare truncating variants in *UACA*, *SFXN4*, *TWSG1*, *PSPH*, *NUDT7*, *ZNF490*, *PRSS37*, *CCDC18*, *PRADC1*, *MRPL3*, and *AKR1C4*	[61]
		25	Illumina Genome Analyzer II	Agilent SureSelect Human All Exon Kit v1	Potential mutation hot spots were confirmed in *ADAR*, *DCAF12L2*, *GLT1D1*, *ITGA7*, *MAP1B*, *MRGPRX4*, *PSRC1*, *RANBP2*, *RPS6KL1*, *SNCAIP*, *TCEAL6*, *TUBB6*, *WBP5*, *VEGFB*, and *ZBTB2*	[62]
		40	Illumina GAIIx	Agilent 36 Mb or All Human Exon chip	Two variants in *CENPE* and *KIF23*, located within previously reported CRC linkage regions	[63]
		1	Illumina GAII platform	NimbleGen 2.1 M Human Exome Array	Identified 12 nonsynonymous somatic SNVs in the adenoma and 42 nonsynonymous somatic SNVs in the adenocarcinoma	[64]
		474	SOLiD	NimbleGen SeqCap EZ Exome 2.0	*MLH1* expression status and frequencies of *APC*, *KRAS*, and *BRAF* mutation in CRC may provide a useful diagnostic tool	[65]
		224	SOLiD or Illumina HiSeq	NimbleGen SeqCap EZ Exome 2.0	Comprehensive molecular characterization of human colon and rectal cancer	[66]
	Hepatocellular carcinoma	30	Illumina HiSeq	-	Frequent mutations in *TP53*, *CTNNB1*, *AXIN1*, *BAP1* and *IDH1*.	[67]
		47	HiSeq2500	Agilent SureSelect XT Human All Exon 50 Mb kits	Demonstrated the importance of *CTNNB1* mutations and *NFE2L2-KEAP1* pathway activation in hepatoblastoma	[68]
		1	Illumina HiSeq	NimbleGen human solution capture	Pediatric hepatocellular carcinoma due to somatic *CTNNB1* and *NFE2L2* mutations in the setting of inherited bi-allelic *ABCB11* mutations	[69]
		231	HiSeq2000	Agilent SureSelect 50 Mb	*RB1* mutations can be used as a prognostic molecular biomarker for resectable hepatocellular carcinoma	[70]

		250	HiSeq2000	SureSelect Human All Exon Kit v2	Recurrent somatic mutation activating *FRK*, a Src-like kinase	[71]
		2 from same patient	HiSeq2000	NimbleGen	*UBE3C* is a candidate oncogene involved in tumor development and progression	[72]
		2	HiSeq2000	Agilent SureSelect XT Human All Exon v.2 44 Mb capture kit	A novel homozygous p.Ser171Phe *TALDO1* variant identified in a family with a cirrhosis	[73]
		7	Illumina Genome Analyzer IIx	NimbleGen SeqCap EZ Human Exome Library v2.0	Somatic mutations accumulate in *LEPR* in cirrhotic liver with chronic hepatocellular carcinoma in hepatitis C virus infection	[74]
		87	HiSeq2000	Agilent SureSelect	The *NFE2L2-KEAP1* and *MLL* pathways are recurrently mutated in multiple cohorts of hepatocellular carcinoma	[75]
		2	Illumina HiSeq	Agilent array	A novel nonsense mutation in exon 26 of *APOB* (p.K2240X) may be responsible for cirrhosis and liver	[76]
		1	Illumina Genome Analyzer IIx	NimbleGen SeqCap EZ Human Exome v 2.0	Seven nonsynonymous somatic variants in *SPATA21*, *PPCS*, *CDH12*, *OR1L3*, *PCK2*, *HUWE1*, and *PHF16*	[77]
		10	Illumina Genome Analyzer II or SOLiD	NimbleGen Human Exome 2.1 M Arrays or Agilent SureSelect	*VCAM1* and *CDK14* may confer growth and infiltration capacity to hepatocellular carcinoma	[78]
		24	HiSeq2000	Agilent SureSelect Human All Exon Kit v2	Recurrent alterations in *ARID1A*, *RPS6KA3*, *NFE2L2*, and *IRF2*	[79]
		1	Illumina GAIIx	Agilent SureSelect	First high-resolution characterization of a virus-associated cancer genome	[80]
	Breast cancer	89	SOLiD	NimbleGen SeqCap EZ Exome v2.0	Rare mutations in *RINT1* predispose carriers to breast and lynch syndrome	[81]
		98	HiSeq2000	Agilent SureSelectXT Human All Exon v3	Highly recurrent *MED12* somatic mutations in breast fibroadenoma	[82]

(continued)

Table 2 (continued)

Cancer type	Subtype	No. of samples	Sequencing platform	Capture technology	Important findings	Reference
		10	HiSeq2000	NimbleGen SeqCap EZ Exome version 3.0	Dynamic conversion between differentiation states of breast cancer stem cell populations in vivo	[83]
		16 samples from three families	HiSeq2000	TruSeq Exome Enrichment Kit	Genetic predispositions in many BRCAx familial breast cancer families can be family specific	[84]
		1	HiSeq2000	NimbleGen 2.1 M-probe sequence capture array	Loss of heterozygosity observed for multiple genes of the *CECAM*, *MMP*, and *ZNF* families	[85]
		8	HiSeq2000	NimbleGen SeqCap EZ Exome v2.0	*BRCA1*+ can cause genome instability with both germ line and somatic mutations in non-breast cells	[86]
		2	HiSeq2000	Agilent SureSelect Human All Exon Kit	Enriched variations in TEKT4 confer breast cancer resistance to paclitaxel	[87]
		6	HiSeq2000	NimbleGen SeqCap EZ Exome v2.0	High prevalence of *GPRC5A* germ line mutations in BRCA1-mutant breast cancer patients	[88]
		11	HiSeq2000	Agilent SureSelect Human All Exon v4	Activating mutations in *ESR1* are a key mechanism in acquired endocrine resistance in breast cancer therapy	[89]
		507	HiSeq2000	Agilent SureSelect All Exome v2.0	Comprehensive mutational profiles of human breast tumors	[90]
		103	Illumina GA II	Hybrid Capture	Recurrent mutations in the *CBFB* transcription factor gene and deletions of its partner *RUNX1*	[91]
		46	HiSeq2000	NimbleGen SeqCap EZ v2.0	Eighteen significantly mutated genes were identified	[92]
		50	Illumina GAIIx	Agilent SureSelect n All Exon Kit	Identified large number of variants	[93]
		13	SOLiD	NimbleGen exome-capture protocol v.1.2	Mutations in *XRCC2* increase the risk of breast cancer	[94]

	Lung Cancer	105	HiSeq2000	NimbleGen 44 M Human Exome	Identified *MLL2* gene as one of the most significantly mutated genes	[95]
		2	HiSeq2000	NimbleGen SeqCap EZ Exome Plus 64 M	Truncating mutations were detected in 8 cancer genes	[96]
		15	Illumina GAII	Agilent Exome Capture	Identified mutations in 11 members of the NF-κB pathway	[97]
		15	HiSeq2000	NimbleGen SeqCap EZ v2.0	*MEN1*, *PSIP1*, and *ARID1A* genes are recurrently mutated	[98]
		70	HiSeq2000	TruSeq Exome Enrichment Kit	Identified 27 genes potentially implicated in the pathogenesis of lung adenocarcinoma	[99]
		343	Hiseq2000	NimbleGen 2.1 M Human Exome Array	Novel *PROM1* and *CRTC2* mutations, which could promote lung cancer development	[100]
		10	Hiseq2000	TruSeq Exome Enrichment Kit	Recurrent somatic mutations in *EGFR*, *BCHE*, and *TP53* genes	[101]
		16	Hiseq2000 or Hiseq2500	Agilent SureSelect All Exon V4	CNVs at certain genomic loci are selected for the metastasis of cancer	[102]
		4 cell lines	Hiseq2000	Oncopanel Capture Kit	Acquired resistance to dasatinib in lung cancer cell lines conferred by *DDR2* gatekeeper mutation and *NF1* loss	[103]
		42	HiSeq2000	Agilent SureSelect Human All Exome Kit (38 Mb or 50 Mb)	*SOX2* is frequently amplified gene in small-cell lung cancer	[104]
		31	Illumina Genome Analyzer IIx	Agilent SureSelect	*CSMD3* frequently mutated	[105]
Sarcoma	Liposarcoma	2	SOLiD	Agilent SureSelect	Frequent alterations and epigenetic silencing of differentiation pathway genes in liposarcomas	[106]
	Kaposi sarcoma	1	Illumina Genome Analyzer IIx	Agilent SureSelect Human All Exon	*STIM1* T-cell deficiency precipitated the development of lethal Kaposi sarcoma	[107]

(continued)

Table 2 (continued)

Cancer type	Subtype	No. of samples	Sequencing platform	Capture technology	Important findings	Reference
	Thymoma	1	HiSeq2500	Agilent SureSelect Exome	*ASXL1* and *DNMT3A* mutation in a cytogenetically normal B3 thymoma	[108]
	Angiosarcoma	11	HiSeq2000	Agilent SureSelect Human All Exon 50 Mb	Identified recurrent *PTPRB* and *PLCG1* mutations in angiosarcoma	[109]
	Rhabdomyosarcoma	1	HiSeq2000	Agilent SureSelectXT Human All Exon Kit	Aberrant *CDK4* amplification in refractory rhabdomyosarcoma	[110]
		1	HiSeq2000	Agilent SureSelect	*MYOD1* gene is frequently mutated	[111]
Myeloma		84	HiSeq2000	Agilent SureSelect Human Exon Kit	Identified truncations of *SP140*, *LTB, ROBO1*, and clustered missense mutations in *EGR1*	[112]
		36	HiSeq2000	Agilent SureSelect Human All Exon 50 Mb	Intraclonal heterogeneity is a typical feature of the disease	[113]
		5	Illumina GAIIx	NimbleGen SeqCap EZ Exome Library	Increased rate of mutations in receptor tyrosine kinases (RTKs) and associated signaling effectors	[114]
Leukemia	Primary CNS lymphoma	9		Agilent SureSelectXT Human All Exon V4	Identified new targets of aberrant somatic hyper mutations in *KLHL14*, *OSBPL10*, and *SUSD2*	[115]
	Acute lymphoblastic leukemia (ALL)	42	HiSeq2000	Agilent SureSelect Human Exome v5	Identified driver mutations in *KRAS*, *NRAS*, and *JAK2*	[116]
		56	Illumina Genome Analyzer IIx	Agilent SureSelect Human All Exon 50 Mb Kit	RAG-mediated recombination is the predominant driver of oncogenic rearrangement in *ETV6-RUNX1* acute lymphoblastic leukemia	[117]
		34	Illumina Genome Analyzer IIx	Agilent SureSelect Human All Exon V4	Identified high frequency of mutations in spliceosome genes	[118]

		2	HiSeq1000	Agilent SureSelect Human All Exon Kit	Hyperdiploid acute lymphoblastic leukemia arises in a pre-B cell in utero, and mutational changes necessary for clinical ALL accumulate subclonally and postnatally	[119]
	T-cell prolymphocytic leukemia (T-PLL)	32	HiSeq2000	NimbleGen EZCap 3	High-frequency mutational activation of the *IL2RG-JAK1-JAK3-STAT5B* axis in the pathogenesis of T-PLL	[120]
	Acute myeloid leukemia (AML)	2	HiSeq2000	NimbleGen 2.1 M Human Exome Array	*TGM6* as a novel familial AML-associated gene	[121]
		1	Illumina Genome Analyzer II	NimbleGen Sequence Capture Human Exome 2.1	A novel somatic mutation in *MMP8* gene	[122]
		1	HiSeq2000	Agilent SureSelect Human All Exon	This study highlights the development of AML in an adult with CBL syndrome	[123]
	Chronic lymphocytic leukemia (CLL)	1	HiSeq2000	SureSelect Human All Exon 50 Mb	*FAT1* gene significantly mutated	[124]
	T-cell large granular lymphocytic (T-LGL)	116	HiSeq2000	NimbleGen SeqCap EZ Exome Library v2.0	Novel somatic mutations in large granular lymphocytic leukemia affecting the STAT-pathway and T-cell activation	[125]
	Infant leukemia (IL)	23	HiSeq2000	TruSeq Exome Enrichment Kit	Comprehensive mutational profiling of infant leukemia	[126]
	Hairy-cell leukemia variant (HCLv)	10	HiSeq2000	Agilent SureSelect Human All Exon	High prevalence of *MAP2K1* mutations in variant and *IGHV4-34-expressing* hairy-cell leukemias	[127]

7 Conclusions

The application of next-generation sequencing has become a powerful tool for cancer research, yielding important biological insights and enabling systematic profiling based on genomic information. Although sequencing entire cancer genomes provides a comprehensive picture of the genome, resequencing an entire genome with high coverage is expensive and also generates enormous amounts of data that are difficult to process; it is also difficult to interpret variants found in noncoding portions of the genome. Exome sequencing offers a cost-effective and viable alternative for cancer genomic applications, which demand sequencing with high coverage from a large number of cancer samples. The major limitations of exome sequencing are that exome capture requires complex library-preparation procedures that are laborious and time consuming and that regions of interest with high or low GC content are difficult to capture. As the cost of sequencing continues to decrease markedly, advances in capturing technologies will simultaneously allow the sequencing of a large number of different cancer samples and thereby advance our knowledge of cancer genomics.

References

1. Boveri T. Concerning the origin of malignant tumours by Theodor boveri translated and annotated by henry Harris. J Cell Sci. 2008;121 Suppl 1:1–84.
2. Stehelin D, Varmus HE, Bishop JM, Vogt PK. DNA related to the transforming gene(s) of avian sarcoma viruses is present in normal avian DNA. Nature. 1976;260:170–3.
3. Taub R, Kirsch I, Morton C, Lenoir G, Swan D, Tronick S, et al. Translocation of the c-myc gene into the immunoglobulin heavy chain locus in human Burkitt lymphoma and murine plasmacytoma cells. Proc Natl Acad Sci U S A. 1982;79:7837–41.
4. Knudson AG. Mutation and cancer: statistical study of retinoblastoma. Proc Natl Acad Sci U S A. 1971;68:820–3.
5. Dulbecco R. A turning point in cancer research: sequencing the human genome. Science. 1986;231:1055–6.
6. International Human Genome Sequencing Consortium. Finishing the euchromatic sequence of the human genome. Nature. 2004;431:931–45.
7. Vassy JL, Lautenbach DM, McLaughlin HM, Kong SW, Christensen KD, Krier J, et al. The MedSeq Project: a randomized trial of integrating whole genome sequencing into clinical medicine. Trials. 2014;15:85.
8. Asan, Xu Y, Jiang H, Tyler-Smith C, Xue Y, Jiang T, et al. Comprehensive comparison of three commercial human whole-exome capture platforms. Genome Biol. 2011;12:R95.
9. Pruitt KD, Tatusova T, Maglott DR. NCBI Reference Sequence (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. Nucleic Acids Res. 2005;33:D501–4.
10. Flicek P, Amode MR, Barrell D, Beal K, Brent S, Carvalho-Silva D, et al. Ensembl 2012. Nucleic Acids Res. 2012;40:D84–90.
11. Pruitt KD, Harrow J, Harte RA, Wallin C, Diekhans M, Maglott DR, et al. The consensus coding sequence (CCDS) project: identifying a common protein-coding gene set for the human and mouse genomes. Genome Res. 2009;19:1316–23.

12. Mamanova L, Coffey AJ, Scott CE, Kozarewa I, Turner EH, Kumar A, et al. Target-enrichment strategies for next-generation sequencing. Nat Methods. 2010;7:111–8.
13. Turner EH, Ng SB, Nickerson DA, Shendure J. Methods for genomic partitioning. Annu Rev Genomics Hum Genet. 2009;10:263–84.
14. Hodges E, Xuan Z, Balija V, Kramer M, Molla MN, Smith SW, et al. Genome-wide in situ exon capture for selective resequencing. Nat Genet. 2007;39:1522–7.
15. Wu W, Choudhry H. Next generation sequencing in cancer research. Berlin: Springer Verlag; 2013.
16. Saiki RK, Gelfand DH, Stoffel S, Scharf SJ, Higuchi R, Horn GT, et al. Primer-directed enzymatic amplification of DNA with a thermostable DNA polymerase. Science. 1988;239:487–91.
17. Bustamante CD, Fledel-Alon A, Williamson S, Nielsen R, Hubisz MT, Glanowski S, et al. Natural selection on protein-coding genes in the human genome. Nature. 2005;437:1153–7.
18. Sjöblom T, Jones S, Wood LD, Parsons DW, Lin J, Barber TD, et al. The consensus coding sequences of human breast and colorectal cancers. Science. 2006;314:268–74.
19. Bashiardes S, Veile R, Helms C, Mardis ER, Bowcock AM, Lovett M. Direct genomic selection. Nat Methods. 2005;2:63–9.
20. Gnirke A, Melnikov A, Maguire J, Rogov P, LeProust EM, Brockman W, et al. Solution hybrid selection with ultra-long oligonucleotides for massively parallel targeted sequencing. Nat Biotechnol. 2009;27:182–9.
21. Noonan JP, Coop G, Kudaravalli S, Smith D, Krause J, Alessi J, et al. Sequencing and analysis of Neanderthal genomic DNA. Science. 2006;314:1113–8.
22. Clark MJ, Chen R, Lam HYK, Karczewski KJ, Chen R, Euskirchen G, et al. Performance comparison of exome DNA sequencing technologies. Nat Biotechnol. 2011;29:908–14.
23. Sulonen A-M, Ellonen P, Almusa H, Lepistö M, Eldfors S, Hannula S, et al. Comparison of solution-based exome capture methods for next generation sequencing. Genome Biol. 2011;12:R94.
24. Parla JS, Iossifov I, Grabill I, Spector MS, Kramer M, McCombie WR. A comparative analysis of exome capture. Genome Biol. 2011;12.
25. Chilamakuri CSR, Lorenz S, Madoui M-A, Vodák D, Sun J, Hovig E, et al. Performance comparison of four exome capture systems for deep sequencing. BMC Genomics. 2014;15:449.
26. Ruffalo M, LaFramboise T, Koyutürk M. Comparative analysis of algorithms for next-generation sequencing read alignment. Bioinformatics. 2011;27:2790–6.
27. Cibulskis K, Lawrence MS, Carter SL, Sivachenko A, Jaffe D, Sougnez C, et al. Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples. Nat Biotechnol. 2013;31:213–9.
28. Koboldt DC, Zhang Q, Larson DE, Shen D, McLellan MD, Lin L, et al. VarScan 2: somatic mutation and copy number alteration discovery in cancer by exome sequencing. Genome Res. 2012;22:568–76.
29. Larson DE, Harris CC, Chen K, Koboldt DC, Abbott TE, Dooling DJ, et al. SomaticSniper: identification of somatic point mutations in whole genome sequencing data. Bioinformatics. 2012;28:311–7.
30. Xu H, DiCarlo J, Satya RV, Peng Q, Wang Y. Comparison of somatic mutation calling methods in amplicon and whole exome sequence data. BMC Genomics. 2014;15:244.
31. Ng PC, Henikoff S. SIFT: Predicting amino acid changes that affect protein function. Nucleic Acids Res. 2003;31:3812–4.
32. Adzhubei IA, Schmidt S, Peshkin L, Ramensky VE, Gerasimova A, Bork P, et al. A method and server for predicting damaging missense mutations. Nat Methods. 2010;7:248–9.
33. Biesecker LG, Green RC. Diagnostic clinical genome and exome sequencing. N Engl J Med. 2014;370:2418–25.
34. Rabbani B, Tekin M, Mahdieh N. The promise of whole-exome sequencing in medical genetics. J Hum Genet. 2014;59:5–15.

35. Ku CS, Cooper DN, Polychronakos C, Naidoo N, Wu M, Soong R. Exome sequencing: dual role as a discovery and diagnostic tool. Ann Neurol. 2012;71:5–14.
36. Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, Baldwin J, et al. Initial sequencing and analysis of the human genome. Nature. 2001;409:860–921.
37. Venter JC, Adams MD, Myers EW, Li PW, Mural RJ, Sutton GG, et al. The sequence of the human genome. Science. 2001;291:1304–51.
38. Wood LD, Parsons DW, Jones S, Lin J, Sjöblom T, Leary RJ, et al. The genomic landscapes of human breast and colorectal cancers. Science. 2007;318:1108–13.
39. Ding L, Getz G, Wheeler DA, Mardis ER, McLellan MD, Cibulskis K, et al. Somatic mutations affect key pathways in lung adenocarcinoma. Nature. 2008;455:1069–75.
40. Greenman C, Stephens P, Smith R, Dalgliesh GL, Hunter C, Bignell G, et al. Patterns of somatic mutation in human cancer genomes. Nature. 2007;446:153–8.
41. Johnson AM, Zuhlke KA, Plotts C, McDonnell SK, Middha S, Riska SM, et al. Mutational landscape of candidate genes in familial prostate cancer. Prostate. 2014;74:1371–8.
42. VanderWeele DJ, Brown CD, Taxy JB, Gillard M, Hatcher DM, Tom WR, et al. Low-grade prostate cancer diverges early from high grade and metastatic disease. Cancer Sci. 2014;105:1079–85.
43. Kim T-M, Jung S-H, Baek I-P, Lee S-H, Choi Y-J, Lee J-Y, et al. Regional biases in mutation screening due to intratumoural heterogeneity of prostate cancer. J Pathol. 2014;233:425–35.
44. Scott AF, Mohr DW, Ling H, Scharpf RB, Zhang P, Liptak GS. Characterization of the genomic architecture and mutational spectrum of a small cell prostate carcinoma. Genes (Basel). 2014;5:366–84.
45. Lohr JG, Adalsteinsson VA, Cibulskis K, Choudhury AD, Rosenberg M, Cruz-Gordillo P, et al. Whole-exome sequencing of circulating tumor cells provides a window into metastatic prostate cancer. Nat Biotechnol. 2014;32:479–84.
46. Spans L, Helsen C, Clinckemalie L, Van den Broeck T, Prekovic S, Joniau S, et al. Comparative genomic and transcriptomic analyses of LNCaP and C4-2B prostate cancer cell lines. PLoS One. 2014;9:e90002.
47. Van Allen EM, Foye A, Wagle N, Kim W, Carter SL, McKenna A, et al. Successful whole-exome sequencing from a prostate cancer bone metastasis biopsy. Prostate Cancer Prostatic Dis. 2014;17:23–7.
48. Menon R, Deng M, Rüenauver K, Queisser A, Peifer M, Pfeifer M, et al. Somatic copy number alterations by whole-exome sequencing implicates YWHAZ and PTK2 in castration-resistant prostate cancer. J Pathol. 2013;231:505–16.
49. Fitzgerald LM, Kumar A, Boyle EA, Zhang Y, McIntosh LM, Kolb S, et al. Germline missense variants in the BTNL2 gene are associated with prostate cancer susceptibility. Cancer Epidemiol Biomarkers Prev. 2013;22:1520–8.
50. Lindberg J, Mills IG, Klevebring D, Liu W, Neiman M, Xu J, et al. The mitochondrial and autosomal mutation landscapes of prostate cancer. Eur Urol. 2013;63:702–8.
51. Barbieri CE, Baca SC, Lawrence MS, Demichelis F, Blattner M, Theurillat J-P, et al. Exome sequencing identifies recurrent SPOP, FOXA1 and MED12 mutations in prostate cancer. Nat Genet. 2012;44:685–9.
52. Spans L, Atak ZK, Van Nieuwerburgh F, Deforce D, Lerut E, Aerts S, et al. Variations in the exome of the LNCaP prostate cancer cell line. Prostate. 2012;72:1317–27.
53. Yu J, Wu WKK, Li X, He J, Li X-X, Ng SSM, et al. Novel recurrently mutated genes and a prognostic mutation signature in colorectal cancer. Gut. 2015;64:636–645.
54. Nieminen TT, O'Donohue M-F, Wu Y, Lohi H, Scherer SW, Paterson AD, et al. Germline mutation of RPS20, encoding a ribosomal protein, causes predisposition to hereditary non-polyposis colorectal carcinoma without DNA mismatch repair deficiency. Gastroenterology. 2014;147:595–8.
55. Mouradov D, Sloggett C, Jorissen RN, Love CG, Li S, Burgess AW, et al. Colorectal cancer cell lines Are representative models of the main molecular subtypes of primary cancer. Cancer Res. 2014;74:3238–47. Available from: http://cancerres.aacrjournals.org/content/early/2014/04/22/0008-5472.CAN-14-0013.

56. Lee SY, Haq F, Kim D, Jun C, Jo H-J, Ahn S-M, et al. Comparative genomic analysis of primary and synchronous metastatic colorectal cancers. PLoS One. 2014;9.
57. Gala MK, Mizukami Y, Le LP, Moriichi K, Austin T, Yamamoto M, et al. Germline mutations in oncogene-induced senescence pathways Are associated with multiple sessile serrated adenomas. Gastroenterology. 2014;146(2):520–9.
58. Tahara T, Yamamoto E, Suzuki H, Maruyama R, Chung W, Garriga J, et al. Fusobacterium in colonic flora and molecular features of colorectal carcinoma. Cancer Res. 2014;74:1311–8.
59. Cajuso T, Hänninen UA, Kondelin J, Gylfe AE, Tanskanen T, Katainen R, et al. Exome sequencing reveals frequent inactivating mutations in ARID1A, ARID1B, ARID2 and ARID4A in microsatellite unstable colorectal cancer. Int J Cancer. 2014;135:611–23.
60. Tahara T, Yamamoto E, Madireddi P, Suzuki H, Maruyama R, Chung W, et al. Colorectal carcinomas with CpG island methylator phenotype 1 frequently contain mutations in chromatin regulators. Gastroenterology. 2014;146:530–5.
61. Gylfe AE, Katainen R, Kondelin J, Tanskanen T, Cajuso T, Hänninen U, et al. Eleven candidate susceptibility genes for common familial colorectal cancer. PLoS Genet. 2013;9:e1003876.
62. Gylfe AE, Kondelin J, Turunen M, Ristolainen H, Katainen R, Pitkänen E, et al. Identification of candidate oncogenes in human colorectal cancers with microsatellite instability. Gastroenterology. 2013;145:540–3. e22.
63. DeRycke MS, Gunawardena SR, Middha S, Asmann YW, Schaid DJ, McDonnell SK, et al. Identification of novel variants in colorectal cancer families by high-throughput exome sequencing. Cancer Epidemiol Biomarkers Prev. 2013;22:1239–51.
64. Zhou D, Yang L, Zheng L, Ge W, Li D, Zhang Y, et al. Exome capture sequencing of adenoma reveals genetic alterations in multiple cellular pathways at the early stage of colorectal tumorigenesis. PLoS One. 2013;8:e53310.
65. Donehower LA, Creighton CJ, Schultz N, Shinbrot E, Chang K, Gunaratne PH, et al. MLH1-silenced and non-silenced subgroups of hypermutated colorectal carcinomas have distinct mutational landscapes. J Pathol. 2013;229:99–110.
66. Network CGA. Comprehensive molecular characterization of human colon and rectal cancer. Nature. 2012;487:330–7.
67. Jhunjhunwala S, Jiang Z, Stawiski EW, Gnad F, Liu J, Mayba O, et al. Diverse modes of genomic alterations in hepatocellular carcinoma. Genome Biol. 2014;15:436.
68. Eichenmüller M, Trippel F, Kreuder M, Beck A, Schwarzmayr T, Häberle B, et al. The genomic landscape of hepatoblastoma and their progenies with HCC-like features. J Hepatol. 2014;61(6):1312–20.
69. Vilarinho S, Zeynep Erson-Omay E, Harmanci AS, Morotti R, Carrion-Grant G, Baranoski J, et al. Pediatric hepatocellular carcinoma due to somatic CTNNB1 and NFE2L2 mutations in the setting of inherited bi-allelic ABCB11 mutations. J Hepatol. 2014;61(5):1178–83.
70. Ahn S-M, Jang SJ, Shim JH, Kim D, Hong S-M, Sung CO, et al. A genomic portrait of resectable hepatocellular carcinomas: Implications of RB1 and FGF19 aberrations for patient stratification. Hepatology. 2014;60(6):1972–82.
71. Pilati C, Letouzé E, Nault J-C, Imbeaud S, Boulai A, Calderaro J, et al. Genomic profiling of hepatocellular adenomas reveals recurrent FRK-activating mutations and the mechanisms of malignant transformation. Cancer Cell. 2014;25:428–41.
72. Jiang J-H, Liu Y-F, Ke A-W, Gu F-M, Yu Y, Dai Z, et al. Clinical significance of the ubiquitin ligase UBE3C in hepatocellular carcinoma revealed by exome sequencing. Hepatology. 2014;59:2216–27.
73. Leduc CA, Crouch EE, Wilson A, Lefkowitch J, Wamelink MMC, Jakobs C, et al. Novel association of early onset hepatocellular carcinoma with transaldolase deficiency. JIMD Rep. 2014;12:121–7.
74. Ikeda A, Shimizu T, Matsumoto Y, Fujii Y, Eso Y, Inuzuka T, et al. Leptin receptor somatic mutations are frequent in HCV-infected cirrhotic liver and associated with hepatocellular carcinoma. Gastroenterology. 2014;146:222–35.
75. Cleary SP, Jeck WR, Zhao X, Chen K, Selitsky SR, Savich GL, et al. Identification of driver genes in hepatocellular carcinoma by exome sequencing. Hepatology. 2013;58:1693–702.

76. Cefalù AB, Pirruccello JP, Noto D, Gabriel S, Valenti V, Gupta N, et al. A novel APOB mutation identified by exome sequencing cosegregates with steatosis, liver cancer, and hypocholesterolemia. Arterioscler Thromb Vasc Biol. 2013;33:2021–5.
77. Liu YX, Zhang SF, Ji YH, Guo SJ, Wang GF. Whole-exome sequencing identifies mutated PCK2 and HUWE1 associated with carcinoma cell proliferation in a hepatocellular carcinoma patient. Oncology. 2012;4(4):847–51.
78. Huang J, Deng Q, Wang Q, Li K-Y, Dai J-H, Li N, et al. Exome sequencing of hepatitis B virus-associated hepatocellular carcinoma. Nat Genet. 2012;44:1117–21.
79. Guichard C, Amaddeo G, Imbeaud S, Ladeiro Y, Pelletier L, Maad IB, et al. Integrated analysis of somatic mutations and focal copy-number changes identifies key genes and pathways in hepatocellular carcinoma. Nat Genet. 2012;44:694–8.
80. Totoki Y, Tatsuno K, Yamamoto S, Arai Y, Hosoda F, Ishikawa S, et al. High-resolution characterization of a hepatocellular carcinoma genome. Nat Genet. 2011;43:464–9.
81. Park DJ, Tao K, Le Calvez-Kelm F, Nguyen-Dumont T, Robinot N, Hammet F, et al. Rare mutations in RINT1 predispose carriers to breast and Lynch syndrome-spectrum cancers. Cancer Discov. 2014;4:804–15.
82. Lim WK, Ong CK, Tan J, Thike AA, Ng CCY, Rajasegaran V, et al. Exome sequencing identifies highly recurrent MED12 somatic mutations in breast fibroadenoma. Nat Genet. 2014;46:877–80.
83. Klevebring D, Rosin G, Ma R, Lindberg J, Czene K, Kere J, et al. Sequencing of breast cancer stem cell populations indicates a dynamic conversion between differentiation states in vivo. Breast Cancer Res. 2014;16:R72.
84. Wen H, Kim YC, Snyder C, Xiao F, Fleissner EA, Becirovic D, et al. Family-specific, novel, deleterious germline variants provide a rich resource to identify genetic predispositions for BRCAx familial breast cancer. BMC Cancer. 2014;14:470.
85. Li H, Yang B, Xing K, Yuan N, Wang B, Chen Z, et al. A preliminary study of the relationship between breast cancer metastasis and loss of heterozygosity by using exome sequencing. Sci Rep. 2014;4:5460.
86. Xiao F, Kim YC, Snyder C, Wen H, Chen PX, Luo J, et al. Genome instability in blood cells of a BRCA1+ breast cancer family. BMC Cancer. 2014;14:342.
87. Jiang Y-Z, Yu K-D, Peng W-T, Di G-H, Wu J, Liu G-Y, et al. Enriched variations in TEKT4 and breast cancer resistance to paclitaxel. Nat Commun. 2014;5:3802.
88. Sokolenko AP, Bulanova DR, Iyevleva AG, Aleksakhina SN, Preobrazhenskaya EV, Ivantsov AO, et al. High prevalence of GPRC5A germline mutations in BRCA1-mutant breast cancer patients. Int J Cancer. 2014;134:2352–8.
89. Robinson DR, Wu Y-M, Vats P, Su F, Lonigro RJ, Cao X, et al. Activating ESR1 mutations in hormone-resistant metastatic breast cancer. Nat Genet. 2013;45:1446–51.
90. Network CGA. Comprehensive molecular portraits of human breast tumours. Nature. 2012;490:61–70.
91. Banerji S, Cibulskis K, Rangel-Escareño C, Brown KK, Carter SL, Frederick AM, et al. Sequence analysis of mutations and translocations across breast cancer subtypes. Nature. 2012;486:405–9.
92. Ellis MJ, Ding L, Shen D, Luo J, Suman VJ, Wallis JW, et al. Whole-genome analysis informs breast cancer response to aromatase inhibition. Nature. 2012;486:353–60.
93. Snape K, Ruark E, Tarpey P, Renwick A, Turnbull C, Seal S, et al. Predisposition gene identification in common cancers by exome sequencing: insights from familial breast cancer. Breast Cancer Res Treat. 2012;134(1):429–33.
94. Park DJ, Lesueur F, Nguyen-Dumont T, Pertesi M, Odefrey F, Hammet F, et al. Rare mutations in XRCC2 increase the risk of breast cancer. Am J Hum Genet. 2012;90:734–9.
95. Yin S, Yang J, Lin B, Deng W, Zhang Y, Yi X, et al. Exome sequencing identifies frequent mutation of MLL2 in non-small cell lung carcinoma from Chinese patients. Sci Rep. 2014;4:6036.
96. Renieri A, Mencarelli MA, Cetta F, Baldassarri M, Mari F, Furini S, et al. Oligogenic germline mutations identified in early non-smokers lung adenocarcinoma patients. Lung Cancer. 2014;85:168–74.

97. Chapman MA, Lawrence MS, Keats JJ, Cibulskis K, Sougnez C, Schinzel AC, et al. Initial genome sequencing and analysis of multiple myeloma. Nature. 2011;471:467–72.
98. Fernandez-Cuesta L, Peifer M, Lu X, Sun R, Ozretić L, Seidel D, et al. Frequent mutations in chromatin-remodelling genes in pulmonary carcinoids. Nat Commun. 2014;5:3518.
99. Ahn JW, Kim HS, Yoon J-K, Jang H, Han SM, Eun S, et al. Identification of somatic mutations in EGFR/KRAS/ALK-negative lung adenocarcinoma in never-smokers. Genome Med. 2014;6:18.
100. He Y, Li Y, Qiu Z, Zhou B, Shi S, Zhang K, et al. Identification and validation of PROM1 and CRTC2 mutations in lung cancer patients. Mol Cancer. 2014;13:19.
101. Zhao Y, Yang J, Chen Z, Gao Z, Zhou F, Li X, et al. Identification of somatic alterations in stage I lung adenocarcinomas by next-generation sequencing. Genes Chromosomes Cancer. 2014;53:289–98.
102. Ni X, Zhuo M, Su Z, Duan J, Gao Y, Wang Z, et al. Reproducible copy number variation patterns among single circulating tumor cells of lung cancer patients. Proc Natl Acad Sci. 2013;110:21083–8.
103. Beauchamp EM, Woods BA, Dulak AM, Tan L, Xu C, Gray NS, et al. Acquired resistance to dasatinib in lung cancer cell lines conferred by DDR2 gatekeeper mutation and NF1 loss. Mol Cancer Ther. 2014;13:475–82.
104. Rudin CM, Durinck S, Stawiski EW, Poirier JT, Modrusan Z, Shames DS, et al. Comprehensive genomic analysis identifies SOX2 as a frequently amplified gene in small-cell lung cancer. Nat Genet. 2012;44:1111–6.
105. Liu P, Morrison C, Wang L, Xiong D, Vedell P, Cui P, et al. Identification of somatic mutations in non-small cell lung carcinomas using whole-exome sequencing. Carcinogenesis. 2012;33:1270–6.
106. Taylor BS, DeCarolis PL, Angeles CV, Brenet F, Schultz N, Antonescu CR, et al. Frequent alterations and epigenetic silencing of differentiation pathway genes in structurally rearranged liposarcomas. Cancer Discov. 2011;1:587–97.
107. Byun M, Abhyankar A, Lelarge V, Plancoulaine S, Palanduz A, Telhan L, et al. Whole-exome sequencing-based discovery of STIM1 deficiency in a child with fatal classic Kaposi sarcoma. J Exp Med. 2010;207:2307–12.
108. Belani R, Oliveira G, Erikson GA, Ra S, Schechter MS, Lee JK, et al. ASXL1 and DNMT3A mutation in a cytogenetically normal B3 thymoma. Oncogenesis. 2014;3:e111.
109. Behjati S, Tarpey PS, Sheldon H, Martincorena I, Van Loo P, Gundem G, et al. Recurrent PTPRB and PLCG1 mutations in angiosarcoma. Nat Genet. 2014;46:376–9.
110. Park S, Lee J, Do I-G, Jang J, Rho K, Ahn S, et al. Aberrant CDK4 amplification in refractory rhabdomyosarcoma as identified by genomic profiling. Sci Rep. 2014;4:3623.
111. Szuhai K, de Jong D, Leung WY, Fletcher CDM, Hogendoorn PCW. Transactivating mutation of the MYOD1 gene is a frequent event in adult spindle cell rhabdomyosarcoma. J Pathol. 2014;232:300–7.
112. Bolli N, Avet-Loiseau H, Wedge DC, Van Loo P, Alexandrov LB, Martincorena I, et al. Heterogeneity of genomic evolution and mutational profiles in multiple myeloma. Nat Commun. 2014;5:2997.
113. Walker BA, Wardell CP, Melchor L, Brioli A, Johnson DC, Kaiser MF, et al. Intraclonal heterogeneity is a critical early event in the development of myeloma and precedes the development of clinical symptoms. Leukemia. 2014;28:384–90.
114. Leich E, Weißbach S, Klein H-U, Grieb T, Pischimarov J, Stühmer T, et al. Multiple myeloma is affected by multiple and heterogeneous somatic mutations in adhesion- and receptor tyrosine kinase signaling molecules. Blood Cancer J. 2013;3:e102.
115. Vater I, Montesinos-Rongen M, Schlesner M, Haake A, Purschke F, Sprute R, et al. The mutational pattern of primary lymphoma of the central nervous system determined by whole-exome sequencing. Leukemia. 2015;29(3):677–685.
116. Nikolaev SI, Garieri M, Santoni F, Falconnet E, Ribaux P, Guipponi M, et al. Frequent cases of RAS-mutated Down syndrome acute lymphoblastic leukaemia lack JAK2 mutations. Nat Commun. 2014;5:4654.

117. Papaemmanuil E, Rapado I, Li Y, Potter NE, Wedge DC, Tubio J, et al. RAG-mediated recombination is the predominant driver of oncogenic rearrangement in ETV6-RUNX1 acute lymphoblastic leukemia. Nat Genet. 2014;46:116–25.
118. Herold T, Metzeler KH, Vosberg S, Hartmann L, Röllig C, Stölzel F, et al. Isolated trisomy 13 defines a homogeneous AML subgroup with high frequency of mutations in spliceosome genes and poor prognosis. Blood. 2014;124:1304–11.
119. Bateman CM, Alpar D, Ford AM, Colman SM, Wren D, Morgan M, et al. Evolutionary trajectories of hyperdiploid ALL in monozygotic twins. Leukemia. 2014;29(1):58–65.
120. Kiel MJ, Velusamy T, Rolland D, Sahasrabuddhe AA, Chung F, Bailey NG, et al. Integrated genomic sequencing reveals mutational landscape of T-cell prolymphocytic leukemia. Blood. 2014;124:1460–72.
121. Pan L-L, Huang Y-M, Wang M, Zhuang X-E, Luo D-F, Guo S-C, et al. Positional cloning and next-generation sequencing identified a TGM6 mutation in a large Chinese pedigree with acute myeloid leukaemia. Eur J Hum Genet. 2015;23(2):218–23.
122. Kim Y, Schulz VP, Satake N, Gruber TA, Teixeira AM, Halene S, et al. Whole-exome sequencing identifies a novel somatic mutation in MMP8 associated with a t(1;22)-acute megakaryoblastic leukemia. Leukemia. 2014;28:945–8.
123. Becker H, Yoshida K, Blagitko-Dorfs N, Claus R, Pantic M, Abdelkarim M, et al. Tracing the development of acute myeloid leukemia in CBL syndrome. Blood. 2014;123:1883–6.
124. Messina M, Del Giudice I, Khiabanian H, Rossi D, Chiaretti S, Rasi S, et al. Genetic lesions associated with chronic lymphocytic leukemia chemo-refractoriness. Blood. 2014;123:2378–88.
125. Andersson EI, Rajala HLM, Eldfors S, Ellonen P, Olson T, Jerez A, et al. Novel somatic mutations in large granular lymphocytic leukemia affecting the STAT-pathway and T-cell activation. Blood Cancer J. 2013;3:e168.
126. Valentine MC, Linabery AM, Chasnoff S, Hughes AEO, Mallaney C, Sanchez N, et al. Excess congenital non-synonymous variation in leukemia-associated genes in MLL- infant leukemia: a Children's Oncology Group report. Leukemia. 2014;28:1235–41.
127. Waterfall JJ, Arons E, Walker RL, Pineda M, Roth L, Killian JK, et al. High prevalence of MAP2K1 mutations in variant and IGHV4-34-expressing hairy-cell leukemias. Nat Genet. 2014;46:8–10.

The Landscape of DNA Virus Associations Across Human Cancers

Jian Chen, Lopa Mishra, and Xiaoping Su

Abstract The human cancer viruses have been found to cause 10–15 % of all human malignancies. High-risk human papillomaviruses (HPV-16 and HPV-18), hepatitis B virus (HBV), and Epstein–Barr virus (EBV) contribute directly to cancer development, including cervical squamous cell cancer, hepatocellular carcinoma, Burkitt's lymphoma, nasopharyngeal carcinoma, adult T cell leukemia, and Kaposi's sarcoma. The role of human cancer viruses in cancer pathogenesis is mediated through various mechanisms, including mutagenic integration into the host genome and expression of oncogenic viral proteins. The elucidation of such mechanisms has played a key role in enhancing our understanding of cancer pathogenesis even as novel aspects of DNA virus biology continue to be unraveled. In this chapter, we give an insight of the main events responsible for the development of malignant tumors upon viral infection. With the availability of high-throughput sequencing and robust bioinformatics tools, it is possible to establish a landscape of viral integration into human cancer genome. Thus, we highlight the utility of RNA-Seq in detecting tumor-associated DNA viruses and identifying viral integration sites that may unravel novel mechanisms of cancer pathogenesis. And we also describe a robust bioinformatics tool VirusSeq and its advantages in this field of study.

1 Introduction

The human cancer viruses have been found to cause 10–15 % of all human malignancies [1, 2]. Seven human viruses, high-risk human papillomaviruses (HPV-16 and HPV-18), hepatitis B virus (HBV), hepatitis C virus (HCV), Epstein–Barr virus (EBV), human T-lymphotropic virus-I (HTLV-I), Kaposi's sarcoma herpesvirus

J. Chen • L. Mishra
Department of Gastroenterology, Hepatology, and Nutrition, The University of Texas MD Anderson Cancer Center, 1515 Holcombe Boulevard, Houston, TX 77030, USA

X. Su (✉)
Department of Bioinformatics and Computational Biology, The University of Texas MD Anderson Cancer Center, 1515 Holcombe Boulevard, Houston, TX 77030-4009, USA
e-mail: xsu1@mdanderson.org

W. Wu, H. Choudhry (eds.), *Next Generation Sequencing in Cancer Research, Volume 2*, DOI 10.1007/978-3-319-15811-2_18

(KSHV), and Merkel cell polyomavirus (MCV), contribute significantly to cancer development worldwide, including cervical cancer, hepatocellular carcinoma, Burkitt's lymphoma, nasopharyngeal carcinoma, adult T cell Leukemia, Kaposi's sarcoma, and Merkel cell carcinoma [2–10]. The role of human cancer viruses in cancer pathogenesis is mediated through various mechanisms, including, for example, mutagenic integration into the host genome and expression of oncogenic viral proteins [11, 12]. The elucidation of such mechanisms has played a key role in enhancing our understanding of cancer pathogenesis even as novel aspects of DNA virus biology continue to be unraveled [13].

Viral etiology is particularly evident in cervical squamous cell carcinoma (CESC), which is almost exclusively caused by high-risk human papillomaviruses (HPV), and in hepatocellular carcinoma (HCC), where infection with hepatitis B virus (HBV) or hepatitis C virus (HCV) is the predominant cause in some countries [14]. In addition, several rare cancers have a strong viral component, including Epstein–Barr virus (EBV)/human herpesvirus (HHV) four in most Burkitt's lymphomas [1]. Huge advances in the prevention of virus-associated cancer have been made through vaccination programs against HPV and HBV, second only to smoke cessation in the number of yearly cancer cases prevented worldwide [15].

One of the best understood causal relationships is between human papillomavirus (HPV) infection and squamous neoplasia of the anogenital and head-and-neck regions [1]. Infection with HPV generally gives rise to warts. Certain types have a strong association with cervical intraepithelial neoplasia (types 6 and 11) whereas other types (principally, types 16 and 18) are present in over 90 % of tumors. There are many reports of different papillomavirus types being detected in cancers of the head, neck, and mouth [1]. Although initially it was thought that HPV16 was confined to the genital tract there are reports of HPV16 being found in the other tumors. Genital strains are closely associated with cervical intraepithelial neoplasia and cervical, vulval, and anogenital cancer, diagnosed by abnormal cytology and pathology. It is on the cervical and anogenital cancers that work has concentrated although cancers of the head, neck, and oral cavity have also a strong association with certain HPV strains, including HP16. HPV can immortalize keratinocytes, but another step is required for full transformation to an oncogenic phenotype. HPV does not code for a virus-encoded oncogene. Oncogenesis is associated with the two early proteins E6 and E7 which, respectively, bind to cell cycle control gene products, the Rb and p53 proteins. In epidermodysplasia verruciformis, host-mediated immunity is significantly impaired. This disease is mainly associated with human papillomavirus types 5 and 8. The exact role of the host immune response in patients with cervical intraepithelial neoplasia and cervical carcinoma is not clear but increases in both. Precancerous lesions and cervical cancer occur in immunosuppressed patients. Proliferation of peripheral blood lymphocytes is observed after stimulation with HPV16 L1, and E6 and E7 proteins. Rodents immunized with L1, E6, or E7 are protected against syngeneic tumor transplants transfected with L1, E6, or E7 by CD8+ lymphocytes. In cervical carcinoma, the human papillomavirus genome is usually detected as an integrated fragment. Deletions do occur but E1, E6, and E7 are retained and can be expressed. Recommendations for prevention include the use of condoms and avoidance of early age of first intercourse and of multiple sexual

partners. Vaccine production is currently under consideration as Bovine papillomavirus type 4 (BPV-4) vaccine has been useful in controlling cancer of the bovine alimentary tract. Several suitable T cell recognition epitopes have been located in E6 and E7 and may be suitable for peptide vaccines.

HCC is one of the leading causes of cancer-related mortality in the world and is strongly associated with chronic HBV or HCV infection [16]. Hepatitis virus infection causes chronic liver injury and subsequent progression to severe fibrosis and cirrhosis. The presence of cirrhosis is a major risk factor for the development of HCC. However, HCC can occur in the absence of cirrhosis, suggesting that both HBV and HCV may be directly involved in hepatocarcinogenesis. Chronic HBV infection accounts for approximately 50 % of all cases of hepatocellular carcinoma and virtually all childhood cases. The HBV genome was frequently detected in chronic hepatitis B carriers and patients with HCC [17, 18]. The integration of HBV into the host genome induces DNA deletions, translocations, and mutations in various chromosome positions. In contrast to HBV, HCV is an RNA virus that is unable to reverse transcribe to DNA. Various HCV proteins, including the core, envelope, and nonstructural protein, have been shown to possess oncogenic properties [19]. It has been reported that proteins encoded by HCV RNA are involved in the manipulation of diverse cellular functions, including apoptosis, proliferation, endoplasmic reticulum (ER) stress, etc.

The Epstein–Barr virus (EBV) (also known as human herpesvirus 4; HHV4) is a DNA virus that infects over 90 % of the world's population before adolescence. It has been associated with a wide variety of human malignancies of epithelial, hematolymphoid, and mesenchymal derivation [11, 20, 21]. Gastric carcinoma associated with EBV appears to comprise a distinct entity that is predominant in younger male individuals [22, 23]. This subset of gastric carcinoma, 8–10 % of cases, is more prevalent in Caucasian and Hispanic patients than Asians, and it shows no association with *Helicobacter pylori* infection [22]. In these cases, EBV appears to play a direct oncogenic role through genome-wide alteration of promoter methylation [24], microRNA (miRNA) expression, and expression of genes involved in cell motility and transformation pathways [25]. The integration status of EBV in gastric carcinoma remains poorly understood.

Our current knowledge of virus–tumor associations is based largely on data gathered with low-throughput methodologies in the pregenomic era. However, massively parallel sequencing is now showing promise for efficient unbiased detection of viruses in tumor tissue. This recently led to the discovery of a new polyomavirus as the cause of most Merkel cell carcinomas [9], where essential virus–host interactions are currently being targeted in clinical drug trials [26]. Recent studies describe techniques for detection of viruses using high-throughput RNA or DNA sequencing [27, 28], and massively parallel sequencing has been used to survey sites of genomic integration of HBV in hepatocellular carcinoma [17, 18]. Recently, viral integration sites were mapped in 17 hepatocellular carcinoma (HCC) and 239 head and neck carcinomas by detecting host–virus fusions in transcriptome sequencing (RNA-seq) data from The Cancer Genome Atlas (TCGA) [29]. These studies provided important insights and clearly demonstrate the potential of the methodology and motivate a broad unbiased survey of viral expression and integration in human cancer.

2 Human Papillomavirus in Malignant Cancers

HPV is a small, 50- to 55-nm-diameter, nonenveloped, double-stranded DNA virus that carries out its life cycle in either mucosal or cutaneous stratified squamous epithelia [30]. The viral genome (8 kb in size) is amplified initially as extrachromosomal circular elements (episomes) but may eventually integrate into the host genome. Over 120 types of HPV have been identified, of which those capable of infecting humans are designated high risk or low risk on the basis of their association with human neoplasms and oncogenic potential. The oncoproteins E5, E6, and E7 are the primary agents responsible for initiation and progression of HPV-associated cancers, and they operate primarily by abrogating negative growth regulators and inducing genomic instability. The integration of HPV DNA into the host cell genome is considered an important step in malignant progression and is commonly identified in noninvasive and invasive carcinomas associated with high-risk types HPV16 and HPV18 [31–33]. HPV integration sites, with a predilection for sites of known genomic fragility, have been found to be distributed randomly over the whole genome in one study [34], and the majority of integrated HPV genomes appear to be actively transcribed [35, 36].

Two hundred and thirty-nine squamous cell carcinomas of the head-and-neck region (HNSCC) available in the TCGA database were analyzed [37]. HPV transcripts were detected in 36 tumors as the following: 30 tumors with HPV16, five tumors with HPV33, and one tumor with HPV35. Among all cases with HPV transcripts, E7 was expressed in 22, E6 in 20, E1 in 17, and E4 in eight tumors. In 24 tumors, HPV transcripts encoding key viral proteins/oncoproteins were integrated in the tumor genome, with the majority in association with known genes (Figs. 1 and 2). Tumors with HPV integration harbored the following types: 19 HPV16, 4 HPV33, and 1 HPV35. Of the tumors with HPV integration, 18 have both E6 and E7 integration sites, four have only E7 integration sites, and two have only E6 integration sites. The detected HPV status correlated with perfect sensitivity and specificity with known clinicopathologic variables and with established methods for HPV detection (colorimetric in situ hybridization and/or p16ink4a expression) [38]. For this HNSCC data set, the sensitivity for HPV16 detection was 100 %, with 95 % confidence intervals (CI) of 67.6–100 %, and specificity for HPV16 detection was 100 %, with 95 % CI of 90.4–100 %, as reported previously [39].

HPV16 was also detected in two tumors of histologic types rarely associated with this virus. From a group of 219 lung squamous cell carcinoma tumors, one case harbored HPV16, where E1, E6, and E7 transcripts were highly expressed and integrated in NROB1 (also known as DAX1), a gene involved in steroidogenesis and cell cycle regulation. In this case, no E2, E4, E5, L1, or L2 transcripts were detected. Additionally, one of 253 endometrial carcinoma tumors harbored HPV16, where E1, E2, E4, E5, E6, and E7 transcripts were highly expressed, and no L1 or L2 transcripts were detected.

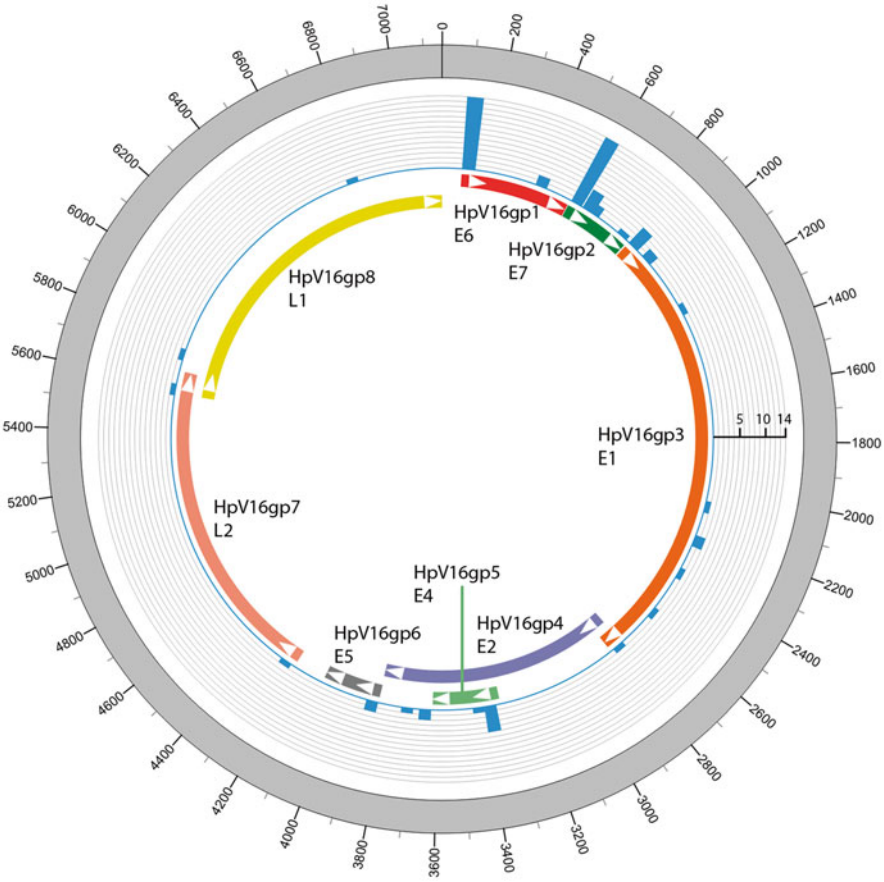


Fig. 1 Visualization of HPV16 integration breakpoints in the HPV16 genome. The frequency of integration breakpoints at different loci in the HPV16 genome is shown as a *blue* histogram. The scale bar indicates the number of tumors. The locations of the genes encoding HPV16 E6 (*red*), E7 (*dark green*), E1 (*orange*), E2 (*purple*), E4 (*green*), E5 (*gray*), L2 (*light orange*), and L1 (*yellow*) proteins are shown. Genomic positions are numbered

Recently, a large survey also showed that 96.6 % of Cervical Squamous Cell Carcinomas (CESC) is associated with HPV [40]. Twelve HPV types, all previously described as associated, were found in 84 positive cervical tumors, with HPV16 and HPV18 expectedly being predominant (65.5 and 13.1 % of positive cases, respectively). Head and neck squamous cell carcinoma showed 14.1 % HPV association, with 83.7 and 14.0 % of positive tumors attributed to HPV16 and HPV33, respectively. Less common but previously observed associations included HPV6b and high-risk types in bladder urothelial carcinoma (BLCA) and uterine endometroid carcinoma (UCEC) [40].

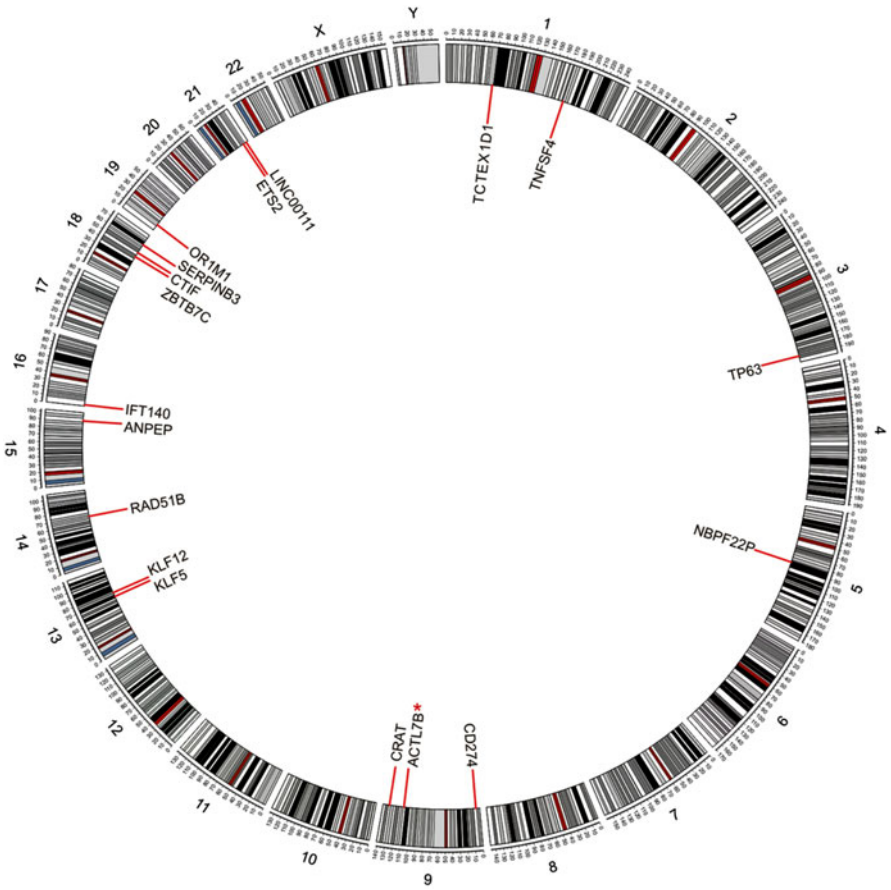


Fig. 2 Integration sites of HPV16 in head-and-neck squamous cell carcinoma tumors in the human genome (hg19). Chromosome numbers are shown (*, detected in two cases)

3 HBV in HCC

The incidence of HCC is rising at an alarming rate in the United States and worldwide; it is predicted to be the cause of 100 million deaths through the twenty-first century (GLOBOCAN 2012 updated June 18, 2014: Estimated Cancer Incidence, Mortality and Prevalence Worldwide in 2012 http://globocan.iarc.fr/Pages/fact_sheets_population.aspx). Chronic viral hepatitis and cirrhosis are major risk factors for HCC. An estimated two billion people worldwide are infected with Hepatitis B, and a further 3 % of the world's population is infected with Hepatitis C Virus (HCV). Identifying a small manageable subset of high-risk patients and developing

nontoxic chemopreventive strategies is urgent. The current lack of translational progress in HCC can be attributed mainly to the difficulty of recruiting a large number of HCC patients, a factor of natural history, and clinical features of the disease.

The HBV virion consists of partially double-stranded DNA packaged with a core protein (HBcAg) and DNA polymerase within envelope proteins (HBsAg) [41]. Integration of viral DNA into the genome of HCC cells has been demonstrated in several studies, and insertional mutagenesis has been identified as a critical step in HBV-mediated HCC pathogenesis [17, 42]. Integration sites were initially thought to be distributed randomly throughout the host genome, but data supporting a more deliberate process that preferentially involves transcribed regions of critical genes have been reported [43–48]. A genome-wide association study identified 1p36.22 as a new susceptibility locus for hepatocellular carcinoma in chronic hepatitis B virus carriers [49]. Most recently, it was reported that genetic variants in STAT4 and HLA-DQ genes confer risk of hepatitis B virus-related hepatocellular carcinoma [50].

Recently, we analyzed 69 HCC tumors available in the TCGA database and detected HBV transcripts in 16 (23 %) tumors [37]. Eight of these patients had serologic evidence of HBV infection, and one patient was HBV (and hepatitis C virus) seronegative. Serologic data on the remaining patients were either negative or not available. Virus integration was identified in 15 (94 %) of these tumors. Our data demonstrate frequent HBV integration within previously identified genes, namely, TERT (5 tumors) and MLL4 (3 tumors), suggesting that these sites are particularly susceptible to HBV insertion. Integration of two or more HBV genes was detected in eight tumors, whereas in the remaining seven tumors integration of only one HBV gene was detected. Interestingly, the latter group included the three tumors with MLL4 involvement and two with TERT involvement. Several HBV factors have also been implicated in hepatocarcinogenesis, including the HBx gene, the pre-S2/S gene, and the HBV spliced protein [17]. HBx is indispensable in hepatocarcinogenesis and only promotes persistent viral infection by enhancing HBV gene expression and replication, but also leads to genome instability through suppression of p53-regulated DNA repair [51]. Other insertion sites were restricted to single cases. Of the tumors with HBV integration, 11 have X protein integration sites, eight have S protein integration sites, six have core/E antigen integration sites, four have pre-S protein integration sites, four have polymerase I integration sites, and three have polymerase two integration sites. Integration of two or more HBV genes was detected in eight tumors, whereas in the remaining seven tumors integration of only one HBV gene was detected. Interestingly, the latter group included the three tumors with MLL4 involvement and two with TERT involvement. We additionally identified HBV transcripts in one case among 460 clear-cell renal cell carcinoma tumors analyzed. In this tumor, as well as in one HCC tumor, we detected HBV S protein transcripts integrated into GLI2, generally considered a marker of activation of the sonic hedgehog signaling pathway, which has been shown to play a role in HCC [51, 52].

4 EBV in Gastric Carcinoma

The Epstein–Barr virus (EBV) (also known as human herpesvirus 4; HHV4) is a double-stranded DNA virus that infects over 90 % of the world population before adolescence. This virus has been associated with a wide variety of human cancers: undifferentiated nasopharyngeal cancer, gastric carcinoma, Burkitt's lymphoma, and Hodgkin's disease, wherein the virus persists in latent phase usually in episomal form. It remains unclear whether EBV integration into the host genome plays a role in cancer. A subset of gastric carcinomas is associated with EBV, and in such tumors the virus has been associated with genome-wide alteration of host promoter methylation, miRNA expression, and expression of genes involved in cell motility and transformation pathways.

We recently analyzed 71 cases of gastric carcinoma in the TCGA database and detected EBV transcripts in four (5.6 %) tumors [37]. Of the four tumors with unequivocal EBV association, all harbored transcripts encoding A73, RPMS1, BARF0, BALF3, BALF4, BALF5, LF1, LF2, and BILF1. In the single head and neck squamous cell carcinoma tumor associated with EBV, the most abundant transcripts included those that encode BARF1/2, BdRF1, BMRF2, BLF1, BNLF2b, BBLF1, BMRF1, BLRF2, and BNLF2a. None of the tumors analyzed in this study had evidence of EBV integration into the host genome.

5 Detection of DNA Viruses and Their Integration Sites by Next-Generation Sequencing

About 12 % of all human cancers are known to be caused by viruses [53], thus, the detection of viruses in human cancer tissue has significant clinical implications in oncology. The advent of next-generation sequencing (NGS) technologies using paired-end reads allows for the detection of viruses in human cancer tissue at unprecedented levels of efficiency and precision. Several groups have developed computational tools for pathogen/virus discovery by exploiting the great amount of NGS data obtained from human tissue [27, 54]. These groups have implemented a computational subtraction analysis, which has also been used to discover a new polyomavirus associated with most cases of Merkel cell carcinoma [9]. Although detecting viruses in human tissue is important in clinical oncology, investigating virus integration sites in host cell chromosomes is equally valuable since insertional mutagenesis is one of the most critical steps in the pathogenesis of HBV-mediated HCC [47]. NGS data have been used to map the HBV integration sites in HCC samples [17, 18].

We developed VirusSeq (Fig. 3) for detection of DNA viruses and their integration sites using next-generation sequencing of human cancer tissues [29, 37]. VirusSeq starts with computational subtraction of human sequences followed by generation of a set of nonhuman sequences (e.g., viruses, etc.) on NGS data.

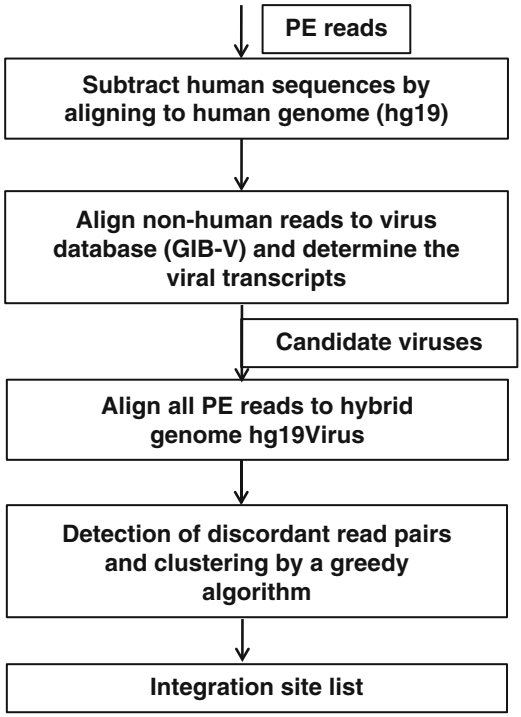


Fig. 3 VirusSeq workflow

Once raw PE reads from whole genome/transcriptome resequencing are aligned/mapped to human genome reference, any read with more than half read length mapped to the human reference genome is removed along with its paired mate in this subtraction step. Thus, a set of nonhuman sequences is generated after human sequence subtraction. In the second step, VirusSeq determines whether the nonhuman sequences match any known viral sequences by searching a comprehensive database that includes all known viral sequences (Genome Information Broker for Viruses; GIB-V, http://gib-v.genes.nig.ac.jp/) and quantifies virus representation by a measure of the virus genome coverage (or overall count of mapped reads) to determine the existence of viruses in human samples. Furthermore, VirusSeq excludes nontranscribed viral genome elements to eliminate/reduce the potential of nonsense reads or inclusion of nontranscribed viral genomic elements. The expression level of each viral transcript is measured by the normalized depth of coverage within each viral transcript. The cutoff of viral gene expression detection is empirically determined by profiling the distribution of viral gene expression levels across multiple cancer-associated viruses (e.g., HPV16, HPV33, EBV) and multiple patient samples. Any viral expression level below cutoff is treated as no expression.

VirusSeq is also able to detect virus integration sites: The genomes of viruses with detectable expression level detected in previous steps are concatenated into a single genome named chrVirus with related annotation of each virus in refFlat format [29]. A new hybrid reference genome named hg19Virus is built by combining

hg19 and chrVirus. All paired-end reads without computational subtraction are again mapped to this reference (hg19Virus). If the paired-end reads are uniquely mapped with one end to hg19 and with the other end to chrVirus, it is reported as a discordant read pair. All discordant reads are annotated by using the genes and viruses defined in the curated refFlat file. VirusSeq then clusters the remaining discordant read pairs that support the same integration (fusion) event (e.g., HBV-MLL4) and selects them as fusion candidates. VirusSeq implements a greedy search-based dynamic clustering process to accurately determine the exact fusion junction between human gene and virus. Specifically, the boundary for each discordant read cluster of candidate fusion is estimated on the basis of discordant read mapping locations and orientations with fragment length distribution as a constraint of cluster size, which is measured by using reads' genomic location excluding intronic sizes if mapped reads are located across adjacent exons in a candidate fusion. For the forward-aligned discordant reads in a fusion candidate, the clustering process starts with the most right read, and the genomic coordinate for the most right read is used to define the in silico fusion junction excluding the outliers within the discordant read cluster. In order to remove outliers within a cluster, VirusSeq implements the robust "extreme studentized deviate" (ESD) multiple-outlier procedure. If the outliers come from right end of cluster, the outliers are removed and the clustering process restarts with new in silico fusion junction after exclusion of outliers. If the outliers come from left end of cluster, the cluster size is reset with in silico fusion junction intact by excluding the outlier reads. For the reverse-aligned discordant reads, the clustering process starts with the most left read, and the genomic coordinate for the most left read is used to define the in silico fusion junction with same outlier detection/removal processing step. For either side of candidate fusion partner (gene vs. virus), this clustering process is performed independently. This greedy search-based dynamic clustering process accurately determines the exact fusion junction between human gene and virus. Meanwhile, an in silico sequence by using the consensus of reads within discordant read clusters for each fusion candidate is generated to help PCR primer design, which facilitates quick PCR validation.

VirusSeq was used to analyze RNA-Seq data of 239 cases of HNSCC available in the TCGA database with the sensitivity at 100 % (8/8) with a 95 % CI of 67.6–100 %, and the specificity at 100 % (36/36) with a 95 % CI of 90.4–100 %.

6 Discussions

The pathogenetic role of DNA viruses has been well established in some cancers, particularly in squamous cell neoplasms of the anogenital and head-and-neck regions and hepatocellular carcinoma. The quest to identify similar associations in other malignant cancers has consumed significant efforts over the past decades. One of the key findings so far by next-generation sequencing of TCGA tumor samples [37, 40] is the absence of an association between all known DNA viruses and some of the most prevalent human cancers, including acute myeloid leukemia; cutaneous melanoma;

low- and high-grade gliomas of the brain; and adenocarcinomas of the breast, colon and rectum, lung, prostate, ovary, kidney, and thyroid. Based on these results and unless novel pathogenic DNA viruses are discovered, we believe that the yield of future searches for DNA viruses in these types of cancers is likely to be very low.

The capacity of our algorithm VirusSeq [29] to detect viral integration points within the host genome is a significant advantage that might alter the manner by which tumor–virus associations are studied in the future. The value of this capacity might be best illustrated in the rare cases of urothelial carcinoma with HPV. Our approach provides a discovery framework to identify tumors whose pathogenesis might be driven, at least in part, by virus-mediated/induced genomic perturbations.

Next-generation sequencing has paved the way to a greater understanding of virus-associated tumors, thanks to the study of the molecular complexity of multicentric lesions and intratumoral heterogeneity in whole tumor genomes. It already helps to better understand epidemiology in highlighting relevant associations between viruses and cancers. Detailed analysis of somatic mutation signatures and DNA virus genome integration sites will clarify the molecular basis of carcinogenesis. Integration of large genomic data sets with functional annotation will provide a new horizon for human cancer diagnosis, prognosis, and treatment in the near future. At an individual scale, identification of molecular oncogenic events is leaning toward a molecular-level personalized medicine.

References

1. Parkin DM. The global health burden of infection-associated cancers in the year 2002. Int J Cancer. 2006;118(12):3030–44.
2. Moore PS, Chang Y. Why do viruses cause cancer? Highlights of the first century of human tumour virology. Nat Rev Cancer. 2010;10(12):878–89.
3. Blumberg BS, Alter HJ, Visnich S. A “new” antigen in leukemia sera. JAMA. 1965;191:541–6.
4. Boshart M, et al. A new type of papillomavirus DNA, its presence in genital cancer biopsies and in cell lines derived from cervical cancer. EMBO J. 1984;3(5):1151–7.
5. Chang Y, et al. Identification of herpesvirus-like DNA sequences in AIDS-associated Kaposi’s sarcoma. Science. 1994;266(5192):1865–9.
6. Choo QL, et al. Isolation of a cDNA clone derived from a blood-borne non-A, non-B viral hepatitis genome. Science. 1989;244(4902):359–62.
7. Durst M, et al. A papillomavirus DNA from a cervical carcinoma and its prevalence in cancer biopsy samples from different geographic regions. Proc Natl Acad Sci U S A. 1983;80(12):3812–5.
8. Epstein MA, Achong BG, Barr YM. Virus particles in cultured lymphoblasts from Burkitt’s lymphoma. Lancet. 1964;1(7335):702–3.
9. Feng H, et al. Clonal integration of a polyomavirus in human Merkel cell carcinoma. Science. 2008;319(5866):1096–100.
10. Poiesz BJ, et al. Detection and isolation of type C retrovirus particles from fresh and cultured lymphocytes of a patient with cutaneous T-cell lymphoma. Proc Natl Acad Sci U S A. 1980;77(12):7415–9.
11. Ng SB, Khoury JD. Epstein-Barr virus in lymphoproliferative processes: an update for the diagnostic pathologist. Adv Anat Pathol. 2009;16(1):40–55.

12. Poreba E, Broniarczyk JK, Gozdzicka-Jozefiak A. Epigenetic mechanisms in virus-induced tumorigenesis. Clin Epigenetics. 2011;2(2):233–47.
13. Boss IW, Plaisance KB, Renne R. Role of virus-encoded microRNAs in herpesvirus biology. Trends Microbiol. 2009;17(12):544–53.
14. Williams R. Global challenges in liver disease. Hepatology. 2006;44(3):521–6.
15. Strong K, et al. Preventing cancer through tobacco and infection control: how many lives can we save in the next 10 years? Eur J Cancer Prev. 2008;17(2):153–61.
16. Ding J, Wang H. Multiple interactive factors in hepatocarcinogenesis. Cancer Lett. 2014;346(1):17–23.
17. Sung WK, et al. Genome-wide survey of recurrent HBV integration in hepatocellular carcinoma. Nat Genet. 2012;44(7):765–9.
18. Jiang Z, et al. The effects of hepatitis B virus integration into the genomes of hepatocellular carcinoma patients. Genome Res. 2012;22(4):593–601.
19. Fung J, Lai CL, Yuen MF. Hepatitis B and C virus-related carcinogenesis. Clin Microbiol Infect. 2009;15(11):964–70.
20. Deyrup AT. Epstein-Barr virus-associated epithelial and mesenchymal neoplasms. Hum Pathol. 2008;39(4):473–83.
21. Thompson MP, Kurzrock R. Epstein-Barr virus and cancer. Clin Cancer Res. 2004;10(3): 803–21.
22. Lee JH, et al. Clinicopathological and molecular characteristics of Epstein-Barr virus-associated gastric carcinoma: a meta-analysis. J Gastroenterol Hepatol. 2009;24(3):354–65.
23. Takada K. Epstein-Barr virus and gastric carcinoma. Mol Pathol. 2000;53(5):255–61.
24. Zhao J, et al. Genome-wide identification of Epstein-Barr virus-driven promoter methylation profiles of human genes in gastric cancer cells. Cancer. 2013;119(2):304–12.
25. Marquitz AR, et al. Infection of Epstein-Barr virus in a gastric carcinoma cell line induces anchorage independence and global changes in gene expression. Proc Natl Acad Sci U S A. 2012;109(24):9593–8.
26. Arora R, Chang Y, Moore PS. MCV and Merkel cell carcinoma: a molecular success story. Curr Opin Virol. 2012;2(4):489–98.
27. Isakov O, Modai S, Shomron N. Pathogen detection using short-RNA deep sequencing subtraction and assembly. Bioinformatics. 2011;27(15):2027–30.
28. Kostic AD, et al. Genomic analysis identifies association of Fusobacterium with colorectal carcinoma. Genome Res. 2012;22(2):292–8.
29. Chen Y, et al. VirusSeq: software to identify viruses and their integration sites using next-generation sequencing of human cancer tissue. Bioinformatics. 2013;29(2):266–7.
30. Moody CA, Laimins LA. Human papillomavirus oncoproteins: pathways to transformation. Nat Rev Cancer. 2010;10(8):550–60.
31. Alazawi W, et al. Changes in cervical keratinocyte gene expression associated with integration of human papillomavirus 16. Cancer Res. 2002;62(23):6959–65.
32. Pett MR, et al. Selection of cervical keratinocytes containing integrated HPV16 associates with episome loss and an endogenous antiviral response. Proc Natl Acad Sci U S A. 2006;103(10):3822–7.
33. Dall KL, et al. Characterization of naturally occurring HPV16 integration sites isolated from cervical keratinocytes under noncompetitive conditions. Cancer Res. 2008;68(20):8249–59.
34. Wentzensen N, Vinokurova S, von Knebel Doeberitz M. Systematic review of genomic integration sites of human papillomavirus genomes in epithelial dysplasia and invasive cancer of the female lower genital tract. Cancer Res. 2004;64(11):3878–84.
35. Ziegert C, et al. A comprehensive analysis of HPV integration loci in anogenital lesions combining transcript and genome-based amplification techniques. Oncogene. 2003;22(25): 3977–84.
36. Smith PP, et al. Viral integration and fragile sites in human papillomavirus-immortalized human keratinocyte cell lines. Genes Chromosomes Cancer. 1992;5(2):150–7.
37. Khoury JD, et al. Landscape of DNA virus associations across human malignant cancers: analysis of 3,775 cases using RNA-Seq. J Virol. 2013;87(16):8916–26.

38. Liang C, et al. Biomarkers of HPV in head and neck squamous cell carcinoma. Cancer Res. 2012;72(19):5004–13.
39. Rozenblatt-Rosen O, et al. Interpreting cancer genomes using systematic host network perturbations by tumour virus proteins. Nature. 2012;487(7408):491–5.
40. Tang KW, et al. The landscape of viral expression and host gene fusion and adaptation in human cancer. Nat Commun. 2013;4:2513.
41. Dandri M, Locarnini S. New insight in the pathobiology of hepatitis B virus infection. Gut. 2012;61 Suppl 1:i6–17.
42. Fujimoto A, et al. Whole-genome sequencing of liver cancers identifies etiological influences on mutation patterns and recurrent mutations in chromatin regulators. Nat Genet. 2012;44(7): 760–4.
43. Murakami Y, et al. Large scaled analysis of hepatitis B virus (HBV) DNA integration in HBV related hepatocellular carcinomas. Gut. 2005;54(8):1162–8.
44. Amaddeo G, et al. Integration of tumour and viral genomic characterisations in HBV-related hepatocellular carcinomas. Gut. 2014;0:1–10.
45. Dejean A, et al. Hepatitis B virus DNA integration in a sequence homologous to v-erb-A and steroid receptor genes in a hepatocellular carcinoma. Nature. 1986;322(6074):70–2.
46. Wang J, et al. Hepatitis B virus integration in a cyclin A gene in a hepatocellular carcinoma. Nature. 1990;343(6258):555–7.
47. Paterlini-Brechot P, et al. Hepatitis B virus-related insertional mutagenesis occurs frequently in human liver cancers and recurrently targets human telomerase gene. Oncogene. 2003;22(25): 3911–6.
48. Ferber MJ, et al. Integrations of the hepatitis B virus (HBV) and human papillomavirus (HPV) into the human telomerase reverse transcriptase (hTERT) gene in liver and cervical cancers. Oncogene. 2003;22(24):3813–20.
49. Zhang H, et al. Genome-wide association study identifies 1p36.22 as a new susceptibility locus for hepatocellular carcinoma in chronic hepatitis B virus carriers. Nat Genet. 2010;42(9): 755–8.
50. Jiang DK, et al. Genetic variants in STAT4 and HLA-DQ genes confer risk of hepatitis B virus-related hepatocellular carcinoma. Nat Genet. 2013;45(1):72–5.
51. Wang Y, et al. Hedgehog signaling pathway regulates autophagy in human hepatocellular carcinoma cells. Hepatology. 2013;58(3):995–1010.
52. Kim Y, et al. Selective down-regulation of glioma-associated oncogene 2 inhibits the proliferation of hepatocellular carcinoma cells. Cancer Res. 2007;67(8):3583–93.
53. Zur Hausen H. The search for infectious causes of human cancers: where and why. Virology. 2009;392(1):1–10.
54. Kostic AD, et al. PathSeq: software to identify or discover microbes by deep sequencing of human tissue. Nat Biotechnol. 2011;29(5):393–6.

Using Next-Generation Sequencing to Reveal Patterns of Chromosomal Alterations in Oral Verrucous Lesions

Manar Samman and Neeraj Sethi

Abstract Oral verrucous carcinoma is considered to be a histological subtype of oral squamous cell carcinoma, which generally follows a more indolent clinical course. It presents a specific clinical challenge in that it can be difficult to reach a definitive diagnosis and harbours the potential to transform into the more aggressive squamous cell carcinoma. In this chapter, we will discuss the background of this disease and the potential and previous role of next-generation sequencing (NGS) in head and neck cancer. The use of low coverage NGS to produce copy number variation (CNV) data and demonstrate how different computational methods can be applied to that data to analyse patterns and identify targets of interest. The application of NGS to detect and determine the prevalence of human papillomavirus in this disease will also be discussed.

1 Introduction

Head and neck squamous cell carcinoma (HNSCC) is the sixth most common cancer in the world, with an occurrence incidence of ~600,000 cases in each year and a 5-year survival rate of only ~50 % [1]. They are a heterogeneous group of tumours occurring anywhere from the lips to the trachea. These are biologically similar tumours in that the vast majority (>90 %) are squamous cell carcinoma, but clinically different in their presentation and complications of the disease. Oral cavity squamous cell carcinoma (OSCC) is found mainly in older men who are exposed to

M. Samman, M.Sc. (✉)
Leeds Institute of Cancer and Pathology, St James' University Hospital, Wellcome Trust Brenner Building, Leeds LS9 7TF, UK

King Fahad Medical City, Riyadh, Saudi Arabia
e-mail: m.samman@leeds.ac.uk

N. Sethi, M.B.Ch.B., M.R.C.S., D.O.H.N.S.
Leeds Institute of Cancer and Pathology, St James' University Hospital, Wellcome Trust Brenner Building, Leeds LS9 7TF, UK
e-mail: neerajsethi@doctors.org.uk

W. Wu, H. Choudhry (eds.), *Next Generation Sequencing in Cancer Research, Volume 2*, DOI 10.1007/978-3-319-15811-2_19

known risk factors such as tobacco and alcohol [2]. OSCC remains a challenging disease to treat, with disease-free survival rates of 58 % [3]. Oral verrucous carcinoma is considered a histological subtype of oral carcinoma with a relatively indolent clinical course [4–6].

Similarly to Fearon and Vogelstein's description of colorectal carcinogenesis, oral cancer has been demonstrated to occur with the stepwise accumulation of genetic abnormalities [7]. Alterations in genes regulating DNA synthesis and repair, cell cycle progression, and cell division are fundamental to this process [8]. The consequences of these genetic and molecular alterations are changes in the epithelial tissue phenotype, which may be histologically recognised as epithelial dysplasia, and eventually representing cell proliferation and differentiation dysregulation [9, 10]. However, precise, predictive assessments of individual oral cancers and precancers clinical behaviour and progression still remain indefinable in clinical practice [8]. Some oral squamous cell carcinomas arise in obvious normal mucosa, but many are preceded by clinically apparent premalignant lesions, mainly leukoplakia (white patch), erythroplakia (red patch), or speckled leukoplakia (white and red patches) [5].

Oral leukoplakia clinical phenotypes can range from thin homogeneous well-defined bordered white plaques to thick verrucous lesions [4]. In 1980, the term verrucous hyperplasia of the oral mucosa was coined by Shear and Pindborg [11]. Oral verrucous hyperplasia (OVH) is a whitish or pink mass or an oral mucosal plaque with a papillary or verrucous surface [12]. Shear and Pindborg, they described that 29 % of OVH lesions also showed histological features of OVC. Very few studies have been published on OVH, and the malignant transformation potential of verrucous hyperplasia lesions has not been inspected in detail [11]. In 2007, a follow-up study based in a Taiwanese hospital indicated that ten out of 324 patients with OVH developed oral cancers (two progressed to OVC and eight progressed to OSCC) in an average time of 54.6 months [13]. Similarly, a retrospective study that was conducted in 2009 in Taiwan hospital clinics reported that the annual malignant transformation rate of oral verrucous hyperplasia to OSCC is around 5.2 per 100 patient-years [14]. OVH is therefore considered a histological precursor of oral verrucous carcinoma (OVC) [6] that may transform into either an OVC or an OSCC [12]. OVH lesions are more common in 4th to 5th decade male patients; they occur mostly on the buccal mucosa and the tongue and are usually highly associated with cigarette smoking, alcohol drinking, and the areca quid chewing habits [6, 12].

In 1948, Ackerman defined OVC; it is also known as Ackerman's tumour or verrucous carcinoma of Ackerman [15]. OVC has been described as a low grade, slow growing, non-metastasizing, rare variant of OSCC [16]. It constitutes 2–10 % of OSCC [17]. Generally speaking, OVC clinico-histopathological diagnosis is usually exclusionary and extremely difficult, though it has a better prognosis compared to other carcinomas [18]. Histologically, OVC consists of thickened, club-shaped papillae and blunt stromal invaginations of well-differentiated squamous epithelium with marked keratinization, with the squamous epithelium lacking cytological criteria of malignancy. OVC invades underlying stroma with a pushing, rather than infiltrating front [16].

The aetiology of OVC is not well known [18], though it has been suggested that OVC develops from premalignant lesion [11, 19]. Smoking appears to be highly associated with the development of OVC [20]. In Asia, bidis and cigarettes smoking is known to be associated with leukoplakia, and areca quid, paan, and miang chewing habits have also been found [21]. Since a verrucous appearance is suggestive of viral aetiology, a number of investigations to study the putative association between HPV and OVC have been undertaken [22, 23]. These have reported a wide range in the incidence of HPV in OVC (30–100 %) leading to its actual role in OVC pathogenesis being controversial and inconclusive. This variation can be attributed to the deficiency of standardised detection procedures and the difficulty in defining complete histological criteria for OVH and OVC cases. Furthermore, the rarity of these types of lesions makes it difficult to study and investigate, and most previous studies or case reports have been made on small number of cases.

Another challenge in the establishing the diagnosis in this disease is that OVH resembles OVC both histologically and clinically. Routine histological examination of haematoxylin and eosin (H&E) stained sections is currently the most reliable method to distinguish between these entities, which is based on determining the endophytic and invasive growth pattern of OVC, from the exophytic growth pattern associated with OVH [24]. In1980, Shear and Pindborg described the histopathological key point features of oral verrucous lesions [11]. However, the differentiation of these lesions is often difficult with poorly orientated specimens, small biopsies, and, particularly, with biopsies that fail to show the margin of the lesion [24]. Therefore a more discriminatory method of distinguishing OVC from OVH as well as OSCC is needed.

During the past half-decade, the development of next-generation sequencing (NGS) technologies has enabled high sensitivity and resolution studies of cancer genomes through whole-exome and whole-genome sequencing approaches [25]. In head and neck cancer, three studies have collectively performed whole-exome sequencing on 151 tumours from multiple subsites. These studies confirmed that in HPV-negative tumours, *TP53* is almost universally aberrant and discovered *NOTCH1* to be the second most commonly mutated gene in head and neck cancer [26–28]. In addition these studies revealed a relatively low level of overlap in recurrently mutated genes between tumours, though they did discover that 31 % of their cohort contained phosphoinositide 3-kinase (*PI3K*) pathway mutations. NGS has also been established to be an effective, sensitive method for testing for HPV with the advantage that it can test for all subtypes of HPV in a sample [29].

Furthermore, NGS techniques offer considerable benefits for copy number variation (CNV) analysis, including precise delineation of the genomic breakpoints and higher resolution (can detect single-base insertions or deletions) of copy number changes [30]. It enables the estimation of tumour-to-normal copy number ratio at a genomic locus through counting the number of reads at this locus in normal and tumour samples [30]. Nevertheless, sequencing data can be produced even with nanogram amounts of DNA extracted from formalin-fixed paraffin-embedded (FFPE) materials [31].

Since OVC reveals minimal cytological atypia, besides the nuclear hyperchromatism of focal basal cells, distinction of OVC from OVH cannot only rely on the cytological features [32, 33]. For this reason, new experimental approaches are needed to improve the pathological diagnostic criteria for OVH and OVC and for better understanding of the development and progression of those lesions. We will present the methodology and analysis of NGS copy number data to distinguish between the genomic damage pattern in OVH and OVC and to analyse a subset of oral verrucous lesions (including VC and VH cases) for the presence of HPV subtypes and all characterised human viral genomes.

2 Sample Selection and Characteristics of OVH and OVC

The rarity of OVC and OVH lesions is an obstacle to studying this disease. Therefore, only by collaborating with several international cancer centres were we able to obtain the largest cohort of OVC and OVH samples. These included the Leeds Teaching Hospitals Pathology Archive; the Pathology Division, University of Torino; the Department of Pathology and Laboratory Medicine, at National Guard Hospital, Saudi Arabia; and the Department of Histopathology, Queen Victoria Hospital, East Grinstead, UK. All pathological materials used for this from each case were available in the form of archival formalin-fixed paraffin-embedded (FFPE) tumour blocks. Samples from Turin, Italy, were taken as sections on glass slides (10 μm sections onto 10 plain glass slides from each block).

In total, 92 OVC and OVH samples were obtained, making this the largest cohort in the literature. The original diagnoses were confirmed by Dr. Alec High (reference head and neck pathologist). World Health Organisation (WHO) definition and criteria were used for the histological diagnosis of OVH and OVC [16]. Verrucous appearing, but clearly "invasive" squamous lesions were classified as verrucous SCC and excluded [16]. Because of different reasons such as low yields of the extracted DNA and failed library preparations, 73 cases out of 92 succeeded for NGS copy number analysis. From the 73 cases, a total of 16 OVH patients were identified, ranging from 52 to 80 years old, and a total of 57 OVC patients were identified, ranging from 46 to 96 years old.

3 Nucleic Acid Extraction

A head and neck pathologist, on a single 5 μm H & E slide created from each tumour FFPE block, identified the areas of the highest tumour cell purity. This area was then microdissected from seven consecutive 10 μm slides. The nucleic acid was extracted from this tissue using the Qiagen AllPrep kit. The DNA was sonically sheared to ensure no fragments larger than 200bo were present in each sample, and the NEBNext Library Prep Master Mix Set for Illumina was used to create sequencing libraries. These were then multiplexed at 40 per lane on the Illumina HiSeq 2000.

4 Generation of Digital Karyograms

High-resolution mapping of CNV involves sequenced reads aligned to a reference genome [34]. The distribution of the aligned reads is then analysed on a genomic segmental window-by-window basis to define alterations in read depth between the reference genomes and tests [35]. When compared to the control sample, a reduction in sample read depth across a window suggests a loss in genomic component; an increase in read depth represents a gain [36]. The threshold of ploidy at which a significant copy number alteration is to be "called" needs to be set, and in the following analyses, this was set at 0.05.

Here, DNA was sequenced at a coverage between 0.1 and 0.5, and the copy number (CN) was calculated and analysed. Briefly, sample reads were arranged and organised by chromosome and position. The ratio of test to control reads was calculated across the genome in equal sized windows averaging 200 reads in some cases and 400 reads in others. A control sample was pooled from a group of 20 British normal individuals downloaded from the 1000 Genomes Project[38] (ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp, last accessed 3rd March 2011). For all samples, sufficient sequencing reads were obtained, and hence, digital karyograms were constructed from these data using the CNAnorm programme in which CNV can be analysed [37].

Generally speaking, sequence variations are usually not distributed uniformly within genomes [38]. Nonetheless, CNVs that are enriched in simple tandem repeats occur more often towards centromeres and telomeres and are not elevated in guanine and cytosine content or SNPs [38]. For the purposes of this analysis, copy number alterations in all centromere and telomere chromosomal regions were excluded [39].

5 Analysing CNV Data

5.1 Visual Inspection

5.1.1 Comparison of Individual Karyograms

Visual inspection of the 73 patient (OVHs and OVCs) genomic copy number karyograms demonstrated regions of gain and loss along the whole genome in OVC cases. In general, OVC karyograms showed different types of copy number patterns, in terms of both, complexity of the damage, and the proportion of genomes involved. This pattern ranged from whole chromosome gain to amplified or lost chromosome arms and regions. As an illustration, Fig. 1 below demonstrates a CN karyogram for a histologically normal oral epithelium tissue with no chromosomal gain or loss.

An example of a typical digital karyogram generated from an OVH sample (Fig. 2) is shown below. The blue arrow points to gain in Chr7. Horizontal lines above the centre demonstrate regions of gain, and those below the centre demonstrate regions of loss. In general, little gain and loss were found in OVH cases. Visual

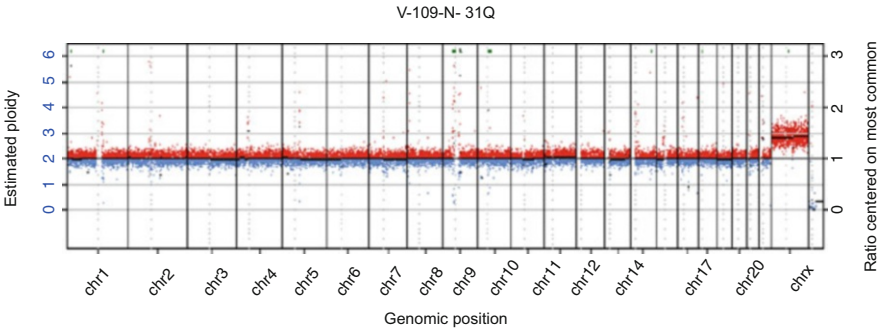


Fig. 1 Representative genomic profiling of histologically normal oral epithelium by NGS CN analysis. Each data point represents one window of approximately 200 reads. Genomic position is on the *x*-axis and tumour/normal ratio is on the *y*-axis. The black lines are regions of common copy number between breakpoints. Windows of gain and loss are *red* and *blue*, respectively

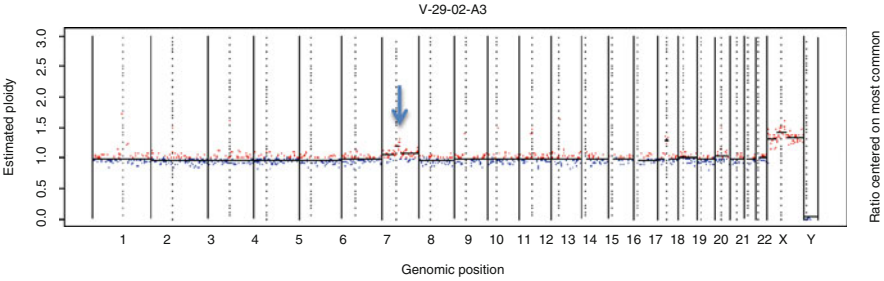


Fig. 2 Representative genomic profiling of OVH by NGS CN analysis. Each data point represents one window of approximately 200 reads. Genomic position is on the *x*-axis and tumour/normal ratio is on the *y*-axis. The *black lines* are regions of common copy number between breakpoints. Windows of gain and loss are *red* and *blue*, respectively. *Blue arrow* points to gain in chromosome 7

examination of the 16 individual patient genomic copy number traces revealed a very low level of genomic damage in OVH samples compared to oral verrucous tumours (N: 57), indicating that the genomic profile of these cases has minimal chromosomal abnormalities and is more similar to normal.

These findings are surprising since it has been well known that OVH shares similar clinical and histological morphology to OVC, and the clinical differentiation of the verrucous hyperplastic lesions from OVC is often difficult [6, 11, 40]. From what has been found here and despite the similar clinical and histological features that OVH and OVC share, the analysis of OVH individual CN karyograms showed that these lesions have different genomic profile from OVC with very low, narrow levels of DNA aneuploidy.

Figure 3 shows an example of the karyogram profile generated from an OVC sample. Blue arrows point at gains in Chr2, Chr7, Chr10, Chr16, and Chr17. Horizontal lines above the centre demonstrate regions of gain, and those below the centre demonstrate regions of loss. OVC karyograms appear in an early stage of DNA near-diploid aneuploidy. In addition, gains at 7q, 16q, and 17q (represented by red

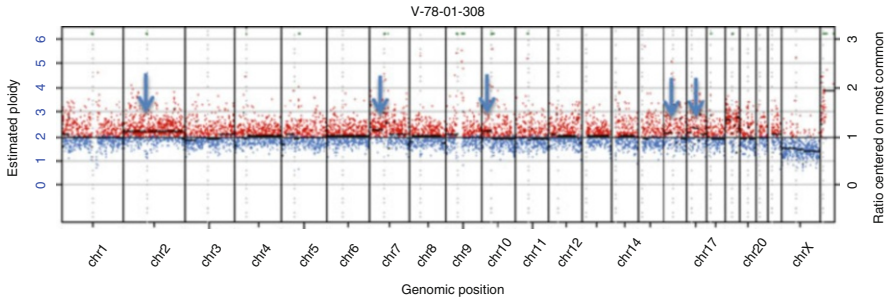


Fig. 3 Representative genomic profiling of OVC by NGS CN analysis. Each data point represents one window of approximately 200 reads. Genomic position is on the *x*-axis and tumour/normal ratio is on the *y*-axis. The *black lines* are regions of common copy number between breakpoints. Windows of gain and loss are *red* and *blue*, respectively. *Blue arrows* point to regions of chromosomal gain

with black lines) were detected frequently in OVC cohort, suggesting that these CN alterations may be involved in the development of OVC. On the other hand, deletion trends were minimally found in OVCs, suggesting that overexpression of oncogenes is most likely to be involved in the development of OVC.

5.1.2 Comparison of Cumulative Frequency Karyograms

Frequency karyograms were produced using an in-house built programme that takes all BED files (.bed) from the CN analysed sample lists. The selected CN threshold was of 0.05 above or below and was considered a gain or loss. In general, visual examination of OVH (N: 16) genomic CN frequency karyogram (Fig. 4a) noticeably illustrates the very low level of CN alterations in OVHs compared to OVCs, indicating that the genomic profile of these cases has minimal chromosomal abnormalities and is most similar to normal. In addition to a genome-wide view, the individual chromosome cumulative frequency plots can be viewed to enable a higher resolution examination. Visual examination was performed on chromosome plots in order to investigate the genomic locations of chromosomal segments with altered CN in OVHs. Gains mapped at chromosome 7q11.2 and 7q22 (represented by red colour) were noticed in OVHs at a frequency of ~50 %, suggesting that this CN alteration may be related to development of OVH. These results are different from a study in 2001, which reported high frequency of allelic loss in 20/25 OVH cases at loci on 3p, 9p, 4q, 8p, 11q, 13q, and 17p chromosome arms (one or more than one arm) and then suggested that loss of heterozygosity (LOH) on these arms may explain the malignant potential of OVH lesions [40]. Allelic loss without CN loss is possible; however, it is unlikely not to identify any in OVH cohort here at all these loci. Nonetheless, it is important to keep in mind the inability of LOH techniques to identify chromosomal gains, which differ from array-based comparative genomic hybridization (aCGH) or NGS CN analysis capabilities in detecting both chromosomal losses and gains [41]. Consequently, these CN approaches are preferable to LOH if gene amplifications and chromosomal gains are to be assessed alongside losses.

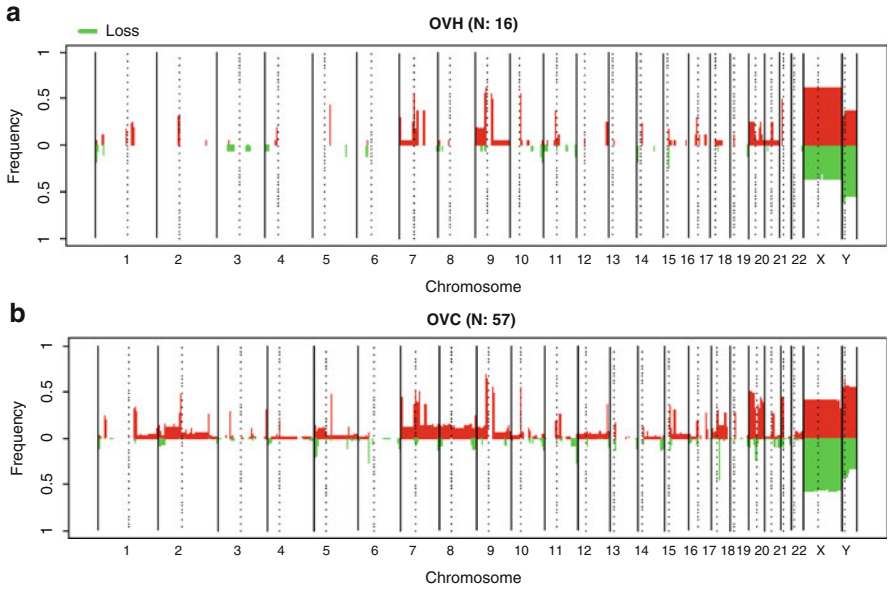


Fig. 4 Frequency of genomic gain and loss for OVH (**a**) and OVC (**b**). Genomic position is on the *x*-axis, frequency (%) of gains (*red*), and losses (*green*) are shown on the *y*-axis

Visual examination of OVC (N: 57) genomic CN frequency karyogram (Fig. 4b) revealed a higher level of CNA compared to OVH. In OVCs, there is no loss at chromosome 3p and gain at 3q and a lower frequency of gain of 5p and 8q. Furthermore, gains mapped at chromosome 7p22, 7q11.2, and 7q22 (represented by red colour) were observed in OVCs at a frequency of ~50 %, in addition to gains mapped at chromosomes 3p21 (at a frequency of ~30 %), 15q15 (at a frequency of ~30 %), 16q22 (at a frequency of ~25 %), and 17q23 (at a frequency of ~25 %) as well as losses on chromosomes 6p21 (at a frequency of ~25 %) and 17q12 (at a frequency of ~50 %) represented by green colour, suggesting that these CN alterations may be involved in the development of OVC.

Gains at chromosome arms 7q, 16q, and 17q were detected in OVCs at a frequency of 50 % and have not been previously identified as a common CN altered chromosome lesions in oral cancer. Nevertheless, deletion trends were minimally found in OVC's frequency karyogram, suggesting that overexpression of oncogenes is most likely to be involved in the development of OVC. However, in 2001, a study was conducted to investigate the frequency of allelic loss in oral verrucous lesions, including 17 OVC samples [40]. They reported high frequency of allelic loss at loci on 3p, 9p, 4q, 8p, 11q, 13q, and 17p chromosome arms (one or more than one arm) then suggested that LOH on these arms may explain the malignant potential of OVCs [40]. From their findings, two chromosomal regions were comparable here to CN aberrations outcome in OVC cohort (losses in chromosomes 8p23.3 and 9p21). Though, loss of chromosome 8p23.3 (which is a telomere region as well) was at a frequency of ~10 %, and loss of chromosome 9p21 was at a frequency of ~5 % in this

OVC study cohort. In addition, the previous study included 17 OVC samples, while here, 57 OVC samples were included. Again, it is important to keep in mind the inability of LOH techniques to identify chromosomal gains, which differ from aCGH or NGS CN analysis capabilities in detecting chromosomal losses and gains [41].

5.2 Computational Approaches

5.2.1 Genomic Identification of Significant Targets in Cancer (GISTIC) 2.0

The GISTIC algorithm identifies likely somatic driver CN alterations through evaluating the amplitude and frequency of amplified or deleted observed events [42]. GISTIC has been used and applied to many cancer types, including lung and esophageal squamous carcinoma [43], colorectal carcinoma [44], melanoma [45], and ovarian carcinoma [46] and has facilitated the identification of several new amplification targets (including *SOX2* [43], *CDK8* [44], *NKX2-1* [47], and *VEGFA* [48], besides deletions (*EHMT1* [49]).

Here, the CN profiles of OVH and OVC were characterised by several approaches using the GISTIC2.0 algorithm including amplification and deletion plots of CNAs, the identification of amplification and deletion genes within CN altered regions, and segmented CN heat maps. The default parameters were used in the GISTIC analysis. A number of regions of recurrent CN gains and losses were evident in the GISTIC analysis in OVH and OVC cohorts and matched the generated frequency karyograms CN aberrations described above. Genomic positions of the most significant amplification and deletion peaks (from the GISTIC analysis) including the list of genes contained in them for OVH and OVC samples were identified in Tables 1 and 2. The results were then further analysed by running the gene

Table 1 Lists of genes located in the most common regions of recurrent DNA copy number change in OVH

Focal events	Genomic position	Genes mapping within region
Focal deletion (*q*-value: 0.032774)	5q31.1, wide peak boundaries: chr5:133200002-137600000	CAMLG, IL9, LECT2, SMAD5, NEUROG1, NPY6R, PITX1, PPP2CA, SKP1, SPOCK1, TCF7, TGFBI, UBE2B, VDAC1, WNT8A, NME5, CDC23, MYOT, CXCL14, H2AFY, SMAD5-AS1, DDX46, KIF20A, SEC24A, BRD8, HNRNPA0, PHF15, FBXL21, KLHL3, PKD2L2, SAR1B, CDKL3, FAM13B, C5orf15, TRPC7, TXNDC15, PCBD2, CDKN2AIPNL, C5orf24, C5orf20, SLC25A48, LOC340073, LOC340074, CATSPER3, LOC389332, TIFAB, VTRNA2-1, MIR874, MIR3661, MIR4461
Focal deletion (*q*-value: 0.015741)	17q11.2, wide peak boundaries: chr17: 30000002-33600000	hsa-mir-632, ACCN1, LIG3, MYO1D, PSMD11, RAD51D, CCL1, CCL2, CCL7, CCL8, CCL11, CCL13, SH3GL1P1, ZNF207, CDK5R1, CCT6B, SUZ12, TMEM98, NLE1, FNDC8, RHOT1, C17orf79, UTP6, C17orf75, ZNF830, LRRC37B, RFFL, TMEM132E, SPACA3, SLC35G3, UNC45B, SLFN5, RHBDL3, C17orf102, ARGFXP2, MIR632, AA06, RAD51L3-RFFL

Table 2 Lists of genes located in the most common regions of recurrent DNA copy number change in OVC

Focal events	Genomic position	Genes mapping within region
Focal amplification (*q*-value: 4.68E-05)	3p21.31, wide peak boundaries: chr3:46400002-50000000	AMT, APEH, RHOA, SLC25A20, CAMP, CDC25A, CCR5, COL7A1, DAG1, CELSR3, GPX1, IMPDH2, LAMB2, LTF, MAP4, MST1, MST1R, MYL3, PFKFB4, PLXNB1, PRKAR2A, PTH1R, QARS, SMARCC1, TCTA, TDGF1, UBA7, USP4, UQCRC1, BSN, CCRL2, IP6K1, RBM6, NME6, TRAIP, ARIH2, CSPG5, USP19, WDR6, TREX1, DHX30, SCAP, LAMB2P1, NBEAL2, KLHL18, NDUFAF3, PTPN23, SETD2, PRSS50, GMPPB, SHISA5, CCDC72, ZNF589, IP6K2, NCKIPSD, P4HTM, C3orf75, QRICH1, DALRD3, RNF123, KIF9, CCDC71, SLC26A6, CAMKV, LRRC2, CCDC51, RTP3, ATRIP, NICN1, MON1A, UCN2, CCDC12, KLHDC8B, ALS2CL, TMIE, FBXW12, FLJ39534, CCDC36, PRSS42, C3orf62, PRSS45, AMIGO3, CDHR4, FAM212A, MIR191, TMEM89, MIR425, SPINK8, C3orf71, LOC646498, CCR2, NRADDP, LOC100132146, BSN-AS2, PRSS46, MIR1226, MIR711, MIR4271, MIR4793, MIR4443
Focal amplification (*q*-value: 1.47E-08)	7q11.23, wide peak boundaries: chr7:72800002-76400000	CLDN4, CLDN3, ELN, GTF2I, GTF2IP1, HIP1, HSPB1, LIMK1, MDH2, PMS2L2, PMS2P5, PMS2P3, POR, RFC2, CCL24, STX1A, EIF4H, CLIP2, LAT2, YWHAG, ZP3, FZD9, BAZ1B, BCL7B, GTF2IRD1, CCL26, POMZP3, TBL2, MLXIPL, STYXL1, STAG3L1, RHBDD2, UPK3B, WBSCR16, ABHD11, TMEM120A, GTF2IRD2, DNAJC30, DTX2, WBSCR22, WBSCR28, SRCRB4D, WBSCR27, VPS37D, NSUN5P1, ABHD11-AS1, SRRM3, TRIM73, TRIM74, SPDYE8P, GATSL1, GTF2IRD2B, STAG3L2, SPDYE5, FDPSL2A, NCF1, NCF1C, SNORA14A, MIR590, GATSL2, LOC100093631, POM121C, LOC100133091, MIR4284, MIR4651
Focal amplification (*q*-value: 4.93E-11)	7p22.2, wide peak boundaries: chr7:3600002-7200000	ACTB, PMS2, RAC1, FSCN1, ZNF12, AIMP2, CYTH3, KIAA0415, KDELR2, WIPI2, EIF2AK1, CCZ1, RNF216, ZNF853, ZDHHC4, RADIL, PAPOLB, RBAK, C7orf26, FBXL18, USP42, TNRC18, C7orf70, SDK1, FOXK1, MMD2, DAGLB, CCZ1B, SLC29A4, RSPH10B, LOC389458, GRID2IP, ZNF815, RNF216P1, PMS2CL, ZNF890P, OCM, MIR58, RSPH10B2, LOC100131257, RBAK-LOC389458, MIR4656
Focal amplification (*q*-value: 0.0015106)	7q21.3, wide peak boundaries: chr7:96800002-102800000	ACHE, ASNS, AZGP1, AP1S1, CUX1, CYP3A7, CYP3A4, CYP3A5, EPHB4, EPO, GNB2, AGFG2, LRCH4, MCM7, NPTX2, OCM2, SERPINE1, PCOLCE, PMS2P1, POLR2J, TAC1, TAF6, TFR2, TRIP6, VGF, ZAN, ZNF3, ZKSCAN1, ZSCAN21, TRRAP, BUD31, PLOD3, AP4M1, ATP5J2, MUC12, ARPC1B, RASA4, LRRC17, POP7, ZNHIT1, ARPC1A, SH2B2, STAG3, CPSF4, COPS6, PDAP1, LMTK2, ZKSCAN5, CLDN15, BRI3, TECPR1, PTCD1, FBXO24, PILRB, PILRA, FIS1, ACTL6B, SRRT, ALKBH4, ZCWPW1, C7orf43, BAIAP2L1, MEPCE, SLC12A9, ACN9, SMURF1, MOSPD3, GIGYF1, RABL5, CYP3A43, ZNF655, PVRIG, GAL3ST4, PRKRIP1, ORAI2, OR2AE1, TSC22D4, TRIM56, ARMC10, ZNF394, MYH16, TRIM4, MYL10, EMID2, MUC17, BHLHA15, ZNF498, FAM200A, PPP1R35, GPC2, LRWD1, FAM185A, FBXL13, NAPEPLD, TMEM130, NYAP1, CNPY4, POLR2J2, MBLAC1, ZNF789, MOGAT3, GJC3, GATS, NAT16, MGC72080, C7orf59, KPNA7, C7orf61, UFSP1, MIR106B, MIR25, MIR93, SPDYE3, SPDYE2, POLR2J3, AZGP1P1, SPDYE6, RPL19P12, LOC100129845, UPK3BL, LOC100289187, LOC100289561, SPDYE2L, SAP25, MIR4285, MIR3609, ATP5J2-PTCD1, MIR4653, MIR4467, MIR4658, LOC100630923, CYP3A7-CYP3AP1

Focal amplification (*q*-value: 0.0012753)	15q15.1, wide peak boundaries: chr15:40000002-45600000	B2M, BUB1B, CAPN3, CKMT1B, EPB42, GANC, GCHFR, PDIA3, ITPKA, IVD, LTK, MAP1A, MFAP1, PLCB2, RAD51, SORD, SPINT1, SRP14, TP53BP1, TYRO3, EIF3J, JMJD7-PLA2G4B, SNAP23, SLC28A2, TGM5, PPIP5K1, LCMT2, SERF2, GPR176, CHP, OIP5, BAHD1, MAPKBP1, RTF1, MGA, VPS39, CCNDBP1, C15orf63, TMEM87A, RPAP1, RPUSD2, TUBGCP4, EHD4, DUOX2, NDUFAF1, NUSAP1, SPTBN5, CTDSPL2, DUOX1, DLL4, INO80, PPP1R14D, HAUS2, FAM82A2, DNAJC17, PAK6, CASC5, STARD9, VPS18, ZFP106, CHAC1, WDR76, TMEM62, SPG11, ELL3, ZFYVE19, FRMD5, DISP2, C15orf57, C15orf23, BMF, SHF, DUOXA1, CHST14, CASC4, TGM7, CATSPER2, PLA2G4E, TRIM69, C15orf43, ZSCAN29, TTBK2, CDAN1, STRC, ADAL, EXD1, FSIP1, RHOV, UBR1, PATL2, PLA2G4F, LRRC57, PLA2G4D, MRPL42P5, C15orf52, DUOXA2, EIF2AK4, CATSPER2P1, CKMT1A, SERINC4, C15orf62, C15orf56, PHGR1, LOC645212, MIR626, MIR627, LOC728758, OIP5-AS1, LOC100131089, ANKRD63, JMJD7, PLA2G4B, MIR1282, MIR4310, LOC100505648, SERF2-C15ORF63
Focal amplification (*q*-value: 0.0010611)	16q22.1, wide peak boundaries: chr16:66000002-70400000	AARS, AGRP, CA7, CBFB, CDH1, CDH3, CDH5, CDH16, CTRL, NQO1, DYNC1LI2, E2F4, HAS3, HSD11B2, HSF4, LCAT, NFATC3, PSKH1, PSMB10, RRAD, SLC9A5, SLC12A4, SNTB2, TERF2, TK2, TRADD, CES2, NAE1, NOL3, SLC7A6, ATP6V0D1, NUTF2, CTCF, NFAT5, WWP2, DDX19B, CES3, EDC4, PLA2G15, PLEKHG4, LRRC29, VPS4A, NOB1, TMEM208, FHOD1, ZDHHC1, PARD6A, CKLF, NIP7, FAM96B, TPPP3, PRMT7, DUS2L, CHTF8, PDPR, LRRC36, DDX19A, FBXL8, SMPD3, DDX28, TSNAXIP1, THAP11, PDP2, RANBP10, PDF, DPEP2, DPEP3, ACD, FAM65A, TMCO7, ELMO3, ESRP2, CENPT, C16orf70, CYB5B, GFOD2, C16orf48, SLC7A6OS, COG8, B3GNT9, CIRH1A, CMTM1, EXOSC6, NRN1L, CMTM3, ZFP90, RLTPR, KCTD19, CMTM4, CMTM2, BEAN1, TMED6, CCDC79, CES4A, EXOC3L1, PDXDC2P, CLEC18C, CLEC18A, C16orf86, MIR140, MIR328, KIAA0895L, LOC729513, MIR1538, MIR1972-1, MIR1972-2, LOC100505865, LOC100506083, CKLF-CMTM1
Focal amplification (*q*-value: 0.00024633)	17q22, wide peak boundaries: chr17:56000002-61200000	CA4, CLTC, LPO, MPO, TRIM37, SEP4, RAD51C, RPS6KB1, SRSF1, SUPT4H1, TBX2, VEZF1, EPX, PPM1D, MTMR4, BZRAP1, TBX4, MRC2, MED13, APPBP2, TLK2, PPM1E, TANC2, OR4D1, RNFT1, TUBD1, PTRH2, BCAS3, RNF43, MKS1, SMG8, MSX2P1, PRR11, TEX14, INTS2, HEATR6, DHX40, VMP1, BRIP1, USP32, HSF5, OR4D2, C17orf64, DYNLL2, EFCAB3, MARCH10, C17orf47, GDPD1, METTL2A, NACA2, SKA2, YPEL2, C17orf82, MIR142, MIR21, MIR301A, TBC1D3P2, LOC645638, TBC1D3P1-DHX40P1, LOC653653, SCARNA20, MIR454, MIR548W, LOC100506779, MIR4729, MIR4737, MIR4736
Focal deletion (*q*-value: 3.44E-05)	5q31.1, wide peak boundaries: chr5:133200002-137600000	CAMLG, IL9, LECT2, SMAD5, NEUROG1, NPY6R, PITX1, PPP2CA, SKP1, SPOCK1, TCF7, TGFBI, UBE2B, VDAC1, WNT8A, NME5, CDC23, MYOT, CXCL14, H2AFY, SMAD5-AS1, DDX46, KIF20A, SEC24A, BRD8, HNRNPA0, PHF15, FBXL21, KLHL3, PKD2L2, SAR1B, CDKL3, FAM13B, C5orf15, TRPC7, TXNDC15, PCBD2, CDKN2AIPNL, C5orf24, C5orf20, SLC25A48, LOC340073, LOC340074, CATSPER3, LOC389332, TIFAB, VTRNA2-1, MIR874, MIR3661, MIR4461
Focal deletion (*q*-value: 5.70E-14)	6p21.2, wide peak boundaries: chr6:38400002-41200000	DNAH8, GLO1, GLP1R, MOCS1, NFYA, KCNK5, APOBEC2, DAAM2, TREM2, SAYSD1, LRFN2, TREML2, KCNK16, KCNK17, LOC221442, C6orf130, KIF6, TSPO2, UNC5CL, TREML1, TREML3, FLJ41649, TDRG1, LOC100131047
Focal deletion (*q*-value: 0.015741)	17q11.2, wide peak boundaries: chr17:30000002-33600000	hsa-mir-632, ACCN1, LIG3, MYO1D, PSMD11, RAD51D, CCL1, CCL2, CCL7, CCL8, CCL11, CCL13, SH3GL1P1, ZNF207, CDK5R1, CCT6B, SUZ12, TMEM98, NLE1, FNDC8, RHOT1, C17orf79, UTP6, C17orf75, ZNF830, LRRC37B, RFFL, TMEM132E, SPACA3, SLC35G3, UNC45B, SLFN5, RHBDL3, C17orf102, ARGFXP2, MIR632, AA06, RAD51L3-RFFL
Focal deletion (*q*-value: 6.26E-06)	17q21.33, wide peak boundaries: chr17:48400002-55200000	CHAD, COX11, HLF, NME1, NME2, TRIM25, COIL, AKAP1, DGKE, ABCC3, CACNA1G, SPAG9, NOG, TOM1L1, TOB1, MMD, UTP18, MRPL27, LUC7L3, MBTD1, LINC00483, EPN3, TMEM100, RSAD1, LRRC59, CA10, PCTP, SCPEP1, XYLT2, SPATA20, ACSF2, MYCBPAP, KIF2B, ANKRD40, WFIKKN2, EME1, ANKFN1, MTVR2, STXBP4, LOC253962, C17orf67, RNF126P1, LOC400604, NME1-NME2, MIR3614

lists against 13 enriched Kyoto encyclopaedia of gene and genomes (KEGG) pathways, which are more related to head and neck cancers, as well as cancer gene census and Stransky mutation list (76 previously identified genes in HNSCCs harbouring high statistically significant mutations) [27].

5.2.2 Applying GISTIC to OVH CNV Data

Regions of significant gains or losses were identified using GISTIC algorithm. Two chromosomal regions (deletions) from the CNAs identified by GISTIC analysis were significantly altered in OVH patients' genomes according to this analysis (orange highlighted chromosomal positions in Fig. 5b). These two deletion regions that surpass the significance threshold are in chr 5q31.1 and 17q12 with a frequency less than 20 %. Surprisingly, no significant amplification regions were detected by

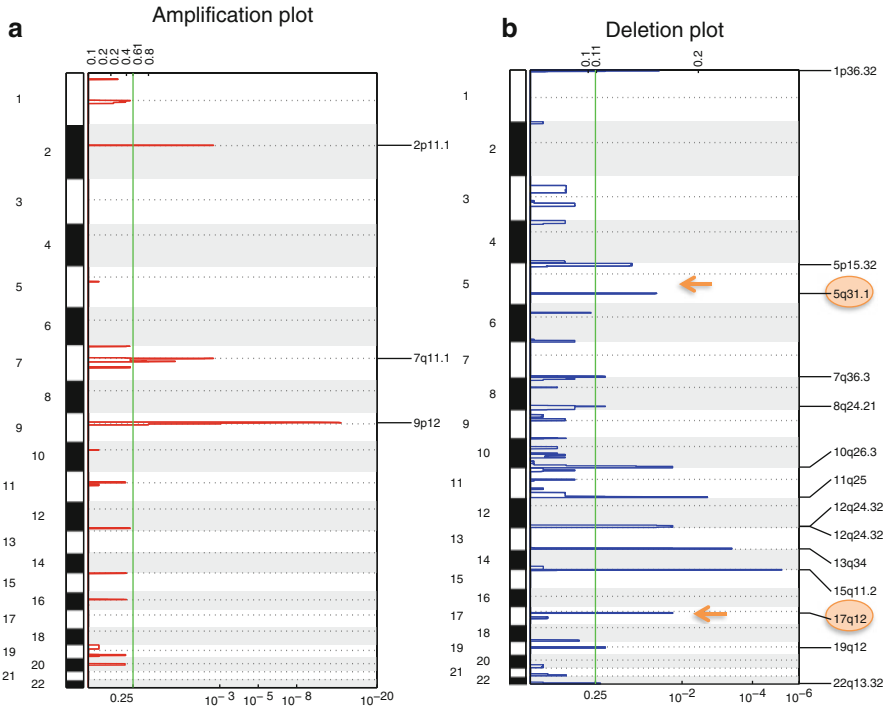


Fig. 5 Genome-wide amplification and deletion plots of CNAs in OVH. Genomic positions are indicated along the *y*-axis with centromere locations showed by *dotted lines*. Amplification (*red*) and deletion (*blue*) GISTIC plots show *q*-values (*bottom* on the *x*-axis), the G-scores that consider the frequency of the aberration occurrence as well as its amplitude across samples (*top*), and the significance threshold is indicated by the *green line* at 0.25, with respect to amplifications and deletions for all markers over the entire analysed region. *Blue arrows* and *circles* point to regions with significant gain, and *orange arrows* and *circles* point to regions with significant loss

GISTIC analysis, and the chromosomal regions shown in the amplification plot below are centromeres (e.g. Chr 7q11.1). In general, visual examination of OVH genomic CN plots noticeably illustrates the very low level of CN alterations in OVHs compared to OVC genomic CN plots.

5.2.3 Applying GISTIC to OVC CNV Data

Again, regions of significant gains or losses were identified using GISTIC algorithm. Ten chromosomal regions (seven amplifications and three deletions) from the CNAs identified by GISTIC analysis were significantly altered in OVC genomes (Fig. 6a, b). The seven most significant amplifications from GISTIC peaks (Fig. 6a, blue circles) that also surpass the significance threshold include chromosomes 3p21.31, 7p22.2, 7q11.23, 7q22.1, 15q15.2, 16q22.1, and 17q23.2. Gains mapped

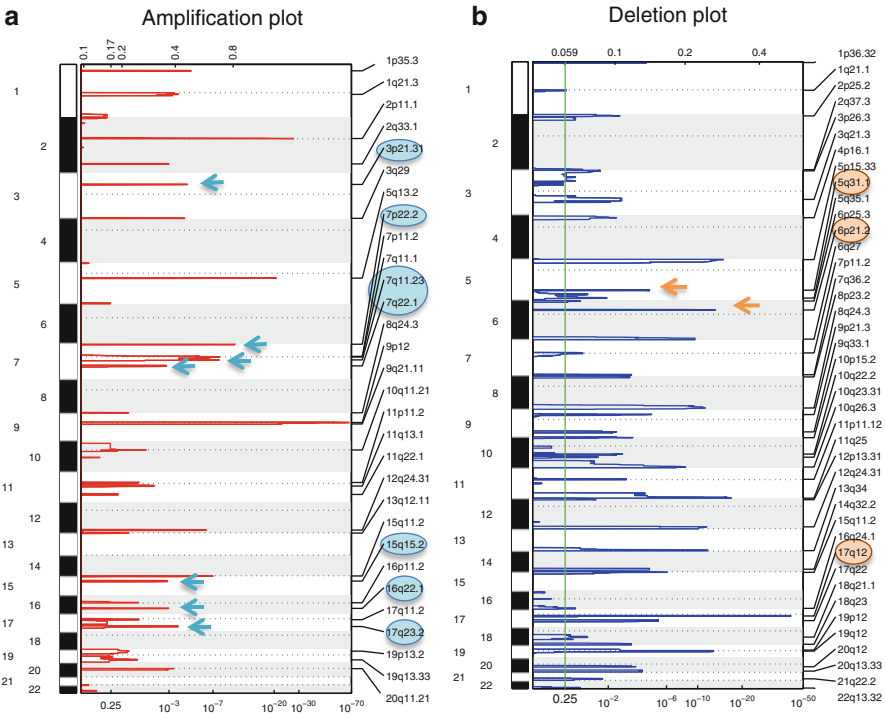


Fig. 6 Genome-wide amplification and deletion plots of CNAs in OVC. Genomic positions are indicated along the *y*-axis with centromere locations showed by *dotted lines*. Amplification (*red*) and deletion (*blue*) GISTIC plots show *q*-values (*bottom* on the *x*-axis), the G-scores that consider the frequency the aberration occurrence as well as its amplitude across samples (*top*), and the significance threshold is indicated by the *green line* at 0.25, with respect to amplifications and deletions for all markers over the entire analysed region. *Blue arrows* and *circles* point to regions with significant gain, and *orange arrows* and *circles* point to regions with significant loss

at chromosome 3p21 (at a frequency of ~50 %), 7p22 (at a frequency of ~75 %), 7q11.2 (at a frequency of ~70 %), 7q22 (at a frequency of ~35 %), 15q15 (at a frequency of ~40 %), 16q22 (at a frequency of ~40 %), and 17q23 (at a frequency of ~45 %) were observed as well in OVC frequency karyograms although with different frequencies (described above). The variation in the frequencies between GISTIC analyses CN plots and the frequency karyograms generated from OVC individual CN karyograms can be attributed to the non-specific visual examination method and human eye errors, as there was no algorithm to give the exact frequency percentage at the time of my analysis.

The three most significant deletions from GISTIC peaks (Fig. 6b, orange circles) that also surpass the significance threshold include chromosomes 5q31.1 (at a frequency of ~15 %), 6p21.2 (at a frequency of ~25 %), and 17q12 (at a frequency of ~15 %). Losses on chromosomes 6p21 and 17q12 were also observed in OVC frequency karyograms but with different frequencies (refer to Sect. 5.1.2). Again, the variation in the frequencies between GISTIC analyses CN plots and the frequency karyograms generated from OVC individual CN karyograms can be attributed to the inaccurate visual estimation and human eye errors, as there was no algorithm to give the exact frequency percentage at the time of this analysis.

A number of regions of recurrent CN gain and loss were evident in the GISTIC analysis (Figs. 5 and 6). Genomic positions of amplification and deletion peaks (identified in the GISTIC analysis) are listed below (Tables 1 and 2) in order to explain the next step, including the list of genes contained in them. Focal event regions were selected from the highlighted deletion circles in Fig. 5 and the highlighted amplifications and deletions circled in Fig. 6. The threshold for q-values is 0.25; regions with q-values lower than this number were considered significant, and genes within those regions were further investigated.

5.2.4 Assessment of the List of Genes Identified via GISTIC

The GISTIC method was used to identify the most significant amplifications and deletions as described previously. Two deletion peaks were identified in OVH, and these regions had a large gene lists. The results were then further analysed by running the gene lists against 13 enriched KEGG pathways (a large database project for metabolic pathways), including KEGG P13K, KEGG WNT signalling pathway, KEGG cell cycle, KEGG calcium signalling pathway, KEGG VEGF signalling pathway, KEGG MAPK pathway, KEGG DNA replication pathway, KEGG phosphatidylinositol signalling system, KEGG P53 signalling pathway, KEGG NOTCH signalling pathway, KEGG JAK-STAT signalling pathway, KEGG ERBB signalling, and KEGG hedgehog, which are more related to head and neck cancers, as well as cancer gene census and Stransky mutation list (76 previously identified genes in HNSCCs harbouring high statistically significant mutations) [27].

Out of eight key genes hits (refer to Fig. 7), four genes were involved in KEGG WNT signalling pathway (36 % of the CN altered genes in OVH cohort were involved in this pathway). In addition, two genes were involved in KEGG cell cycle

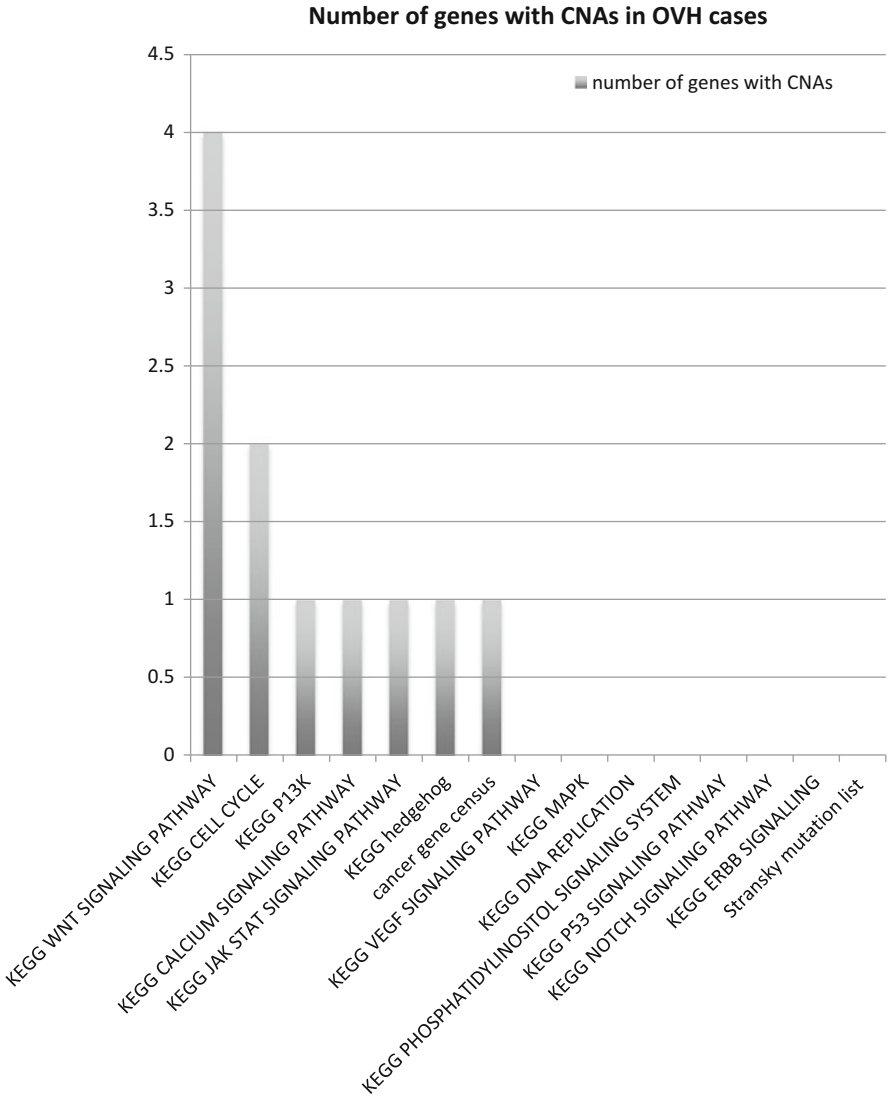


Fig. 7 A graphical representation of 13 enriched KEGG pathways, cancer gene census, and Stransky mutation list [27] on the *x*-axis ranked by the number of genes with CNAs from OVH samples in each pathway and list on the *y*-axis

pathway (18 % of the CN altered genes in OVH cohort were involved in this pathway), and one cancer gene was located as well among the eleven genes list (*SUZI2*). Table 3 lists all key genes founded to be in CN altered regions with the highest significant losses in OVH cohort. OVH illustrates very low level of CNAs, and as a result, a low number of genes were identified.

Table 3 Lists of all key genes founded to be CN altered within regions with the highest significant gains and losses in OVH, OVC, and OSCC cohorts

	Genes	
	OVH	OVC
KEGG WNT signalling pathway	***PPP2CA*** ***SKP1*** ***TCF7*** ***WNT8A***	*RAC1* *RHOA* *NFATC3* *NFAT5* *PLCB2* *CHP* ***DAAM2*** ***PPP2CA*** ***SKP1*** ***TCF7*** ***WNT8A***
KEGG JAK-STAT signalling pathway	***IL9***	*EPO* ***IL9***
KEGG MAPK		*RAC1* *JMJD7-PLA2G4B* *CHP* *PLA2G4E* *PLA2G4B* ***CACNA1G***
KEGG DNA replication		*RFC2* *MCM7*
KEGG NOTCH signalling pathway		*DTX2* *DLL4*
KEGG phosphatidylinositol signalling system		*ITPKA* *PLCB2* ***DGKE***
KEGG P13K	***PPP2CA***	*RAC1* *YWHAG* *LAMB2* *RPS6KB1* *EPO* *GNB2* ***CHAD*** ***PPP2CA***
KEGG hedgehog	***WNT8A***	***WNT8A***
KEGG calcium signalling pathway	***VDAC1***	*FZD9* *ITPKA* *PLCB2* *CHP* ***CACNA1G*** ***VDAC1***
KEGG ERBB signalling		*RPS6KB1* *PAK6*
KEGG cell cycle	***SKP1*** ***CDC23***	*YWHAG* *CDC25A* *E2F4* *BUB1B* *MCM7* ***SKP1*** ***CDC23***

(continued)

Table 3 (continued)

	Genes	
	OVH	OVC
KEGG P53 signalling pathway		*SHISA5* *PPM1D* *SERPINE1*
KEGG VEGF signalling pathway		*RAC1* *NFATC3* *NFAT5* *JMJD7-PLA2G4B* *CHP* *PLA2G4E* *PLA2G4B*
Stransky mutation list		
Cancer gene census	***SUZ12***	*PMS2* *RAC1* *ELN* *HIP1* *SETD2* *CLTC* *RNF43* *BRIP1* *CBFB* *CDH1* *BUB1B* ***HLF*** ***SUZ12***

Genes in italics are genes within amplification regions. Genes in bold italics are genes within deletion regions

Many WNTs are frequently overexpressed in head and neck cancers [50]. However, in OVH cohort, genes involved in WNT signalling pathway were in CN loss regions. *SUZ12* gene is located at chromosome 17q11.2, which has been deleted at a frequency of ~15 % in OVH GISTIC deletion plot. The role of *SUZ12* has been investigated previously in epithelial ovarian cancer cells and revealed high significant expression levels when compared with either fallopian tube epithelium or normal human ovarian surface epithelium, as it inhibits apoptosis by stimulating the proliferation of human epithelial ovarian cancer cells [51]. However, no reports were found to illustrate the role of *SUZ12* when down regulated.

Additionally, twelve peaks were identified in OVC cohort, and these regions had a large gene lists. The results were then further analysed by running the gene lists against KEGG pathways, cancer gene census, and Stransky mutation list. Out of 49 key genes hits (refer to Fig. 8), thirteen genes were cancer genes (17 % of the CN altered genes in OVC cohort were found to be related with cancer). Furthermore, eleven genes were involved in KEGG WNT signalling pathway (15 % of the CN altered genes were involved in this pathway) and eight genes in P13K pathway (10 % of the CN altered genes were involved in this pathway). Seven genes were in KEGG cell cycle pathway and seven genes as well in KEGG VEGF signalling pathway (9 % of the CN altered genes are in these pathways). Similarly, six genes were in the KEGG MAPK pathway, and six genes were involved in KEGG calcium

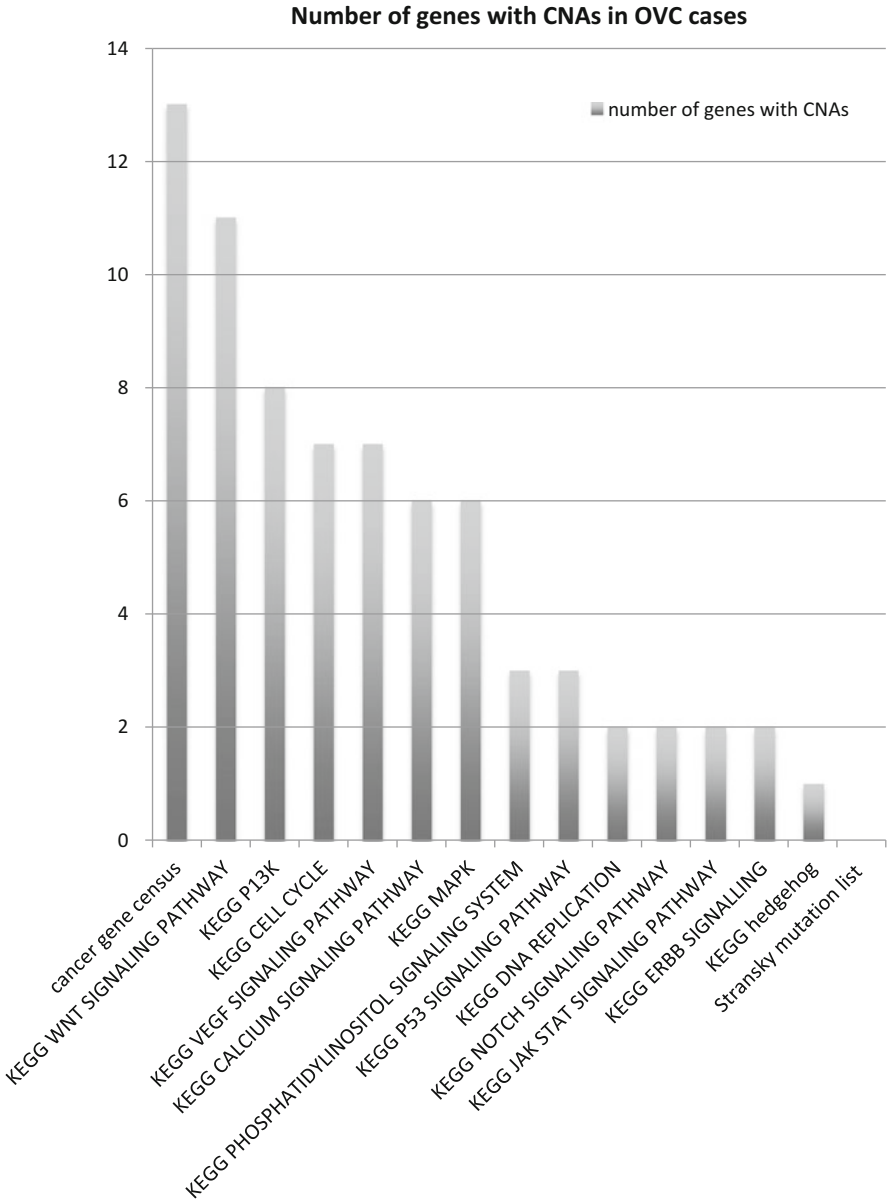


Fig. 8 A graphical representation of 13 enriched KEGG pathways, cancer gene census, and Stransky mutation list [27] on the *x*-axis ranked by the number of genes with CNAs from OVC samples in each pathway and list on the *y*-axis

signalling pathway (8 % of the CN altered genes are in these pathways). Table 3 lists all key genes founded to be in CN altered regions with the highest significant gains and losses in OVC cohort. OVC showed a lower degree of CN alterations compared to OSCCs, and consequently, fewer genes were identified (Table 3).

As can be seen from Table 3, all significant gene hits in OVH group were present in OVC significant gene hit lists, which confirm that OVH is a (histological) precursor for OVCs. In the analysis of OVC gene hit lists, I focused on genes that had a role in head and neck cancers. *CDH1* or epithelial-cadherin gene, located on chromosome 16q22.1, was in a gain chromosomal region and, therefore, is probable to be overexpressed in OVC cohort. The function of E-cadherins has been well established in maintaining junctions. E-cadherin loss enables the disaggregation of malignant cells from one another and promotes metastasis [52, 53]. In human cancers, reduction or loss of E-cadherin expression can be triggered by silencing of the CDH1 promoter, chromosomal deletions, and somatic mutations [52, 53]. However, and in light of the possibility of overexpression of *CDH1* in OVC group, I therefore suggest that this might be a reason behind the fact that OVCs do not metastasise, unlike OSCCs, where the CN data of this cohort showed deletion in chromosome 18q21.3 that harbour *CDH20* gene, which has been reported previously to be involved in tumour invasion regulation [54].

In addition, *MCM7* gene located in chromosome 7q22.1 was in a gain chromosomal region at a frequency of ~50 % according to OVC frequency karyogram (See Fig. 4b) and is therefore probable to be overexpressed in OVC cohort. It has been demonstrated in a previous study that *MCM7* gene is expressed in normal oral mucosa and variably overexpressed in dysplasias and OSCCs [55]. Likewise, *SERPINE1* gene located as well in chromosome 7q22.1 arm that presented a gain is probable to be overexpressed in OVC cohort. In 2005, a study revealed that expression of *SERPINE1* gene in primary head and neck tumours was upregulated compared to normal mucosa [56]. *SERPINE1* overexpression was shown to be essential for the progression of HNSCC and was suggested to play a key role in chromosome 7q21.3–22 karyotypic changes and in oral oncogenesis [57].

5.2.5 Generation of GISTIC Heat Maps for OVH and OVC

Chromosomal alteration regions based on DNA CN changes in OVH and OVC groups are illustrated in the heat maps below generated from GISTIC G-scores analysis (Fig. 9). Visual examination of OVH heat map (Fig. 9a) and OVCs (Fig. 9b) noticeably illustrates the very low level of CNAs in OVHs compared to OVCs, indicating that the genomic profile of these cases has minimal chromosomal abnormalities and is most similar to normal. Nevertheless, gain at chromosome 7q (represented by a lineage red colour) was noticed in OVHs at a frequency of more than 50 %. In addition, as shown in Fig. 4b, gains at chromosome arms 7q, 16q, and 17q (represented by a lineage red colour) and loss at chromosome 5p (represented by a lineage blue colour) were detected in OVCs at a frequency of 50 %. Moreover, deletion trends were minimally found in OVC's heat map. In addition, the thickened, club-shaped papillae and blunt stromal invaginations of well-differentiated squamous epithelium with marked keratinization with the squamous epithelium lacking cytological atypia and histological features of malignancy could be another reason behind the lower level of chromosomal instability in OVCs.

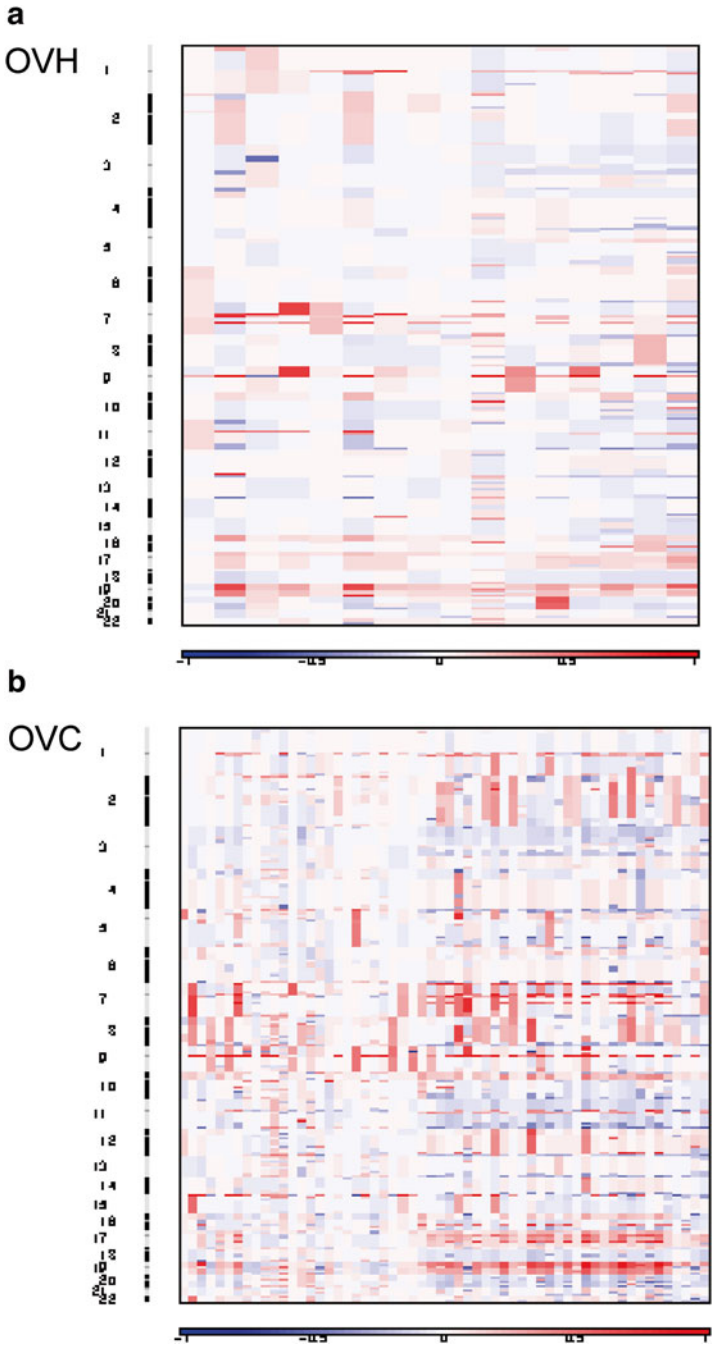


Fig. 9 Heat map images of OVH and OVC based on total segmented DNA copy number variation profiles. Images were analysed using (GISTIC2.0). In each heat map, the samples are arranged from *left* to *right*, and chromosomes are arranged vertically from *top* to *bottom. Red* represents CN gain and *blue* represents CN loss

6 HPV Detection by NGS

Viral load was measured as described in [58]. Briefly, this involved counting the number of reads uniquely aligning to viral genomes. This was scaled to the read depth for the individual sample to calculate the number of KB of viral sequence present per human genome and therefore extrapolates the number of viral genomes per human genome (viral load).

HPV sequencing data from a previous study published by Conway et al. was used to provide positive and negative controls [29]. This technique has been found to have good sensitivity and specificity and has the advantage that it provides HPV subtype, viral load and can be obtained from the same sequencing run which is performed to obtain genomic copy number data. The verrucous cohort was matched with 16 oral and oropharyngeal (OP) cases from the previous study. In Conway et al. data 9, positive HPV cases were detected out of 16 successfully sequenced samples [29]. Sequencing libraries were prepared from all 73 verrucous samples (57 OVC and 16 OVH). HPV-16 sequence was identified in one OVH and one OVC, and HPV-2 sequence was detected in one OVC out of 73 oral verrucous samples at 95 % confidence level with 2.24, 8.16, and 0.33 viral genomes per cell, respectively.

7 Presences of Herpes Virus in Verrucous Samples

Patient tumour DNAs were scanned for all characterised human virus sequences. Human herpesvirus sequences were detected in 21 of 73 verrucous DNA samples, seven OVHs, and 14 OVCs, with viral loads ranging from 0.01 to 0.58 viral genomes per cell. Eleven samples were positive for herpesviruses 1, five were positive for herpesviruses 6B, four were positive for herpesviruses 5, three were positive for herpesviruses 6A, one was positive for herpesviruses 7, and one was positive for herpesviruses 4. Four cases had double herpesviruses infections for herpesviruses 6A and herpesviruses 1, herpesviruses 1 and herpesviruses 6B, herpesviruses 5 and herpesviruses 1, and herpesviruses6A and herpesviruses 6B.

To investigate whether the prevalence of Herpes virus detected was specific to VC cohort in this study, sequencing data from 23 head and neck tumour samples from a previous study published by pre-cancer genomics group [29] were scanned for all characterised human virus sequences. Human herpesvirus sequences were detected in eight out of 23 cases (seven oral and one pharyngeal), with viral loads ranging from 0.0024 to 0.0362 viral genomes per cell and with 0.00702 viral loads standard deviation. From the eight positive samples, four were positive for herpesviruses 1, three were positive for herpesviruses 5, and two were positive for herpesviruses 4. One case had a double herpesviruses infection for herpesviruses 1 and herpesviruses 5.

8 Advantages and Limitations of NGS Copy Number Analysis Technique Used in This Study

It has been previously demonstrated that NGS can provide genomic CN gain and loss details in a cost-effective manner from DNA isolated from different sources, including FFPE tissue blocks stored after histopathological diagnosis, frozen tumour samples, and cell lines [31, 35]. It has been also shown that the resolution of NGS copy number analysis method has a high degree of correlation and comparable with aCGH but gave more information for less money when applied at low multiplexing levels, and it is extremely adjustable [31]. It is also important to keep in mind that aCGH technique has shown difficulty to use with DNA extracted from FFPE materials [31]. Additionally, aCGH requires microgram DNA quantities while NGS can produce CN genomic karyograms from nanogram quantities of DNA (less than 100 ng) [31]. When compared to PCR-based methods such as LOH analysis, NGS produces much more data when performed at high multiplexing levels [31].

The CN analysis method applied in this study provided a digital readout of viral subtypes, loads, as well as tumour karyograms in a single test. It has been also revealed here that good quality CN data can be attained when multiplexing 40 samples on one single lane of an Illumina HiSeq 2500. Multiplexing is an essential aspect in designing research studies according to the required selected resolution, available resources, and accessible sample numbers. Another key point, copy number libraries can be used for several times after being aliquoted. Accordingly, further examination of previously prepared and low-resolution screened libraries can be obtained without the need of additional preparation steps, and hence, data from both screenings on the same sample can be compound to provide a double coverage [31]. Despite the proven utility of next-generation sequencing copy number aberrations detection [31, 35], it cannot detect neutral CN variations (genomic variations that do not cause changes in the amount of the genetic material), such as inversions and balanced translocations [59]. Balanced translocations and inversions that occur in coding region breakpoints can result in a disease phenotype [59]. One of the limitations here was the lack of technical replicates. Though, pre-cancer genomics group previously validated the reproducibility of the same methodology I used here for CN analysis as they did technical replicates back when they first developed the technique [31]. They sequenced the same DNA libraries twice and made libraries from the same DNA twice, and all times, the produced CN karyograms were virtually identical [31].

Another limitation in this study was that we did not check the effect of the fixation procedure on the produced CN genomic profiles by comparing the generated karyograms for DNA extracted from FFPE materials with CN karyograms for DNA extracted from fresh frozen tissue from the same OVC samples. Nevertheless, the rarity of oral verrucous lesions made it really hard to get any fresh frozen OVC samples. DNA is susceptible to degradation in fixative solutions used for tissue preservation in histopathology labs [60], and again, pre-cancer genomics group has previously investigated the effect of fixation on CN karyograms produced from

DNA extracted from FFPE materials when they first developed the method. They compared the CN genomic profiles for DNA extracted from fresh frozen against FFPE materials from the same lung carcinomas [31]. They have shown that the corresponding fixed and fresh CN karyograms for DNA extracted from the same samples were nearly identical [31]. The slight differences were in the magnitude of same CN variants and were attributed to macrodissection of non-cancerous cells in fixed samples, such as inflammatory and stroma cells [31]. Additionally, the lack of paired tumour and normal samples was again another limitation in the current study which if were available would reduce the noise usually associated with CN profiles produced from DNA extracted from FFPE materials [61].

Furthermore, next-generation sequencing was described here as a novel but validated, powerful, high-throughput method to investigate the presence of HPV and all characterised human viral genome loads and subtypes in the largest oral verrucous sample cohort described to date, following careful histological definition for OVC and OVH lesions. Although it is difficult to accurately predict the exact viral load with only a very small number of aligning viral reads, viral loads obtained in this study were clearly much lower than the viral loads obtained in the previously published study of HNSCC by pre-cancer genomics group [29], in which the standard deviation of the viral loads obtained was 37.75 suggesting that the virus was not contributing to disease aetiology. Also, the applied method in the current study was validated before (on the control sample set) by detecting HPV sequences using PCR and by evaluating P16 expression as a marker for HPV infection. It has been shown from the assessment of HPV screening results of the three approaches that NGS method has a high specificity and sensitivity for HPV detection when compared to the two other techniques [29]. Furthermore, it has been previously suggested that PCR methods can be oversensitive [62], while the method used here can provide a better specificity, as demonstrated by the observation that all p16 positive samples were also positive for HPV-16 by sequencing [29]. Moreover, and from the same previous study, HPV-61 was detected in one oral tumour by sequencing and was not detected by any other method, which again shows the ability of this method in detecting all HPV subtypes and loads [29].

Previous studies have relied mostly on PCR and ISH to investigate the presence of HPV subtypes in verrucous lesions without quantitating HPV viral loads [63]. Furthermore, HPV DNA may degrade in paraffin-embedded tissues. Sequencing may be less affected by this than PCR. The standard PCR test for HPV requires a 120-bp fragment to be amplified. DNA libraries are size selected here to be around 200 bp to ensure that enough fragments of <100 bp are sequenced. If an HPV sequence is in one of these, it would be picked up by sequencing but not by PCR. Besides, PCR methods for viral detection are specific to certain subtypes per test. One of the main advantages of using NGS is the fact that sequencing is blind. All known viral subtypes can be quantified in a single test. The method provides a digital readout of viral subtypes and loads with high sensitivity and specificity [29], and the same sequence data can also be reanalysed to produce tumour karyograms. These data are extremely cheap to produce compared to many NGS methods and can be multiplexed to over 40 samples per lane of an Illumina HiSeq.

The power of this method was also shown through the detection of other viruses by screening all verrucous samples for all other known virus sequence genomes. Human herpesviruses were identified in 21/73 of the oral verrucous lesions (28.77 %), although these results have not been confirmed using any other diagnostic test. In addition, the control samples were scanned for all human virus sequences, and eight positive cases were identified out of 23 head and neck samples (34.78 %). In general, Herpes simplex viruses-related infections are among the highest widespread diseases, affecting approximately 60 to 95 % of adult human population [64]. The two human herpesviruses known to be associated with cancer are Kaposi's sarcoma-associated herpesvirus (KSHV) and Epstein-Barr virus (EBV) [65], and these were not detected in oral verrucous samples here. Nonetheless, it is important to point out that the detection of virus DNA in patients' samples does not essentially indicate a viral pathogenic role in a disease. NGS tells nothing about transcriptional activity, so it is not possible to speculate further on the clinical significance of this finding. However, by infecting defence cells, many herpesviruses can persistently arise in different human tissues in the event of inflammation [66], and accordingly, viral genomes accumulate till they become detectable in these infected cells [64]. The finding that herpes sequence could be detected in 28.77 % of oral verrucous lesions while those lesions did not harbour any HPV infection shows further the value of this method. Herpes infection may not be the cause of this disease, but future studies of a similar nature may reveal previously unsuspected oncoviruses to be common in a different tumour type.

9 Summary

In this study, NGS was used on nanogram quantities of DNA isolated from FFPE tissue, in the largest oral verrucous sample cohort described to date. The aim was to identify OVH and OVC genomic characteristic features and distinguish between the genomic damage pattern in OVH and OVC. The current study has demonstrated that NGS CN analysis can be used for more specific assessment and evaluation of OVH and OVC heterogeneity based on the analysis of the whole-genome CN karyograms. The results of this study also suggest that oral verrucous lesions are not associated with HPV or any other human virus.

References

1. Ferlay J, et al. Estimates of worldwide burden of cancer in 2008: GLOBOCAN 2008. Int J Cancer. 2010;127(12):2893–917.
2. Scully C, Bagan JV. Oral squamous cell carcinoma: overview of current understanding of aetiopathogenesis and clinical implications. Oral Dis. 2009;15(6):388–99.
3. Bagan J, Sarrion G, Jimenez Y. Oral cancer: clinical features. Oral Oncol. 2010;46(6):414–7.
4. Kademani D. Oral cancer. Mayo Clin Proc. 2007;82(7):878–87.

5. Cabay RJ, Morton Jr TH, Epstein JB. Proliferative verrucous leukoplakia and its progression to oral carcinoma: a review of the literature. J Oral Pathol Med. 2007;36(5):255–61.
6. Zhu LK, et al. A clinicopathological study on verrucous hyperplasia and verrucous carcinoma of the oral mucosa. J Oral Pathol Med. 2012;41(2):131–5.
7. Califano J, et al. Genetic progression model for head and neck cancer: implications for field cancerization. Cancer Res. 1996;56(11):2488–92.
8. Thomson PJ, et al. Epithelial cell proliferative activity and oral cancer progression. Cell Prolif. 2002;35:110–20.
9. Kushner J, Bradley G, Jordan RCK. Patterns of p53 and Ki-67 protein expression in epithelial dysplasia from the floor of the mouth. J Pathol. 1997;183(4):418–23.
10. Gonzalez-Moles MA, et al. Suprabasal expression of Ki-67 antigen as a marker for the presence and severity of oral epithelial dysplasia. Head Neck. 2000;22(7):658–61.
11. Shear M, Pindborg JJ. Verrucous hyperplasia of the oral-mucosa. Cancer. 1980;46(8): 1855–62.
12. Wang YP, et al. Oral verrucous hyperplasia: histologic classification, prognosis, and clinical implications. J Oral Pathol Med. 2009;38(8):651–6.
13. Hsue SS, et al. Malignant transformation in 1458 patients with potentially malignant oral mucosal disorders: a follow-up study based in a Taiwanese hospital. J Oral Pathol Med. 2007;36(1):25–9.
14. Ho PS, et al. Malignant transformation of oral potentially malignant disorders in males: a retrospective cohort study. BMC Cancer. 2009;9:260.
15. Ackerman LV. Verrucous carcinoma of the oral cavity. Surgery. 1948;23(4):670–8.
16. Barnes L, Eveson JW, Reichart P, Sidransky D. WHO classification of tumours, pathology and genetics of head and neck tumours. Lyon: IARC Press; 2005. p. 174–5.
17. Pentenero M, et al. Distinctive chromosomal instability patterns in oral verrucous and squamous cell carcinomas detected by high-resolution DNA flow cytometry. Cancer. 2011;117(22):5052–7.
18. Ray JG, et al. Oral verrucous carcinoma – a misnomer? Immunohistochemistry based comparative study of two cases. BMJ Case Rep. 2011;2011.
19. Bagan J, et al. Proliferative verrucous leukoplakia: a concise update. Oral Dis. 2010;16(4): 328–32.
20. Alkan A, et al. Oral verrucous carcinoma: a study of 12 cases. Eur J Dent. 2010;4(2):202–7.
21. Chung CH, et al. Oral precancerous disorders associated with areca quid chewing, smoking, and alcohol drinking in southern Taiwan. J Oral Pathol Med. 2005;34(8):460–6.
22. Stokes A, et al. Human papillomavirus detection in dysplastic and malignant oral verrucous lesions. J Clin Pathol. 2012;65(3):283–6.
23. Syrjänen KJ, Stina SS, Syrjänen M. Papillomavirus infections in human pathology. New York: Wiley; 2000.
24. Klieb HB, Raphael SJ. Comparative study of the expression of p53, Ki67, E-cadherin and MMP-1 in verrucous hyperplasia and verrucous carcinoma of the oral cavity. Head Neck Pathol. 2007;1(2):118–22.
25. Haimovich AD. Methods, challenges, and promise of next-generation sequencing in cancer biology. Yale J Biol Med. 2011;84(4):439–46.
26. Agrawal N, et al. Exome sequencing of head and neck squamous cell carcinoma reveals inactivating mutations in NOTCH1. Science. 2011;333(6046):1154–7.
27. Stransky N, et al. The mutational landscape of head and neck squamous cell carcinoma. Science. 2011;333(6046):1157–60.
28. Lui VWY, et al. Frequent mutation of the PI3K pathway in head and neck cancer defines predictive biomarkers. Cancer Discov. 2013;3(7):761–9.
29. Conway C, et al. Next-generation sequencing for simultaneous determination of human papillomavirus load, subtype, and associated genomic copy number changes in tumors. J Mol Diagn. 2012;14(2):104–11.
30. Meyerson M, Gabriel S, Getz G. Advances in understanding cancer genomes through second-generation sequencing. Nat Rev Genet. 2010;11(10):685–96.

31. Wood HM, et al. Using next-generation sequencing for high resolution multiplex analysis of copy number variation from nanogram quantities of DNA from formalin-fixed paraffin-embedded specimens. Nucleic Acids Res. 2010;38(14):e151.
32. Woolgar JA, Triantafyllou A. Pitfalls and procedures in the histopathological diagnosis of oral and oropharyngeal squamous cell carcinoma and a review of the role of pathology in prognosis. Oral Oncol. 2009;45(4–5):361–85.
33. Eversole LR, Papanicolaou SJ. Papillary and verrucous lesions of oral mucous membranes. J Oral Med. 1983;38(1):3–13.
34. Xie C, Tammi MT. CNV-seq, a new method to detect copy number variation using high-throughput sequencing. BMC Bioinformatics. 2009;10:80.
35. Hayes JL, et al. Diagnosis of copy number variation by Illumina next generation sequencing is comparable in performance to oligonucleotide array comparative genomic hybridisation. Genomics. 2013;102(3):174–81.
36. Chiang DY, et al. High-resolution mapping of copy-number alterations with massively parallel sequencing. Nat Methods. 2009;6(1):99–103.
37. Gusnanto A, et al. Correcting for cancer genome size and tumour cell content enables better estimation of copy number alterations from next-generation sequence data. Bioinformatics. 2012;28(1):40–7.
38. Nguyen DQ, Webber C, Ponting CP. Bias of selection on human copy-number variants. PLoS Genet. 2006;2(2):198–207.
39. Sharp AJ, et al. Segmental duplications and copy-number variation in the human genome. Am J Hum Genet. 2005;77(1):78–88.
40. Poh CF, et al. A high frequency of allelic loss in oral verrucous lesions may explain malignant risk. Lab Invest. 2001;81(4):629–34.
41. Mohapatra G, et al. Glioma test array for use with formalin-fixed, paraffin-embedded tissue - Array comparative genomic hybridization correlates with loss of heterozygosity and fluorescence in situ hybridization. J Mol Diagn. 2006;8(2):268–76.
42. Mermel CH, et al. GISTIC2.0 facilitates sensitive and confident localization of the targets of focal somatic copy-number alteration in human cancers. Genome Biol. 2011;12(4):R41.
43. Bass AJ, et al. SOX2 is an amplified lineage-survival oncogene in lung and esophageal squamous cell carcinomas. Nat Genet. 2009;41(11):1238–42.
44. Firestein R, et al. CDK8 is a colorectal cancer oncogene that regulates beta-catenin activity. Nature. 2008;455(7212):547–51.
45. Lin WM, et al. Modeling genomic diversity and tumor dependency in malignant melanoma. Cancer Res. 2008;68(3):664–73.
46. Etemadmoghadam D, et al. Integrated genome-wide DNA copy number and expression analysis identifies distinct mechanisms of primary chemoresistance in ovarian carcinomas. Clin Cancer Res. 2009;15(4):1417–27.
47. Weir BA, et al. Characterizing the cancer genome in lung adenocarcinoma. Nature. 2007;450(7171):893–8.
48. Chiang DY, et al. Focal gains of VEGFA and molecular classification of hepatocellular carcinoma. Cancer Res. 2008;68(16):6779–88.
49. Northcott PA, et al. Multiple recurrent genetic events converge on control of histone lysine methylation in medulloblastoma. Nat Genet. 2009;41(4):465–72.
50. Barker N, Clevers H. Mining the Wnt pathway for cancer therapeutics (vol 5, pg 997, 2006). Nat Rev Drug Discov. 2007;6(3):249.
51. Li H, et al. SUZ12 Promotes Human Epithelial Ovarian Cancer by Suppressing Apoptosis via Silencing HRK. Mol Cancer Res. 2012;10(11):1462–72.
52. Berx G, et al. Mutations of the human E-cadherin (CDH1) gene. Hum Mutat. 1998;12(4):226–37.
53. Onder TT, et al. Loss of E-cadherin promotes metastasis via multiple downstream transcriptional pathways. Cancer Res. 2008;68(10):3645–54.
54. Vermeulen S, et al. Regulation of the invasion suppressor function of the cadherin/catenin complex. Pathol Res Pract. 1996;192(7):694–707.

55. Tamura T, et al. Minichromosome maintenance-7 and geminin are reliable prognostic markers in patients with oral squamous cell carcinoma: immunohistochemical study. J Oral Pathol Med. 2010;39(4):328–34.
56. Chin D, et al. Novel markers for poor prognosis in head and neck cancer. Int J Cancer. 2005;113(5):789–97.
57. Chen YJ, et al. Genome-wide profiling of oral squamous cell carcinoma. J Pathol. 2004;204(3):326–32.
58. Samman M, et al. Next-generation sequencing analysis for detecting human papillomavirus in oral verrucous carcinoma. Oral Surg Oral Med Oral Pathol Oral Radiol. 2014;118(1):117–25. e1.
59. Coughlin CR, Scharer GH, Shaikh TH. Clinical impact of copy number variation analysis using high-resolution microarray technologies: advantages, limitations and concerns. Genome Med. 2012;4:80.
60. Ferrer I, et al. Effects of formalin fixation, paraffin embedding, and time of storage on DNA preservation in brain tissue: a BrainNet Europe study. Brain Pathol. 2007;17(3):297–303.
61. Bhattacharya A, et al. Two distinct routes to oral cancer differing in genome instability and risk for cervical node metastasis. Clin Cancer Res. 2011;17(22):7024–34.
62. Smeets SJ, et al. A novel algorithm for reliable detection of human papillomavirus in paraffin embedded head and neck cancer specimen. Int J Cancer. 2007;121(11):2465–72.
63. Miller CS, Johnstone BM. Human papillomavirus as a risk factor for oral squamous cell carcinoma: a meta-analysis, 1982–1997. Oral Surg Oral Med Oral Pathol Oral Radiol Endod. 2001;91(6):622–35.
64. Brady RC, Bernstein DI. Treatment of herpes simplex virus infections. Antiviral Res. 2004; 61(2):73–81.
65. Everly D et al. 2012. Herpesviruses and cancer. In:Robertson ES. Cancer associated viruses. Springer-Verlag New York, pp. 133–167.
66. Ferreira DC, et al. Identification of herpesviruses types 1 to 8 and human papillomavirus in acute apical abscesses. J Endod. 2011;37(1):10–6.

VIRONOMICS: The Study of Viral Genomics in Human Cancer and Disease

Dirk P. Dittmer, Dongmei Yang, Marcia Sanders, Jie Xiong, Jordan Texier, and Rachele Bigi

Abstract Viruses cause approximately 30 % of all human cancers. New viruses are discovered weekly, as are novel, putative associations between viruses and cancers. Next Generation Sequencing (NGS) has evolved as a new tool to find viruses in cancer and to support virus–cancer associations. Importantly, NGS-based approaches can be applied to clinical samples without the need for intermediate culture of the agent, and the approach is agnostic with regard to the target sequence. This allows for the discovery of entirely novel, as well as novel but evolutionary related viral agents. Since viral genomes are so much smaller than the human genome, they offer unique opportunities and challenges in NGS. Here, we outline some of these challenges and potential bioinformatics solutions using Kaposi Sarcoma-associated herpesvirus (KSHV) as an example. We provide an abbreviated overview about viral cancers as well as NextGen sequencing platforms. This is followed by a summary of open source computing tools as they apply to the bioinformatics analysis of viral contributions to cancer, as well as a virus-specific case study mapping open chromatin regions in KSHV.

D.P. Dittmer (✉)
Department of Microbiology and Immunology, Lineberger Comprehensive Cancer Center, Center for AIDS Research (CfAR), Carolina Center for Genome Sciences, University of North Carolina at Chapel Hill, Chapel Hill, NC 27599-7290, USA
e-mail: ddittmer@med.unc.edu

D. Yang • M. Sanders • J. Texier
Department of Microbiology and Immunology, Lineberger Comprehensive Cancer Center, Center for AIDS Research (CfAR), Carolina Center for Genome Sciences, University of North Carolina at Chapel Hill, Chapel Hill, NC 27599-7290, USA

Program in Global Oncology, Lineberger Comprehensive Cancer Center, and Center for AIDS Research (CfAR), The University of North Carolina at Chapel Hill, Chapel Hill, NC, USA

J. Xiong
The University of North Carolina at Chapel Hill, Chapel Hill, NC, USA

Department of Statistics and Operations Research, The University of North Carolina at Chapel Hill, Chapel Hill, NC, USA

R. Bigi
Department of Experimental Medicine, University of Rome "La Sapienza", Rome, Italy

W. Wu, H. Choudhry (eds.), *Next Generation Sequencing in Cancer Research, Volume 2*, DOI 10.1007/978-3-319-15811-2_20

1 Introduction

1.1 Motivations and Expectations

It is very difficult to write for a modern book in bioinformatics. On the one hand the publishing tradition in molecular biology and cancer biology is almost exclusively focused on publishing in high impact journals and on producing "minimally publishable units" as quickly as possible. Only journal articles have an impact factor and count towards promotion and funding. Cancer biology and cancer bioinformatics is driven almost entirely by the production of novel data and access to novel samples. The majority of time is spent on "wetlab" work and the production of raw reads. On the other hand, most informatics work has migrated to open source biorepositories, github [1], blogs, wikis, or open access, online-only publications such as the "BMC" or "PLoS" series. Almost all information is gleaned from the Web and almost all data are available in the cloud, or more specifically the short read archives (SRA) and db GaP data repositories [2]. Few bioinformaticians spend time reading printed works.

What then is the purpose and utility of this chapter? An edited, peer-reviewed book chapter is better organized than the Web and the authors would argue that the information presented here, can be accessed more timely and be digested more easily than bits and pieces gathered from various websites. We hope that this chapter would save you, the reader, many "clicks."

Writing this book chapter provided us with an opportunity to present scientific details of analysis that are not found in journal articles. Journal articles, because of their word limitation policies, are focused on the end result, not the path towards this end. In this chapter the reader can trace the steps, which we used to obtain a nucleotide resolution map of open chromatin in Kaposi sarcoma associated herpesvirus (KSHV) [3]. We present the tools, commercial and open source, and examples of code.

1.2 Viruses Cause Cancer

Viruses cause 20–30 % of all human cancers [4]. If we can recognize, understand, and eradicate these "oncoviruses," we can cure the corresponding virus-dependent cancers. Because only the tumor cells carry the oncovirus, targeting viral genes will yield drugs and interventions of superior specificity compared to traditional therapies. Vaccination against human papilloma virus (HPV) prevents cervical cancer and has led to a significant reduction in disease burden in the USA, Europe, and Australia [5]. Vaccination against hepatitis B virus (HBV) has led to a significant reduction in liver cancer in all areas where the vaccine has been introduced. New drugs against hepatitis C virus (HCV), which were introduced last year, prevent liver cancer and will replace surgical liver transplants as the standard of care for this disease [6].

In order to develop these interventions, the virus had to be identified in the first place and associated with a specific disease. The discovery of human papilloma virus (HPV), before the introduction of sequencing earned Harald zur Hausen the 2008 Nobel Prize in medicine. His Nobel lecture provides an interesting historical perspective [7]. Once discovered, viral biology has to be explored. Because viruses are extremely multifaceted, the NGS approaches to cancer-associated viruses are very diverse indeed. DNA-based viral genomes (HPV, KSHV) can be analyzed by the same means as the human genome. RNA-based viruses almost always present as a quasi-species with extreme nucleotide variation [8], even a swarm of viruses circulating within the same person. Here, more specialized approaches are needed.

1.3 Aspects of Viral Genomics or "Vironomics"

The twentieth century revolution in molecular biology started with viral genomics. The first restriction map was that of simian virus 40 (SV40). SV40 was also the first completely sequenced genome [9, 10]. Genome-wide transcription mapping was first completed for Adenovirus, and transcriptional analyses of any kind would not be possible without Moloney murine leukemia virus (MMLV) or avian leukosis virus (ALV) reverse transcriptase. The twenty-first century witnessed tremendous advances in host genomics such as polymerase chain reaction (PCR), hybridization-based glass microarrays, bioinformatics, and automated Sanger sequencing that cumulated in the release of the draft human genome sequence. NGS promises to accelerate genomic approaches to cancer virology even further [11]. The $100 viral genome is a reality today and most of the cost is due to viral DNA/RNA isolation and bioinformatics analysis. How do we harness these methods for virus research? What are the specifics of viral genomes and virus lifestyles that we need to consider before blindly transferring tools that were developed and optimized for multi-megabase genomes?

All viruses have small genomes. With few exceptions a virus genome is less than a Megabase (10^6 bp); the HPV genome is less than 10,000 bp and its initial sequence was still printed on paper for the reader to look at by hand [12]. In the early days, the yellow highlighter was the predominant bioinformatics tool, and investigators searched for restriction enzyme sites by reading the printed sequence. Today, GUI-driven visualization tools have replaced the highlighter.

The small genomes of viruses have a distinct advantage for bioinformatics. The entire sequence can be loaded into memory and presented at nucleotide resolution on a desktop computer. This facilitates a visual and interactive approach to viral genomics. Alignments and computer predictions can be verified by biologists, and if need be, corrected manually. Also, because of the small genome, exhaustive, but complete algorithms can be applied. Graph-based de novo assemblers, such as *Newbler*, work as well as de Bruijn graphs-based assemblers for viral genomes [13–15]. For viral genomics, there is no immediate need for heuristic/probabilistic algorithms and the programming overhead that is associated with "big data" is not needed, neither is the "cloud."

The small genomes of viruses mandate a higher standard of genome finishing. In order to submit a viral genome, the assembly has to be 100 % complete and represented as a single continuous sequence [16]. It is expected that all gaps be closed by targeted sequencing. This is in contrast to larger genomes, including the human genome, which was declared finished in 2004, but is not fully sequenced even today [17, 18]. A more recent example of an incomplete draft genome is the panda genome [19, 20].

For viruses, whole genome de novo sequencing efforts can easily be completed on a desktop computer [21, 22]. The devil, however, is in the details. Even today, where personal NGS machines are present in midsize labs, the larger genome centers employ teams of "genome finishers," who have in-depth knowledge of the peculiarities of different hardware used, of each assembly algorithm, and ideally of the biology of their sequencing target. Though the viral genomes are comparatively small, this does not diminish the time and effort needed for "post production." "Post-production" genome finishing easily accounts for ten times the time and effort than what is needed to submit raw sequence reads or a number of unconnected scaffolds. We would emphasize that because viral genomes are small, every nucleotide counts and the accuracy needed to deliver useful genomic information is higher than for eukaryotic genomes.

The ends of the viral genome present a unique problem for viral sequencing. Many viral genomes are linear and the ends may be blocked by unique modifications to the DNA/RNA or by associated proteins [23]. The easiest case is that of circular viral genomes, e.g., episomal herpesvirus genomes maintained as extrachromosomal plasmids in Burkitt Lymphoma, or by integrated viruses, e.g., oncogenic retroviruses. Those can be sequenced easily after adjusting parameters to account for end-to-end connectivity or to detect virus–host DNA fusion events. Neither of those scenarios is found in viral DNA/RNA isolated from virions circulating in body fluids. Sampling body fluids such as blood or saliva, however, is much more applicable than the collection of tumor biopsies, which for many cancers are associated with a small but not infinite risk to the patient. Most packaged viral genomes adopt a strictly linear form. Sanger sequencing as well as NGS needs a free 3′-hydroxyl group to extend and a primer. Only Maxim-Gilbert sequencing can determine free ends of linear DNA genomes [24, 25]. Most viral termini do not represent new sequence information, but are derived from the first consensus sequence that was deposited into GenBank and used to make the sequencing primer. Few studies make the effort to preserve the authentic ends and anneal primers by RNA–DNA or DNA–DNA linker ligation.

2 NextGen Sequencing Platforms

Because the viral genomes are small, often contain unbalanced GC ratios and various repeats, different NGS technologies should be combined for the final assembly. In many cases viral DNA can be amplified by physical means and a fraction of the

reads needed for human genome coverage are needed to cover viral genomes. For a 10,000 base pair (bp) genome, 10,000–100,000 reads can provide appropriate coverage depending on the read length. Viral genome assembly places a premium on read length because of the many repeats within a typical viral genome.

All of the current next-generation sequencers work off of the same basic requirement; sequencing is dependent on high-quality samples to begin with. The enzymatic reactions and sequencing technology will work best when optimal samples are input into the system. Samples must be assessed for quality and quantity before anything can be done. For economic reasons the samples should be as pure as possible. NGS enzymes and sequencing reaction are very sensitive to contamination and concentration, and therefore, this requires different technology than one would use for cloning or polymerase-chain-reaction (PCR). Nanodrop technology is not accepted for next-generation sequencing when assessing DNA quantity. Real-time quantitative PCR, digital PCR, or fluorescent-dye based assays, like Picogreen, are required.

All of the next-generation sequencing begins at the same point: sample preparation. This is commonly referred to as library preparation. It does not matter what platform is used eventually, each requires double-stranded DNA (dsDNA) to be fragmented, the ends repaired, and then sequencing platform specific adapters to be added. Although the basic steps of the library preparation are the same between platforms, the individual reagents are specific to the starting material and sequencing platform. Many platforms are capable of sequencing genomic, cDNA, RNA, BAC, plasmid, and other types of samples, but before the library preparation can take place, all input material must be converted to double-stranded DNA. Many platform kits, such as the RNA-seq kits, include these reagents. After library preparation, different technologies approach NGS differently. Several use amplification in order for the platform to easily read each DNA fragment. Illumina uses "bridge amplification," while Ion Torrent and 454/Roche use emulsion polymerase-chain-reaction (emPCR). Emulsion PCR takes DNA fragments and PCR amplification within an oil bubble (emulsion), afterwards all the amplified fragments are then attached to a bead via hybridization to a biotin–streptavidin-linked oligonucleotide. Success at library preparation and overall NGS depends on several parameters: (1) concentration of starting material (sample), (2) base pair length of original material, (3) desired length of sequencing reads, (4) desired depth of sequencing, and (5) turnaround time. No one approach is optimal with respect to all of these parameters.

2.1 Roche/454 Life Sciences 2008: "One Fragment = One Bead = One Read"

454 Life Sciences produced the first commercially available next-generation sequencer. The best-known version, the FLX (pronounced flex), was available for purchase starting in 2005. The next one, the GS Junior, was available in 2009. It became the first "desktop" sequencer; it was small and fit easily in lab spaces.

The GS Junior still has the longest read length of any of the "personal" sequencers. The workflow for both Roche instruments starts with library preparation, followed by emulsion PCR (emPCR), and finally the sample (and enzymes) end up in individual wells on a plate/chip. 454 sequencing technology is based on one bead into one well. One nucleotide at a time, T, A, C, or G, is passed over the plate and if the complementary nucleotide is incorporated, a pyrophosphate (luciferase) is released and light is detected. The niche for this platform is the ability for long-read lengths, high call accuracy, and "short" NGS run time. Typical read length averages 400–600, with the new "plus" technology increasing averages to 800–1,000 bps. Single run time varies from 10 to 23 h depending on application, with longer time for longer reads.

2.2 Illumina 2009: "Sequencing by Synthesis"

Illumina first started in 2009 with the Genome Analyzer. Since then, Illumina has released several platforms, including new ones released in 2013/2014. Current machines include the MiSeq, HiSeq, and NextSeq, as well as HiSeqX Ten, an expansion of Hiseq. All of the platforms use sequencing by synthesis (SBS) technology. Basically, chips have lanes (as opposed to wells like the other NGS technologies) with two types of oligonucleotides attached. Sample fragments anneal to the first oligonucleotide and are clonally amplified by "bridge amplification" as the other end of the fragment anneals to the second oligo. Post amplification, all four fluorescently labeled nucleotides are passed over the lane for each cycle and competitively incorporated by complementary nucleotide, which elicits a specific fluorescence signal. Fragments can be sequenced in one direction, single-end (SE), or the reverse strand can also be sequenced, paired-end (PE). This produces twice as many reads per fragment. Note that Illumina uses the term "paired-end," whereas Ion Torrent and 454/Roche use the term "mate-pair" to describe sequencing longer than read-length fragment from both ends with potentially unknown sequence of 5–20,000 nucleotides in between. Depending on the platform and single-end vs. paired-end, the read length is limited to 48–200 nucleotides, with only the MiSeq capable of 300 nucleotides. Illumina machines produce around 25–4,000 million reads. Actual sequencing time on the popular HiSeq for a paired-end sequencing can take up to 11 days. On the MiSeq newer "rapid run" parameters will decrease this time to 7 h.

2.3 Life Technologies Ion Torrent (PGM and Proton) 2011: "Ion Semiconductor Chip"

Ion Torrent Sequencers are based on ion semiconductor chip technology. They have two platforms available, the Personal Genome Machine (PGM) and the "Proton." Both rely on the DNA fragments being attached to beads, amplified, and then one

bead is randomly distributed into the sequencing chip well. Instead of fluorescence detection like 454, these platforms depend on a hydrogen ion (H+) being released each time a complementary nucleotide is incorporated, causing a pH change, and thus a voltage change. The instrument records this. Both the PGM and Proton have a typical read length of 200 bp. Recent upgrades to the PGM extend the read-length to 400 nucleotides. Sequencing is completed within 7 h.

2.4 PacBio 2011: "Single Molecule Real Time Sequencing"

Pacific Biosystems (PacBio) released their next-generation sequencer in 2011. PacBio does not use an amplification of the DNA fragments based approach. Instead it uses single molecule real time sequencing (SMRT) technology. Libraries include a hairpin adapter attached to the template and each single library fragment on the SMRT cell ends up in a "zero mode wave guide." This "guide" is essentially a long single well with polymerase immobilized at the bottom. The DNA template feeds through the polymerase and as a complementary phosphor-linked nucleotide is incorporated, the specific fluorophore is released and the sequencer records the specific nucleotide. This technology enables read lengths of 500–20,000 bp, depending on the quality of the original sample and fragmentation settings. The instrument generally only reads the long library strands once, while short fragments are read on multiple passes. The PacBio is one of the quickest next-generation sequencing platforms with a minimum of 30 min on the sequencer. It has the longest read-length, but until recently was plagued by a very high error rate.

3 Bioinformatics of NextGen Sequencing

3.1 Step 1 Alignment to a Reference Genome

There are various open source software tools for NGS data analysis. For any bioinformatics analysis, the first step is to decide which tools to choose. We usually rely only on software that has a large user base and thus community support; it should be under active improvement with new releases on an annual or biannual basis. The first step of the data analysis is to do quality assessment [26]. FastX-Toolkit [27] has a set of tools for data preprocessing including quality statistics, reads filtering and trimming, etc.

After cleaning the raw data, the next step is to align the cleaned NGS reads to the reference genome. Bowtie2 [28, 29] is the most widely used alignment tool due to its speed and accuracy; another one is VELVET [30]. Appendix 1 provides for a short introduction to its command line version and how it can be used to subtract out human genome sequences when the overall goal is to assembler a viral genome. It is regularly released with new features, bug fixes and improvements. There are

large sets of parameters that users can manipulate depending what the targets are. In order to use a reference genome, bowtie2-build is used to build the Bowtie index. Then bowtie2 is used to align the reads to the indexed reference genome.

The output from Bowtie2 is in SAM format [31, 32]. Hence, the next step is to use SAMtools/BCFtools [32] to analyze the alignment or use the unaligned reads for further analysis. The samtools view function converts SAM format to BAM format, then sort function sorts the BAM file. The resultant .bam file is compressed and sorted, which is convenient for both storage and variant discovery. Samtools mpileup function calls on the sorted .bam file and generates a bcf file that stores the likelihood given each possible genotype, then the BCFtools is applied to the bcf file and reports the variants in VCF (variant call format). VCF is a standard format for storing variant data.

3.2 Variant Detection

In order to discover rare mutations for cancer causative genes, the GATK program has become the standard for cancer genomics research [33, 34]. GATK is an abbreviation for "Genome Analysis Toolkit." The typical workflow for variant analysis on NGS data starts with the data preprocessing to make it suitable for variant calling analysis. The first step is to map the raw reads to the reference using Bowtie2 or BWA to generate the alignment in SAM format, then to use PICARD to sort (by coordinates) and convert the SAM to a BAM file, the Picard MarkDuplicates function marks the duplicates reads (dedupping), then calls on BuildBamIndex to index the dedupped BAM to a BAI file. All mapping algorithms from Bowtie2 or BWA tend to generate artifacts, especially for the alignment on the edges of indels. So GATK provides additional tools that realign the reads to clean up these artifacts. The last step of data preprocessing is to do base quality score recalibration. All the GATK variant calling algorithm use the base quality score, the quality score recalibration (BQSR) uses machine-learning method to eliminate systematic errors and gives more accurate base quality score. All in all a larger number of quality control and quality improvement steps are needed for accurate analysis. Variations of these open source tools are incorporated into most commercial software packages and provide the same functionality with an easy to use GUI.

Once the data is ready for variant analysis, we use the GATK variant discovery toolset to discover the meaningful variants. One big challenge for variant discovery is to balance the sensitivity (false negatives) and the specificity (the false positives). In order to achieve the high sensitivity and specificity, GATK variant discovery takes three steps: (1) variant calling (per-sample) (2) joint genotyping (per-cohort) (3) variant filtering (per-cohort).

The HaplotypeCaller [35] is also commonly used for variant calling including SNPs and indels; it is designed to achieve high sensitivity to avoid missing real variants, but the same time, it introduces a certain amount of false positives. GATK uses variant quality score recalibration (VQSR) to filter out the "bad" variants.

Of relevance to cancer virus variant detection, VQSR applies a machine learning method, which does not work well on small datasets and targeted sequencing data. To overcome this limitation we tend to apply hard-filtering using related public databases and in-house generated database. For the joint genotyping step, the initial GATK joint discovery workflow was quite computational intensive. With the new release of GATK version 3.0, there is an enhanced workflow that significantly reduces computational burden. For targeted sequencing projects, we apply instead statistical computing using R code and bioconductor [36] modules to conduct joint genotyping on multiple samples per cohort.

In the variant calling pipeline, it is very important to use known sites to help distinguish true variants from false positives. For cancer genomics, GATK provides sets of known sites as a resource bundle for human genomes. For instance, we typically use SNP data from HapMap as known sites, because the HapMap SNP call set has been validated to a very high degree of confidence [37].

Once the final set of variants is generated, we generally perform preliminary analyses before validating an individual single nucleotide variant (SNV) experimentally through targeted Sanger-based resequencing. Functional annotation is very important to find out if the variants and genotypes are biologically relevant. A popular tool that performs functional annotation is SnpEff [38], which conducts a comprehensive functional analyses and reports effects by type (SNPs, INDELs, etc.), functional class (missense, nonsense, silent), region (exon, intron, intergenic, downstream, upstream, splice-site acceptor, splice-site donor, etc.). The most biologically interesting SNVs based on the SnpEff annotation are subjected to further experimental testing and validation.

3.3 RNAseq, ChIPseq, and Related Techniques

The pipeline for RNA-seq variant analysis is similar to the one for DNA-seq. For differential expression analysis based on RNA-seq, the first step is also an alignment. Both Bowtie2 and BWA can be used for aligning unspliced RNA-seq data. TopHat [39] and STAR [40] are the popular aligners taking spliced messages into account. Some recent commercial alternatives are CLC genomics workbench by Qiagen Inc. or Genious [41] as used in the following examples. By-and-large these programs are simple GUI wrappers for the same open source programs. They provide value because they eliminate the barrier of having to learn command line and to maintain open source installations. After sequence alignment, quantitative analysis and differential expression analysis are conducted. Cufflinks [42, 43] and DESeq (bioconductor package) are the leading tools [44]. RNAseq and ChIPseq follow a similar analyses paradigm, except that for ChIPseq, where there exists an a priori expectation of discrete peaks. Here, additional compensation and aggregations algorithms are employed. This will be discussed in detail below using FAIRE on a human tumor virus as example.

4 Sequence Read Alignment for All Possible Viruses

There is only one human genome and one genome reference as the target for NGS. There are, however, many viruses and multiple can be present in the same sample. Often, the goal is to enumerate just which viruses are present and to determine the so-called "Vironome" [45]. Sometimes this will yield to the discovery of an entirely new cancer virus [46, 47]. Almost always it is a necessary step to "bioinformatically" sieve out all nonviral sequences, human and bacteria, prior to further mapping and assembly. For de-novo assembly of viral genomes cloned into bacterial artificial chromosomes (BAC) bioinformatically subtracting out *E. coli* K12 sequences prior to de novo assembly yielded great improvements in our experience.

Vivonatev is a program written by us, with the goal to manage the mapping of NGS reads to large numbers of reference genomes. With the advance of NGS, the quantity of reference genomes has increased; as has the number of completely sequences strains available in GenBank. This has made mapping NGS reads to all known viral genomes a very difficult task. The purpose of Vivonatev is to automatically generate such a mapping. It uses bowtie2 as the mapping tool and a list of reference/target genomes is provided as input. The program is also capable of extracting, filtering, and/or generating a coverage vector for each reference genome. These coverage vectors can be visualized using R, or even excel, since viral genomes are small. This may be useful for further analysis of the NGS reads. In order to save computing time, Vivonatev is also able to decide whether additional actions are necessary, depending on the mapping ratio.

For each reference in the input list, Vivonatev runs the alignment tool bowtie2. Bowtie2 determines the mapping ratio of the NGS to the reference genome. If the mapping is above a certain ratio (determined by the user, default = 0), Vivonatev performs the following actions (again, determined by the user):

- Filtering: If a read mapped to the reference, it is taken out of the original set of reads. At the end of the run a new filtered set is created in a new file.
- Extracting: If a read maps to the reference, it is copied into a separate file, specific for each reference.
- Covering: A coverage vector of the reference genome is created.

If none of these actions are specified, the output of bowtie2 is printed to a log (the *Log* action can also be specified by the user). All actions are compatible with each other, for example, the user can filter-out the reference genome while extracting the mapped reads to a separate file and generating the coverage vector in the same run.

The list of reference Vivonatev accepts as input must have a specific format. The file is organized as a table, each line corresponding to one reference, with three space-delimited columns: the first item is the name of the reference (usually the latin name of the species, using an underscore to join the two parts of the name); the second item is the location of the fasta file containing the genetic

sequence of the reference genome (either local path or absolute path); the last item is the length of the genome. The last two items are only used by the *coverage* function, therefore, if only extracting or filtering the genome, linking to */dev/null* and 0 would not be harmful to the functioning of the program. Vivonatev also take as (optional) input a directory, containing all the bowtie2 indexes of the references. For example, if the bowtie2 index files for *Escherichia_coli* are placed in *~/genomes/indexes/*, the user must specify *~/genomes/indexes/* as the genome directory. By default, Vivonatev searches for the references in the current working directory.

Vivonatev is available freely under the GNU GPL License. More features are under development. The current version is available at its github repository: http://github.com/jrtex/vivonatev.

5 A Case Study: FAIRE (Formaldehyde-Assisted Isolation of Regulatory Elements) Analysis of Open Chromatin in KSHV

5.1 Molecular Biology

One of the most effective means to discover transcriptional regulatory elements is by identification of nucleosome-depleted regions, also called "open chromatin." Historically this has been achieved by exploiting regional hypersensitivity to nucleases, such as DNase I. Recently, in the laboratory of J. Lieb, at the University of North Carolina, was developed a new methodology to detect regions of open chromatin, termed Formaldehyde-Assisted Isolation of Regulatory Elements (FAIRE) [48, 49]. It is based on the different efficiency of cross-linking between DNA nucleosome-depleted and sequence-specific DNA-binding proteins; specifically, nucleosome-depleted regions are much less efficiently cross-linked to proteins. Cells or tissues are briefly cross-linked with formaldehyde, then lysed and sonicated. Sheared chromatin is subjected to phenol/chloroform extraction: during this step, DNA cross-linked with proteins is trapped in the organic phase while DNA that is nucleosome-depleted is partitioned into the aqueous phase.

5.2 Next Gen Sequencing Set-Up

First, the transcriptional regulatory elements are purified and used to create libraries, which are then sequenced as described above. Both paired and unpaired libraries can be used.

5.3 Data Cleaning and Deposition into Short Reads Archive (SRA)

We used FAIRE to investigate chromatin organization of KSHV during latency, in particular to identify regions of open chromatin [3]. In this part of our review we will go into further details about the bioinformatics analysis of this data set. The data is available from the short reads archive (SRA) at GenBank under accession number: GSE50581.

5.4 Bioinformatics: MACS2 and CLC Genomic Workbench

We performed FAIRE-seq on the KSHV-infected PEL cell lines BCBL1 and aligned the resulting sequence reads to the KSHV reference genome (NC_009333). We used two different programs. First, we used the open source program MACS2 [50] to derive statistically significant nucleosome depletion (FAIRE peaks) at single-nucleotide resolution. Regions of increased coverage density correspond to regions of KSHV open chromatin. Second, we used the commercial program CLC Genomics Workbench (Qiagen Inc.), which has one of the best interfaces for this type of analysis. An alternative to CLC is the recently released Genious software. FAIRE enrichment identified upstream sequences of the constitutively expressed open reading (orf) frame for LANA (Fig. 1), at the constitutively active LANA promoter [51, 52].

5.5 Bioinformatics: Statistical Analysis in R

To expand our observations and to determine if there were differences in chromatin organization in different cell lines, we performed FAIRE-seq on multiple latently cancer cell lines: three B-cells lines (BC1, BCBL1, and BJAB carrying latent KSHV) and two endothelial cell lines (HUVEC carrying latent KSHV and L1-TIVE, also carrying the complete KSHV genome as an extrachromosomal plasmid). We found that regions of open chromatin are conserved across all latent KSHV-infected endothelial and B-cells (Fig. 2) and that the majority of viral promoters and viral genes are populated by closed chromatin and thus inaccessible to transcription factors and RNA polymerase II. The approach used to generate Fig. 2 takes the aligned reads, the

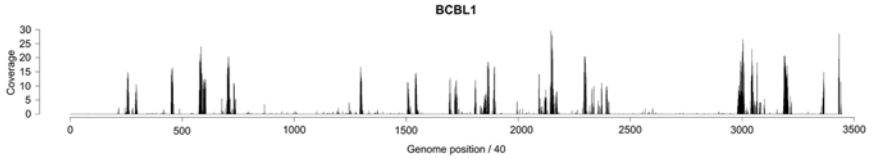


Fig. 1 FAIRE-seq analysis of PEL (BCBL1). Read coverage data for FAIRE across the KSHV genome (BCBL1) is shown on the vertical axis. Genome position/40 is indicated on *bottom*, i.e., a sliding window of 40 is used to aggregate the signal. This is using the traditional plot function in R (see Appendix 2)

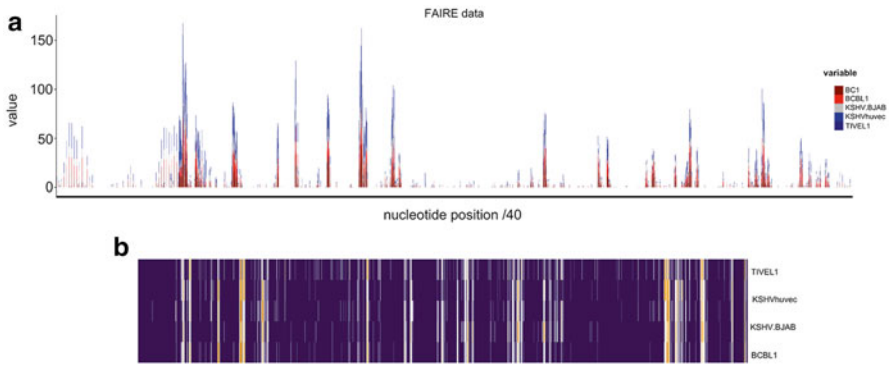


Fig. 2 (**a**) Stacked graph of average coverage across five cells lines. The sum of relative coverages is shown on the vertical and genome position/40 on the horizontal axis, i.e., a sliding window of 40 is used to aggregate the signal. Different cell lines are indicated by different colors. This uses the ggplot function in R (see Appendix 1). (**b**) Heat map of normalized coverage counts across five cell lines is shown. *Darker hues* indicate nucleosome-covered regions; the *lighter colors* regions of nucleosome depletion. The cell lines are indicated as row labels; the columns (*x*-axis) indicate nucleotide position/40

so-called "coverage vectors" and seeks to identify patterns of similarity of coverage. Since all reads are aligned to the same target sequence the nucleotide positions form a consistent base, the *x*-axis, where as the samples constitute the *y*-axis in a x * y matrix. We used CLC genomic workbench for the alignment and exported the coverage at each nucleotide position as an excel file. Interestingly excel was able to work with this amount of information, i.e., a spreadsheet with ~120,000 rows. In hindsight, we would have used .cvs or .txt file and import directly into R. All in all, the input file had 127,696 rows and 6 columns. The reader should be aware of GenBank changing nomenclature and individual nucleotide sequences as well as nucleotide numbers of its reference genomes intermittently as these are manually curated and updated. Nucleotide positions recorded in earlier work do not necessarily correspond to nucleotide positions in the current reference genome release. If the goal of the project is to incorporate both existing and novel NGS information, it is preferable to start from raw reads wherever these are available through the SRA archives.

Appendix 2 presents the R code that was used to produce Fig. 2. Of note, for final publication we applied a nonlinear adjustment of color values using Adobe Photoshop. The R program requires a number of standard libraries and defines a function for reading in one or more excel files, which are the merged using rbind. Note that the nucleotide position is recorded both as numeric value and as ordered factor for the purpose of visualization.

Key to further analysis is appropriate normalization. DNAseq, ChIPSeq, and RNAseq data follow a Poisson distribution, not a normal distribution, as slide array data [53]. Few statistics and few visualization methods use this type of data. Before and after each normalization step we assessed the distribution using the R functions `fitdistr (rain.melt$value, "Poisson")` and `fitdistr (rain.melt$value, "normal")`, as well as graphically. First, we took the cube root of each count data point. As an alternative, the Ascombe transformation may be used. Second, we centered

by median for each experiment and then divided by the IQR/1.349. This adjusts for the different total number of aligned hits and sequencing reads in each experiment. It is commonly called normalization or blocking by biological replicate. Using IQR/1.349 and median provides a more robust measure than mean and standard deviation. Lastly, all negative values were replaced with "0", to establish a biologically significant floor and to eliminate low-level noise. To reduce complexity further and to add visualization, we used a 40 nucleotide sliding window with rollapply from the "zoo" package: `m2 <- rollapply(rain[,1], width = 40, FUN = mean, by = 40).` Lastly, we coerced ggplot into reproducing the coverage diagram.

Acknowledgements This work was supported by Public Health Service grants CA019014 from the National Cancer Institute and AI107810 from the National Institute of Allergy and Infectious Diseases.

Appendix 1: Using Bowtie2 to Subtract Out Human Genome Sequences

Bowtie2 is a great aligner. Like all great programs it is available from the command line, which is both a curse and a blessing. So how do we get started? Let us take apart some sample code (Please make sure to remove the line breaks in any command line as unix does not recognize them).

```
dirkdittmer$ bowtie2 -U ../neisseria_gonorrhoeae/Raw_Reads/140728_
UNC14-SN744_0473_BHAFV9ADXX/AD-GC7_14P_CGTACG_L002_R2_001.fastq -x
hg19 -k 2 -p 8 -5 10 -3 10 --very-fast --end-to-end --un AD-Gc7-14p2.
fa --met-stderr -S output.sam
```

I am the user `dirkdittmer$`. We start in the directory where the bowtie2 target libraries are. More about how to create those is given below. Yet our input files are in a different library. So we are thrown into UNIX path name mudd. If you do not know this, now is a good time to review.

Define the Target Library

Step 1 is to define the target library. The target library is easily defined in our example as such: `-x hg19` Of note, there are many files in the folder hg19 that were created by bowtie. Hg19 refers to the human genome build Chr38. It was downloaded preformatted from Illumina's iGenomes collection at http://support.illumina.com/sequencing/sequencing_software/igenome.html. After download the directory looks like this:

```
dirkdittmer$ ls -ltotal 7950496-rw-r--r--@ 1 dirkdittmer  wheel  960018873 May
2  2012 hg19.1.bt2-rw-r--r--@ 1 dirkdittmer  wheel  716863572 May  2  2012
```

```
hg19.2.bt2-rw-r--r--@ 1 dirkdittmer wheel   3833 May 2 2012 hg19.3.bt2-
rw-r--r--@ 1 dirkdittmer  wheel  716863565 May  2  2012 hg19.4.bt2-rw-r--r--@ 1
dirkdittmer  wheel  960018873 May  2  2012 hg19.rev.1.bt2-rw-r--r--@ 1
dirkdittmer  wheel  716863572 May  2  2012 hg19.rev.2.bt2-rwxr-xr-x@ 1
dirkdittmer  wheel       3189 May  2  2012 make_hg19.sh
```

The important thing to remember is that in order to define the target library/database the bowtie2 tries to match the name before any dot, i.e., `-x hg19`. This is the Index filename prefix (minus the trailing .X.bt2). Note: you need to specify the full or relative path. We can use the abbreviated form, because we start the bowtie2 command from within the target library folder.

Define the Reads to Be Aligned

Because we start bowtie2 from within the library folder, also called index folder, we need to specify the complete path for the reads we want to align. In our case they are in a different folder and the pathname reads as such:

```
-U../neisseria_gonorrhoeae/Raw_Reads/140728_UNC14-SN744_0473_
BHAFV9ADXX/AD-GC7_14P_CGTACG_L002_R2_001.fastq
```

This is the first argument and is required. It can be one file or multiple files separated by comma. The -U flag precedes the input sequence(s) if the input is one or more files containing unpaired reads. For paired reads the first input file needs to be preceded by -1 and the second by -2 as in the following example:

```
dirkdittmer$ bowtie2 -1 ../neisseria_gonorrhoeae/Raw_Reads/140728_
UNC14-SN744_0473_BHAFV9ADXX/AD-GC7_14P_CGTACG_L002_R1_001.fastq
-2    ../neisseria_gonorrhoeae/Raw_Reads/140728_UNC14-SN744_0473_
BHAFV9ADXX/AD-GC7_14P_CGTACG_L002_R2_001.fastq -x hg19 -k 1 -p 8
-5 10 -3 10 --very-fast --end-to-end --un AD-Gc7-14.fa --met-
stderr -S output.sam
```

Setting Flags/Parameters

What other flags do we need to or want to set? Typing bowtie2 -h will pull up the manual in the terminal or you can go to their truly excellent website at http://bowtie-bio.sourceforge.net/bowtie2/manual.shtml. In our example we set: "-k 2". This flag defines how many matches bowtie will search for, i.e., the first or the best *N* alignments. An alternative to setting **-k** is setting **-a**. Setting **-a** will report all

alignments with no upper limit. “-p 8” This defines the number of cores the program will use. “-5 10 -3 10” This defines the number of nucleotides to be trimmed before alignment --very-fast We wanted it fast and as a result slightly less precise. “--end-to-end” This forces the reads to be aligned completely; one end to the other. Not setting this option will align the best central part of the read and allow any number of prefacing or trailing misalignments. “--un AS-Gc7-14p2.fa” This will output unaligned reads as a fastafile to “AS-Gc7–14p2.fa” for further processing. In our examples this was the goal of running bowtie. “-- met -stderr” This outputs the metrices tp stderr, i.e., in most cases the terminal. “-S output.sam.” This is the output file in sam-tools format

Further Examples

This is another example, where we are primarily interested in obtaining unaligned reads, i.e., deplete or filter out reads that align to the human genome. Note the code is using 12 cores -p 12.

```
dittmerrg5:hg19 dirkdittmer$ bowtie2 -U ../neisseria_gonorrhoeae/
Raw_Reads/140728_UNC14-SN744_0473_BHAFV9ADXX/AD-GC7_14P_CGTACG_
L002_R1_001.fastq -x ../hg19/hg19 -k 1 -p 12 -5 10 -3 5 --very-fast
--end-to-end --un AD-Gc7-14.fa --met-stderr -S output.sam
```

The result looks as follows in sterr, i.e., the terminal using the following trimming parameters -5 10 -3 5. `The input was 105543931 reads; of these 105543931 (100.00 %) were unpaired; of these: 13752261 (13.03 %) aligned 0 times 91791670 (86.97 %) aligned exactly 1 time 0 (0.00 %) aligned >1 times with an 86.97 % overall alignment rate`. Note that we do not count reads that aligned >1 times because of -k 1. Let us see how this changes if we trim less from the ends -5 5 -3 5. `105543931 reads; of these: 105543931 (100.00 %) were unpaired; of these: 14459602 (13.70 %) aligned 0 times, 91084329 (86.30 %) aligned exactly 1 time 0 (0.00 %) aligned >1 times 86.30 % overall alignment rate`. What if we do not trim at all, i.e., set -5 0 -3 0, which is also the default setting. This certainly speeds up the run, and the results are: `105543931 reads; of these: 105543931 (100.00 %) were unpaired; of these: 15013710 (14.23 %) aligned 0 times, 90530221 (85.77 %) aligned exactly 1 time, 0 (0.00 %) aligned >1 times, 85.77 % overall alignment rate, Time searching: 00:36:02, Overall time: 00:36:02`. So a little less, but not significantly. To get the time estimate set the time flag --time.

Appendix 2: R Code to Generate Figures

```
R version 3.0.2 (2013-09-25)
Platform: x86_64-apple-darwin10.8.0 (64-bit)

# ------------------------------------------------------------------------------
# STEP 0.a: Read in the libraries.
# ------------------------------------------------------------------------------
library(gdata)
library(reshape2)
library(MASS)
library(RColorBrewer)
library(heatmap.plus)
library(ggplot2)
library(Hmisc)
library(zoo)
# ------------------------------------------------------------------------------
# STEP 0.c: Define function(s)
# ------------------------------------------------------------------------------
# read in multiple spreadsheets each with primers and one or
# more experiments. The experiments are in columns
#
read.rain <- function(x)
{
      filnamelistlength <- length(x)
      mir <- read.delim(x[1], header = TRUE, sep = "\t")
      mir <- na.omit(mir)
      mir <- drop.levels(mir)
      mir$filename <- factor(x[1])
      rain <- mir
      print(nrow(rain))
      if (filnamelistlength< 2) return (rain)
      for (i in 2:filnamelistlength)
      {
      mir <- read.delim(x[i], header = TRUE, sep = "\t")
      mir <- na.omit(mir)
      mir <- drop.levels(mir)
      mir$filename <- factor(x[i])
      rain <- rbind(rain,mir)
      print(nrow(rain))
      }
      return (rain)
}
# ------------------------------------------------------------------------------
# STEP 1 Read in the data
# ------------------------------------------------------------------------------
filenames <- c("toR.txt")
rain <- read.rain(filenames)
table(rain$filename)
# ---------------------------------------------------------
# Safe a copy of the data
#
write.table(rain,file = "CompleteRawDataSet.txt", sep = "\t")
# ---------------------------------------------------------
# Clean the data
#
head(rain)
colnames(rain)
#drop out filename column
rain <- rain[,-ncol(rain)]
rain$PositionFactor <- factor(rain$Position, ordered = TRUE)
# ------------------------------------------------------------------------------
# STEP 2: Data transformation
# ------------------------------------------------------------------------------
# Evaluate Poisson nature of the data
```

(continued)

(continued)

```
#
rain.melt <- melt(rain[,2:6])
readP <- fitdistr(rain.melt$value, "Poisson" )
readP$estimate
readP$sd
readP$vocv
readP$loglik
# ------------------------------------------------------
# Before normalization
#
fitdistr(rain.melt$value, "Poisson" )
fitdistr(rain.melt$value, "normal" )
# ------------------------------------------------------
# After normalization
#
rain.melt$value <- (rain.melt$value +1)^(1/3) - median(((rain.melt$value
+1)^(1/3)), na.rm = FALSE)
rain.melt$value <- rain.melt$value/ (IQR((rain.melt$value +1)^(1/3)) /1.349)
fitdistr(rain.melt$value, "Poisson" )
fitdistr(rain.melt$value, "normal" )
# ------------------------------------------------------
# Normalization by experiment
#
summary(rain)
rain$BC1 <- (rain$BC1 +1)^(1/3) - median(((rain$BC1 +1)^(1/3)), na.rm = FALSE)
rain$BC1 <- rain$BC1/ (IQR((rain$BC1 +1)^(1/3)) /1.349)
rain$BCBL1 <- (rain$BCBL1 +1)^(1/3) - median(((rain$BCBL1 +1)^(1/3)), na.rm =
FALSE)
rain$BCBL1 <- rain$BCBL1/ (IQR((rain$BCBL1 +1)^(1/3)) /1.349)
rain$KSHV.BJAB <- (rain$KSHV.BJAB  +1)^(1/3) - median(((rain$KSHV.BJAB
+1)^(1/3)), na.rm = FALSE)
rain$KSHV.BJAB  <- rain$KSHV.BJAB / (IQR((rain$KSHV.BJAB  +1)^(1/3)) /1.349)
rain$KSHVhuvec <- (rain$KSHVhuvec +1)^(1/3) - median(((rain$KSHVhuvec
+1)^(1/3)), na.rm = FALSE)
rain$KSHVhuvec <- rain$KSHVhuvec/ (IQR((rain$KSHVhuvec +1)^(1/3)) /1.349)
rain$TIVEL1 <- (rain$TIVEL1 +1)^(1/3) - median(((rain$TIVEL1 +1)^(1/3)), na.rm =
FALSE)
rain$TIVEL1 <- rain$TIVEL1/ (IQR((rain$TIVEL1 +1)^(1/3)) /1.349)
rain <- as.data.frame(rain[,2:6])
# ------------------------------------------------------
# Introduce floor of 0 counts
#
rain[rain$BC1 < 0,]           <-  0
rain[rain$BCBL1 < 0,]         <-  0
rain[rain$KSHV.BJAB < 0,]     <-  0
rain[rain$KSHVhuvec < 0,]     <-  0
rain[rain$TIVEL1 < 0,]        <-  0
# ------------------------------------------------------
# write out normalized to to file
#
write.table(rain,file = "normalized.txt", sep = "\t")
# ------------------------------------------------------------------------------
# STEP 3: Data analysis
# ------------------------------------------------------------------------------
# Sliding window: first column
m2 <- rollapply(rain[,1], width = 40, FUN = mean, by = 40)
m2 <- data.frame(m2)
# ------------------------------------------------------
# remaining columns
for (i in 2:ncol(rain))
      {
      m2[,i] <- rollapply(rain[,i], width = 40, FUN = mean, by = 40)
      }
colnames(m2) <- colnames(rain)
rownames(m2) <- as.numeric(rownames(m2))*40
# ------------------------------------------------------
# write out to file
#
write.table(m2,file = "reducedData.txt", sep = "\t")
# ------------------------------------------------------
# Distribution visualization
```

(continued)

```
#
m2$position <- factor(rownames(m2))
m2.melt <- melt(m2)
png(file = "qqPlot_of40ntwindow.png", units = "in", width = 10, height = 8, res
= 300)
{
par(mfrow = c(1,1))
par(cex = 1.5)
par(bty = "o")
qqnorm(m2.melt$value,
pch = 19,
cex = 0.5,
col = "darkblue",
main = "100 bp window")
qqline(m2.melt$value, col = "darkred", lwd = 2, lty = 3)
histSpike(m2.melt$value, add=TRUE, col = "darkgray", frac = 0.3, lwd = 3, side
=2)
}
dev.off()
# ---------------------------------------------------------
# Conversion to all numeric matrix for heatmap
#
m2 <- m2[,-ncol(m2)]
try2m <- as.matrix(m2)
try2m <- t(try2m)
png(file = "heatmapAllScalebyRow.png", units = "in", width = 32, height = 6, res
= 300)
{
brewer.palette <- colorRampPalette(rev(brewer.pal(11, "PuOr")),space = "rgb",
bias = 3, interpolate = "spline")
heatmap.plus(try2m, na.rm = T, scale = "row", col = brewer.palette (55),
        hclustfun=function(m) hclust(m, method="ward"),
        distfun = function(x) dist(x, method = "manhattan"),
        margins = c(5,10), Colv = NA, Rowv = NA, cexCol = 0.5)
}
dev.off()
# ---------------------------------------------------------
# selection of peak locations
m2$sum <- rowMeans(m2)
png(file = "sum.png", units = "in", width = 32, height = 6, res = 600)
{
par(bty="n")
plot(m2$sum,  type = "h")
}
dev.off()
rownames(m2[m2$sum > 10,])
# ---------------------------------------------------------------------------------
# STEP 4: Data visualization
# ---------------------------------------------------------------------------------
# as points
zp1 <- ggplot(m2.melt, aes(x = factor(position, ordered = TRUE), y = value,
color = variable))
zp1 <- zp1 + geom_point(aes = 1)
zp1 <- zp1 + scale_color_manual(values =
c("red","green","blue","gray","orange"))
zp1 <- zp1 + theme_bw()
zp1 <- zp1 + theme(legend.position = "right")  + theme(strip.background =
element_rect(colour = "white"))
zp1 <- zp1 + theme(axis.text.x = element_text(size = 10, angle = 90))  +
theme(axis.text.y = element_text(size = 10, angle = 0))  + theme(axis.title.y =
element_text(size = 12, angle = 0))  + theme(plot.title = element_text(size =
16))  + theme(axis.title.x = element_text(size = 12))  + theme(strip.text.x =
element_text(size = 14))
print(zp1)
# ---------------------------------------------------------
# as lines
zp1 <- ggplot(m2.melt, aes(x = as.numeric(position), y = value, color =
variable))
zp1 <- zp1 + geom_line(aes = 1)
zp1 <- zp1 + scale_color_manual(values =
c("red","green","blue","gray","orange"))

zp1 <- zp1 + theme_bw()
zp1 <- zp1 + theme(legend.position = "right")  + theme(strip.background =
element_rect(colour = "white"))
zp1 <- zp1 + theme(axis.text.x = element_text(size = 10, angle = 90))  +
theme(axis.text.y = element_text(size = 10, angle = 0))  + theme(axis.title.y =
element_text(size = 12, angle = 0))  + theme(plot.title = element_text(size =
16))  + theme(axis.title.x = element_text(size = 12))  + theme(strip.text.x =
element_text(size = 14))
print(zp1)
```

References

1. https://github.com (2014). 2014.
2. http://www.ncbi.nlm.nih.gov/sra (2014). 2014.
3. Hilton IB, Simon JM, Lieb JD, Davis IJ, Damania B, Dittmer DP. The open chromatin landscape of Kaposi's sarcoma-associated herpesvirus. J Virol. 2013;87:11831–42. doi:10.1128/JVI.01685-13.
4. http://monographs.iarc.fr/ENG/Monographs/vol100B/ (2014). Accessed 2014.
5. Smith MA, Canfell K, Brotherton JM, Lew JB, Barnabas RV. The predicted impact of vaccination on human papillomavirus infections in Australia. Int J Cancer. 2008;123:1854–63. doi:10.1002/ijc.23633.
6. Afdhal N, Zeuzem S, Kwo P, Chojkier M, Gitlin N, Puoti M, Romero-Gomez M, Zarski JP, Agarwal K, Buggisch P, Foster GR, Brau N, Buti M, Jacobson IM, Subramanian GM, Ding X, Mo H, Yang JC, Pang PS, Symonds WT, McHutchison JG, Muir AJ, Mangia A, Marcellin P, ION Investigators. Ledipasvir and sofosbuvir for untreated HCV genotype 1 infection. N Engl J Med. 2014;370:1889–98. doi:10.1056/NEJMoa1402454.
7. http://www.nobelprize.org/nobel_prizes/medicine/laureates/2008/hausen-lecture.html (2014). 2014.
8. Ribeiro RM, Li H, Wang S, Stoddard MB, Learn GH, Korber BT, Bhattacharya T, Guedj J, Parrish EH, Hahn BH, Shaw GM, Perelson AS. Quantifying the diversification of hepatitis C virus (HCV) during primary infection: estimates of the in vivo mutation rate. PLoS Pathog. 2012;8:e1002881. doi:10.1371/journal.ppat.1002881.
9. Fiers W, Contreras R, Haegemann G, Rogiers R, Van de Voorde A, Van Heuverswyn H, Van Herreweghe J, Volckaert G, Ysebaert M. Complete nucleotide sequence of SV40 DNA. Nature. 1978;273:113–20.
10. Lebowitz P, Kelly Jr TJ, Nathans D, Lee TN, Lewis Jr AM. A colinear map relating the simian virus 40 (SV40) DNA segments of six adenovirus-SV40 hybrids to the DNA fragments produced by restriction endonuclease cleavage of SV40 DNA. Proc Natl Acad Sci U S A. 1974;71:441–5.
11. Enquist LW, Editors of the Journal of Virology. Virology in the 21st century. J Virol. 2009;83:5296–308. doi:10.1128/JVI.00151-09.
12. Dartmann K, Schwarz E, Gissmann L, zur Hausen H. The nucleotide sequence and genome organization of human papilloma virus type 11. Virology. 1986;151:124–30.
13. Kumar S, Blaxter ML. Comparing de novo assemblers for 454 transcriptome data. BMC Genomics. 2010;11:571. doi:10.1186/1471-2164-11-571.
14. Nederbragt AJ. On the middle ground between open source and commercial software—the case of the Newbler program. Genome Biol. 2014;15:113.
15. Vazquez-Castellanos JF, Garcia-Lopez R, Perez-Brocal V, Pignatelli M, Moya A. Comparison of different assembly and annotation tools on analysis of simulated viral metagenomic communities in the gut. BMC Genomics. 2014;15:37. doi:10.1186/1471-2164-15-37.
16. http://genomea.asm.org (2014).
17. Cole CG, McCann OT, Collins JE, Oliver K, Willey D, Gribble SM, Yang F, McLaren K, Rogers J, Ning Z, Beare DM, Dunham I. Finishing the finished human chromosome 22 sequence. Genome Biol. 2008;9:R78. doi:10.1186/gb-2008-9-5-r78.
18. Zody MC, Garber M, Adams DJ, Sharpe T, Harrow J, Lupski JR, Nicholson C, Searle SM, Wilming L, Young SK, Abouelleil A, Allen NR, Bi W, Bloom T, Borowsky ML, Bugalter BE, Butler J, Chang JL, Chen CK, Cook A, Corum B, Cuomo CA, de Jong PJ, DeCaprio D, Dewar K, FitzGerald M, Gilbert J, Gibson R, Gnerre S, Goldstein S, Grafham DV, Grocock R, Hafez N, Hagopian DS, Hart E, Norman CH, Humphray S, Jaffe DB, Jones M, Kamal M, Khodiyar VK, LaButti K, Laird G, Lehoczky J, Liu X, Lokyitsang T, Loveland J, Lui A, Macdonald P, Major JE, Matthews L, Mauceli E, McCarroll SA, Mihalev AH, Mudge J, Nguyen C, Nicol R, O'Leary SB, Osoegawa K, Schwartz DC, Shaw-Smith C, Stankiewicz P, Steward C, Swarbreck D, Venkataraman V, Whittaker CA, Yang X, Zimmer AR, Bradley A, Hubbard T, Birren BW,

Rogers J, Lander ES, Nusbaum C. DNA sequence of human chromosome 17 and analysis of rearrangement in the human lineage. Nature. 2006;440:1045–9. doi:10.1038/nature04689.
19. Guo L, Yan Q, Yang S, Wang C, Chen S, Yang X, Hou R, Quan Z, Hao Z. Full genome sequence of giant panda rotavirus strain CH-1. Genome Announc. 2013. doi:10.1128/genomeA.00241-12.
20. Li R, Fan W, Tian G, Zhu H, He L, Cai J, Huang Q, Cai Q, Li B, Bai Y, Zhang Z, Zhang Y, Wang W, Li J, Wei F, Li H, Jian M, Li J, Zhang Z, Nielsen R, Li D, Gu W, Yang Z, Xuan Z, Ryder OA, Leung FC, Zhou Y, Cao J, Sun X, Fu Y, Fang X, Guo X, Wang B, Hou R, Shen F, Mu B, Ni P, Lin R, Qian W, Wang G, Yu C, Nie W, Wang J, Wu Z, Liang H, Min J, Wu Q, Cheng S, Ruan J, Wang M, Shi Z, Wen M, Liu B, Ren X, Zheng H, Dong D, Cook K, Shan G, Zhang H, Kosiol C, Xie X, Lu Z, Zheng H, Li Y, Steiner CC, Lam TT, Lin S, Zhang Q, Li G, Tian J, Gong T, Liu H, Zhang D, Fang L, Ye C, Zhang J, Hu W, Xu A, Ren Y, Zhang G, Bruford MW, Li Q, Ma L, Guo Y, An N, Hu Y, Zheng Y, Shi Y, Li Z, Liu Q, Chen Y, Zhao J, Qu N, Zhao S, Tian F, Wang X, Wang H, Xu L, Liu X, Vinar T, Wang Y, Lam TW, Yiu SM, Liu S, Zhang H, Li D, Huang Y, Wang X, Yang G, Jiang Z, Wang J, Qin N, Li L, Li J, Bolund L, Kristiansen K, Wong GK, Olson M, Zhang X, Li S, Yang H, Wang J, Wang J. The sequence and de novo assembly of the giant panda genome. Nature. 2010;463:311–7. doi:10.1038/nature08696.
21. Szpara ML, Parsons L, Enquist LW. Sequence variability in clinical and laboratory isolates of herpes simplex virus 1 reveals new mutations. J Virol. 2010;84:5303–13. doi:10.1128/JVI.00312-10.
22. Wen KW, Dittmer DP, Damania B. Disruption of LANA in rhesus rhadinovirus generates a highly lytic recombinant virus. J Virol. 2009;83:9786–802. doi:10.1128/JVI.00704-09.
23. Dunsworth-Browne M, Schell RE, Berk AJ. Adenovirus terminal protein protects single stranded DNA from digestion by a cellular exonuclease. Nucleic Acids Res. 1980;8:543–54.
24. Lagunoff M, Ganem D. The structure and coding organization of the genomic termini of Kaposi's sarcoma-associated herpesvirus. Virology. 1997;236:147–54.
25. Mocarski ES, Roizman B. Structure and role of the herpes simplex virus DNA termini in inversion, circularization and generation of virion DNA. Cell. 1982;31:89–97.
26. Pearson WR. Flexible sequence similarity searching with the FASTA3 program package. Methods Mol Biol. 2000;132:185–219.
27. http://hannonlab.cshl.edu/fastx_toolkit/ (2014). 2014.
28. Hatem A, Bozdag D, Toland AE, Catalyurek UV. Benchmarking short sequence mapping tools. BMC Bioinformatics. 2013;14:184. doi:10.1186/1471-2105-14-184.
29. Lin Z, Farooqui A, Li G, Wong GK, Mason AL, Banner D, Kelvin AA, Kelvin DJ, Leon AJ. Next-generation sequencing and bioinformatic approaches to detect and analyze influenza virus in ferrets. J Infect Dev Ctries. 2014;8:498–509. doi:10.3855/jidc.3749.
30. Zerbino DR. Using the Velvet de novo assembler for short-read sequencing technologies. Curr Protoc Bioinformatics. 2010;Chapter 11:Unit 11.15. doi:10.1002/0471250953.bi1105s31.
31. Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics. 2009;25:1754–60. doi:10.1093/bioinformatics/btp324.
32. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R, Genome Project Data Processing Subgroup. The sequence alignment/map format and SAMtools. Bioinformatics. 2009;25:2078–9. doi:10.1093/bioinformatics/btp352.
33. McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, Garimella K, Altshuler D, Gabriel S, Daly M, DePristo MA. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res. 2010;20:1297–303. doi:10.1101/gr.107524.110.
34. Yu X, Sun S. Comparing a few SNP calling algorithms using low-coverage sequencing data. BMC Bioinformatics. 2013;14:274. doi:10.1186/1471-2105-14-274.
35. Pirooznia M, Kramer M, Parla J, Goes FS, Potash JB, McCombie WR, Zandi PP. Validation and assessment of variant calling pipelines for next-generation sequencing. Hum Genomics. 2014;8:14. doi:10.1186/1479-7364-8-14.

36. http://www.bioconductor.org (2014). 2014.
37. Manolio TA, Collins FS. The HapMap and genome-wide association studies in diagnosis and therapy. Annu Rev Med. 2009;60:443–56. doi:10.1146/annurev.med.60.061907.093117.
38. Cingolani P, Platts A, le Wang L, Coon M, Nguyen T, Wang L, Land SJ, Lu X, Ruden DM. A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster strain w1118; iso-2; iso-3. Fly. 2012;6:80–92. doi:10.4161/fly.19695.
39. Trapnell C, Pachter L, Salzberg SL. TopHat: discovering splice junctions with RNA-Seq. Bioinformatics. 2009;25:1105–11. doi:10.1093/bioinformatics/btp120.
40. Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, Batut P, Chaisson M, Gingeras TR. STAR: ultrafast universal RNA-seq aligner. Bioinformatics. 2013;29:15–21. doi:10.1093/bioinformatics/bts635.
41. http://www.geneious.com (2014). 2014.
42. Pollier J, Rombauts S, Goossens A. Analysis of RNA-Seq data with TopHat and Cufflinks for genome-wide expression analysis of jasmonate-treated plants and plant cultures. Methods Mol Biol. 2013;1011:305–15. doi:10.1007/978-1-62703-414-2_24.
43. Trapnell C, Williams BA, Pertea G, Mortazavi A, Kwan G, van Baren MJ, Salzberg SL, Wold BJ, Pachter L. Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. Nat Biotechnol. 2010;28:511–5. doi:10.1038/nbt.1621.
44. Zhang ZH, Jhaveri DJ, Marshall VM, Bauer DC, Edson J, Narayanan RK, Robinson GJ, Lundberg AE, Bartlett PF, Wray NR, Zhao QY. A comparative study of techniques for differential expression analysis on RNA-Seq data. PLoS One. 2014;9:e103207. doi:10.1371/journal.pone.0103207.
45. Tamburro KM, Yang D, Poisson J, Fedoriw Y, Roy D, Lucas A, Sin SH, Malouf N, Moylan V, Damania B, Moll S, van der Horst C, Dittmer DP. Vironome of Kaposi sarcoma associated herpesvirus-inflammatory cytokine syndrome in an AIDS patient reveals co-infection of human herpesvirus 8 and human herpesvirus 6A. Virology. 2012;433:220–5. doi:10.1016/j.virol.2012.08.014.
46. Feng H, Shuda M, Chang Y, Moore PS. Clonal integration of a polyomavirus in human Merkel cell carcinoma. Science. 2008;319:1096–100. doi:10.1126/science.1152586.
47. Schowalter RM, Pastrana DV, Pumphrey KA, Moyer AL, Buck CB. Merkel cell polyomavirus and two previously unknown polyomaviruses are chronically shed from human skin. Cell Host Microbe. 2010;7:509–15. doi:10.1016/j.chom.2010.05.006.
48. Simon JM, Giresi PG, Davis IJ, Lieb JD. Using formaldehyde-assisted isolation of regulatory elements (FAIRE) to isolate active regulatory DNA. Nat Protoc. 2012;7:256–67. doi:10.1038/nprot.2011.444.
49. Simon JM, Giresi PG, Davis IJ, Lieb JD. A detailed protocol for formaldehyde-assisted isolation of regulatory elements (FAIRE). Curr Protoc Mol Biol. 2013;Chapter 21:Unit21.26. doi:10.1002/0471142727.mb2126s102.
50. Yang Y, Fear J, Hu J, Haecker I, Zhou L, Renne R, Bloom D, McIntyre LM. Leveraging biological replicates to improve analysis in ChIP-seq experiments. Comput Struct Biotechnol J. 2014;9:e201401002. doi:10.5936/csbj.201401002.
51. Dittmer D, Lagunoff M, Renne R, Staskus K, Haase A, Ganem D. A cluster of latently expressed genes in Kaposi's sarcoma-associated herpesvirus. J Virol. 1998;72:8309–15.
52. Hilton IB, Dittmer DP. Quantitative analysis of the bidirectional viral G-protein-coupled receptor and lytic latency-associated nuclear antigen promoter of Kaposi's sarcoma-associated herpesvirus. J Virol. 2012;86:9683–95. doi:10.1128/JVI.00881-12.
53. Lee S, Chugh PE, Shen H, Eberle R, Dittmer DP. Poisson factor models with applications to non-normalized microRNA profiling. Bioinformatics. 2013;29:1105–11. doi:10.1093/bioinformatics/btt091.

Molecular Typing of Lung Adenocarcinoma on Cytological Samples in the Next-Generation Sequencing Era

Rocco Cappellesso, Ambrogio Fassina, Emilio Bria, Aldo Scarpa, and Matteo Fassan

Abstract Adenocarcinoma (AdC) is the most common subtype of lung cancer, the leading tumor worldwide for incidence and mortality. In the majority of cases, a diagnosis is achieved only in advanced inoperable disease on cytological material obtained from pleural effusion, bronchoalveolar lavage, brushing, or fine-needle aspiration. Current recommendations provide for AdC to be tested for molecular alterations for which are already available targeted agents and many others are in clinical trials. However, conventional sequencing lacks of the necessary sensitivity to detect such molecular alterations in the scant cytological material and produces too many false negative results. Moreover, the number of therapeutically impacting markers that will need to be assessed is expected to rapidly increase. Thus, the application of highly sensitive and multigene probing methods, such as those developed in the context of next-generation sequencing (NGS), has been recently introduced into clinical practice. NGS is able to detect and quantitate multiple gene alterations from limited amounts of DNA, thus improving the diagnostic and prognostic stratification of lung cancer patients, which is essential for personalized cancer therapy. This chapter yields the available data about NGS in this field.

R. Cappellesso, M.D. • A. Fassina, M.D. • M. Fassan, M.D., Ph.D. (✉)
Department of Medicine (DIMED), Surgical Pathology Unit,
University of Padua, Padua, Italy
e-mail: rocco.cappellesso@gmail.com; ambrogio.fassina@unipd.it;
matteo.fassan@gmail.com

E. Bria, M.D.
Department of Medicine, Medical Oncology, University of Verona, Verona, Italy
e-mail: emilio.bria@univr.it

A. Scarpa, M.D.
ARC-NET Research Center and Department of Pathology and Diagnostics,
University of Verona, Verona, Italy
e-mail: aldo.scarpa@univr.it

W. Wu, H. Choudhry (eds.), *Next Generation Sequencing in Cancer Research, Volume 2*, DOI 10.1007/978-3-319-15811-2_21

1 Introduction

Lung cancer is the leading tumor worldwide for incidence and mortality [1]. According to the World Health Classification, lung cancer histologically encompasses two major entities: the small cell lung carcinoma (SCLC) and the more common non-small-cell lung carcinoma (NSCLC) [2]. The latter represents a heterogeneous group of tumors and is subdivided in adenocarcinoma (AdC), squamous cell carcinoma, and large cell carcinoma [2]. Among these subtypes, AdC is the most frequent, may develop in nonsmoker individuals, particularly in women, and is usually located at the periphery of the lung [2].

This site of origin accounts for the common association of this tumor type with pleural effusion and, most of all, for the diagnostic delay. The large majority of AdCs have a poor prognosis because are diagnosed at an advanced inoperable stage. This is mainly due to the lack of symptoms at the beginning of the disease as well as of effective screening methods to date.

2 The Role of Cytology in the Diagnostic Workout of Lung Adenocarcinomas

Cytology is a quick, low-cost, minimally invasive, and repeatable analysis widely employed to achieve a diagnosis of AdC, mainly in patients with advanced diseases or with low performance status that are unable to undergo an open biopsy [3].

In the AdC setting, cytology encompasses different diagnostic applications varying from analysis of the exfoliated cells in fluids contained in the pleural cavities or in bronchoalveolar lavage (BAL) to examination of cells obtained from mass lesions by brushing or fine-needle aspiration (FNA).

In the first case, the excessive fluid accumulated into a pleural cavity is drawn out by thoracentesis and collected for the cytological analysis. If the pleural effusion is due to an AdC that infiltrates the mesothelial lining of the lung, the exfoliated tumor cells can be identified in the fluid. However, AdC cells are usually scant in this kind of specimen and may be much diluted since effusion may range from a volume of about 300 ml to more than 1 l, thus a preliminary cytocentrifugation is mandatory.

BAL usually provides a more adequate and selective specimen to be analyzed. The bronchoscope is introduced into a selected part of the lung where the lavage fluid (100–300 ml) is squirted out in the terminal bronchioles and then recollected for the analysis. During bronchoscopy, cytological samples may be also collected from visible lesions by brushing (including mainly, if not only, superficial tumor cells) and from centrally located deep nodules by ultrasound-guided FNA (i.e., trans-bronchial FNA).

Finally, computed tomography assists the correct targeting of peripheral tumors by fine needles through the thorax (i.e., trans-thoracic FNA). Compared to the previously presented techniques of cell collection, FNA usually provides the largest amount of tumor cells (Fig. 1). It has to be noted that in little or heterogeneous

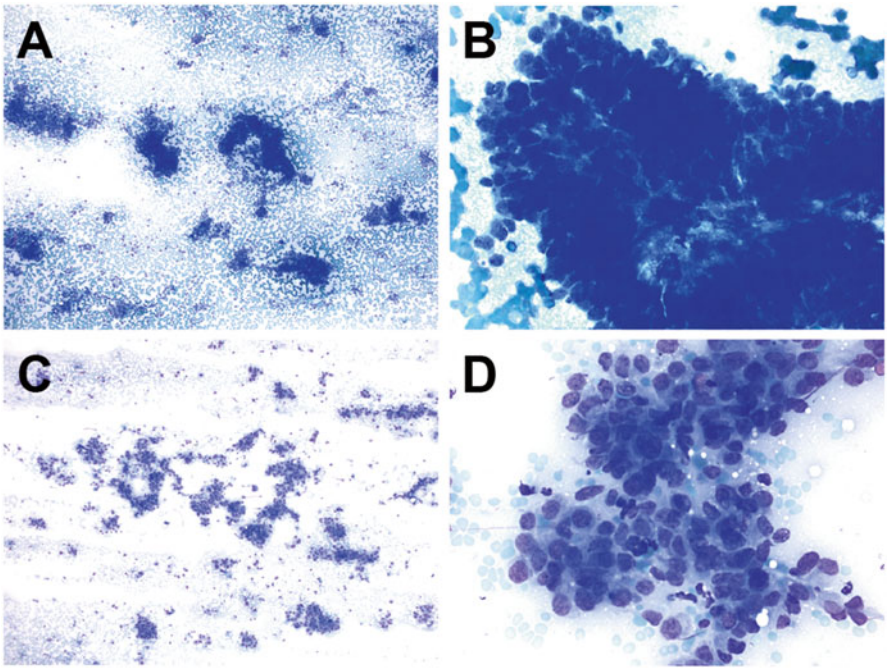


Fig. 1 Representative examples of trans-bronchial ultrasound-guided (**a** and **b**) and trans-thoracic computed tomography-guided (**c** and **d**) fine-needle aspirations of lung adenocarcinomas. May-Grünwald Giemsa stain, original magnifications 40× (**a** and **c**) and 200× (**b** and **d**)

tumors FNA cytology can be even more informative than a core biopsy, since fine needle movements allow to sample different neoplastic areas [4, 5]. Moreover, a rapid on-site evaluation (ROSE) of the aspirates allows to assess cytological material amount and adequacy (revealing if a second FNA pass is needed) and often to reach a cytological diagnosis [4, 6, 7].

The recent advent of AdC-targeting therapies has extremely revolutionized the diagnostic and therapeutic impact of cytology in lung cancer patients. The adequate pre-analytic management of the cytological specimens became of key importance for subsequent—today mandatory—molecular characterization of the tumors. Indeed, all the international recommendations for good practice on lung cancer clearly state the absolute need for an acceptable acquisition and preservation of any residual cytological material for molecular studies [8, 9]. As a result, residual aspirates and needle rinses are usually stock in preservatives or in cell-blocks for further analyses in current clinical practice [3, 10].

However, all these different cytological sampling methods share a common restriction: a low amount of collected cells. Indeed, if a morphological diagnosis of AdC can be based on the alterations of few neoplastic cells, these could not be sufficient to achieve also a comprehensive molecular profiling by using conventional methods. Of note, the number of the molecular determinations to be tested is destined to grow, as explained below.

3 Lung Adenocarcinoma Molecular Scenario and New Targeted Therapies

Recent efforts have been made in order to discovery new biomarkers suitable for the development of novel effective therapies for lung AdC. Such approaches led to the introduction into clinical practice of AdC-specific targeting therapeutics (Table 1).

Among the others, the most important are the tyrosine kinase inhibitors (TKIs) directed against epidermal growth factor receptor (EGFR). Indeed, about 15–20 % of AdC cases harbors mutations involving the tyrosine kinase domain of this receptor [11–14]. Such molecular alterations are represented in approximately 90 % of cases by in-frame deletions in exon 19 or missense mutations in exon 21, such as the common leucine to arginine substitution at codon 858 (L858R), both resulting in a constitutive activation of the receptor [15]. More rarely, activating mutation affects *EGFR* exon 18 [15]. All these molecular alterations confer to AdC a significant sensitivity to EGFR-TKI. Indeed, patients with advanced disease treated with such therapy showed an increase of the overall survival of about 6 months and of the progression free survival of more than 1 month [16].

However, during the treatment can arise drug resistance, usually determined by the occurrence of a second mutation in *EGFR* exon 20, mainly a threonine to methionine substitution at codon 790 (T790M), or, more rarely, due to a mutation in the downstream effector V-Ki-ras2 Kirsten rat sarcoma viral oncogene homologue (*KRAS*) [15].

More recently, the echinoderm microtubule protein like-4/anaplastic lymphoma kinase (*EML4-ALK*) fusion gene has been detected in about 7 % of patients with EGFR-TKI resistance, and thus, crizotinib was added to the therapy of these patients [17–19]. As a result, the current recommendations provide for AdC to be firstly screened for *EGFR* mutations and then for the *EML4-ALK* rearrangement (Fig. 2) [20].

Recent whole-exome and whole-genome sequencing studies revealed new molecules and mechanisms that are involved in AdC (Fig. 3), some of which seems to be therapeutically targetable [12, 21, 22]. In fact, clinical trials are ongoing in subgroups of patients harboring activating mutations of v-Raf murine sarcoma viral oncogene homolog B (*BRAF*), phosphatidylinositol-4,5-bisphosphate 3-kinase catalytic subunit alpha (*PIK3CA*), or *KRAS* [23–25]. Thus, the number of predictive biomarkers to be assessed for novel targeted agents entering into the clinical practice is expected to rapidly increase.

Table 1 Targeted agents currently approved by Food and Drug Administration (FDA) for lung adenocarcinoma treatment

Drug	Target	Mechanism of action
Erlotinib	EGFR with activating mutation	EGFR-TKI
Gefitinib	EGFR with activating mutation	EGFR-TKI
Afatinib	EGFR with activating mutation	EGFR-TKI
Crizotinib	ALK rearrangement	ALK-TKI
Ceritinib	ALK rearrangement	ALK-TKI

ALK anaplastic lymphoma, *EGFR* epidermal growth factor receptor, *TKI* tyrosine kinase inhibitor

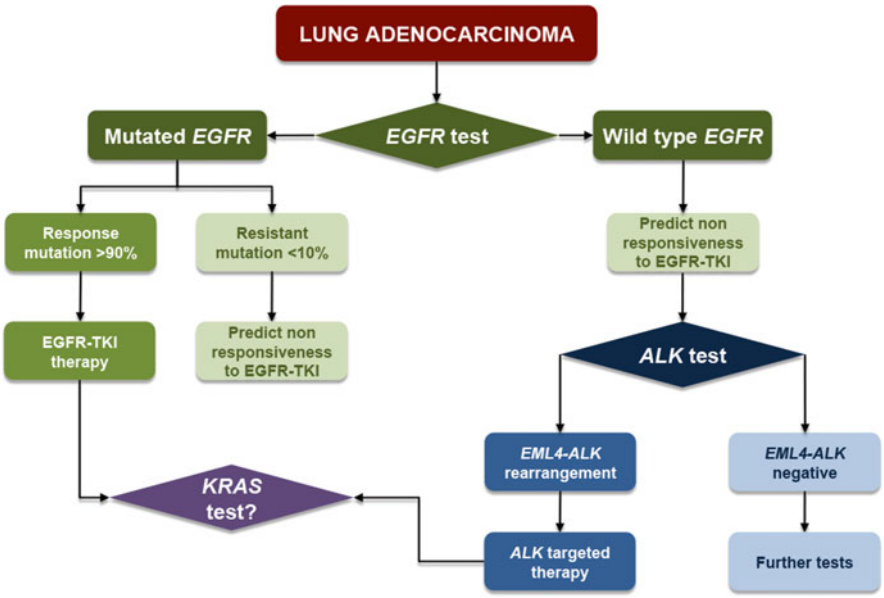


Fig. 2 The current molecular diagnostic workflow for the management of lung adenocarcinoma cytological samples (modified from Cheng et al. [20]). *ALK* anaplastic lymphoma, *BRAF* v-raf murine sarcoma viral oncogene homolog B1, *EGFR* epidermal growth factor receptor, *EML4* echinoderm microtubule protein like-4, *HER2* human epidermal growth factor receptor-2, *KRAS* v-Ki-ras2 Kirsten rat sarcoma viral oncogene homolog, *MEK1* mitogen-activated protein kinase kinase 1, *MET* hepatocyte growth factor receptor, *NRAS* neuroblastoma RAS viral oncogene homolog, *PIK3CA* phosphatidylinositol-4,5-bisphosphate 3-kinase catalytic subunit alpha, *TKI* tyrosine kinase inhibitor

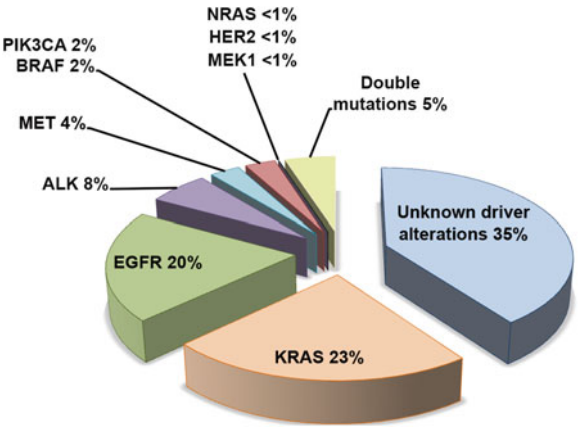


Fig. 3 Driver molecular alterations in lung adenocarcinomas [12–14]. Of note, in a large proportion of cases no driver alteration has been discovered yet. *ALK* anaplastic lymphoma kinase fusion gene, *BRAF* v-raf murine sarcoma viral oncogene homolog B1, *EGFR* epidermal growth factor receptor, *HER2* human epidermal growth factor receptor-2, *KRAS* v-Ki-ras2 Kirsten rat sarcoma viral oncogene homolog, *MEK1* mitogen-activated protein kinase kinase 1, *MET* hepatocyte growth factor receptor, *NRAS* neuroblastoma RAS viral oncogene homolog, *PIK3CA* phosphatidylinositol-4,5-bisphosphate 3-kinase catalytic subunit alpha

4 Role of Next-Generation Sequencing in the Molecular Profiling of Lung Adenocarcinomas

Conventional Sanger sequencing is the most common and widely employed technique for AdC mutational status assessment. However, effectiveness of this method is affected by its low sensitivity, resulting in high false negative rates when applied to cytological samples [3]. Indeed, despite it has been reported that FNA samples and small biopsies usually yield comparable amounts of DNA for molecular testing, this may not be the case for cytological material obtained from pleural effusion, BAL, or brushing [26].

To clear (at least partially) this hurdle, cytological specimens are always checked for relative amount of neoplastic cells and in many cases microdissection is performed to discard the nonneoplastic fraction (mainly inflammatory cells and necrotic debris) and to enrich the tumor cell component. It has been stated that cytological specimens must present at least 25 % of tumor cells to ensure adequate Sanger sequencing [27]. However, this method could be significantly affected also from the low absolute number of neoplastic cells and the presence of a small proportion of mutated cells among the considered tumor cell population.

For these reasons, an increasing number of more sensitive techniques for mutational detection has been developed and are currently used on cytological specimens to molecular characterize lung AdCs. These include restriction fragment length polymorphism (RFLP) and high-resolution melting (HRM) analyses. These types of methods, however, only indirectly highlight mutations and require a subsequent sequencing to confirm and identify the precise identified mutation. On the other hand, techniques such as ARMS-scorpion (TheraScreen), peptide nucleic acid (PNA)-locked PCR clamping, and allele-specific quantitative real-time PCR are based on multiple DNA consuming PCRs to detect only determined specific mutations [11, 28–32].

Thus, all the efforts made ended in procedures guaranteeing only little improvements and that are affected, even if to a lesser degree, by the same limits of Sanger sequencing when they face cytological samples. Indeed, none ensured to detect molecular alterations present only in few tumor cells and exhibited the possibility to analyze a huge number of markers at once.

Next-generation sequencing (NGS) represents the answer to these issues. In fact, NGS allows performing multiple gene analyses from a minute amount of DNA by parallelizing the sequencing process and producing up to millions of sequences concurrently (Fig. 4). Moreover, different NGS panels easily covers all the range of genomic alterations, such as base substitutions, short insertions and deletions, amplifications, homozygous deletions, and gene rearrangements (Fig. 5). The concurrent analysis of multiple gene alterations allows a significant curtailment of time and money to be spent into the analysis [33, 34]. To date, only few pioneer studies tested NGS technology on AdC cytological specimens (Table 2). However, the available data are very promising.

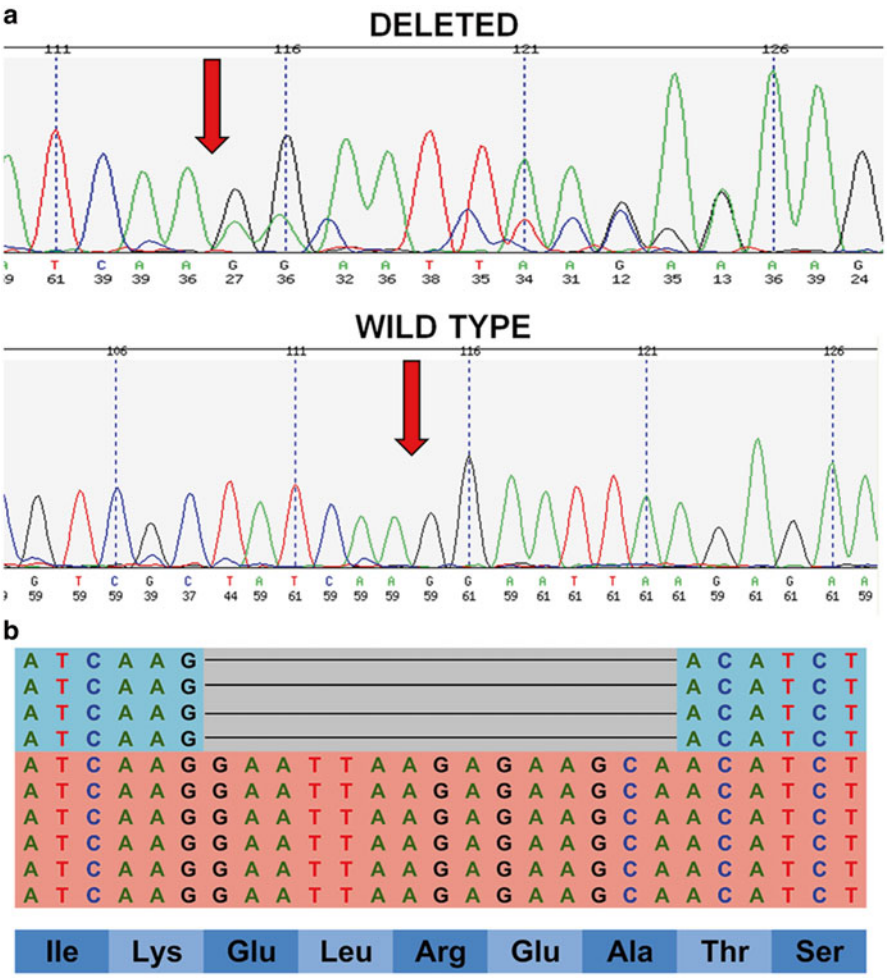


Fig. 4 A representative example of the common *EGFR* deletion E746-A750 in exon 19 as appears in chromatograms using Sanger sequencing (**a**) and in parallel reads aligned to the reference genome using a next-generation sequencing method (**b**)

In their seminal work, Buttitta and colleagues first compared NGS performances in detecting *EGFR* mutations in cytological samples (BAL and pleural effusion) with those of Sanger sequencing [35]. In particular, the author analyzed a series of 33 BALs and 15 pleural effusions corresponding to histologically confirmed AdCs with documented *EGFR* mutations (considered as references). At the cytological evaluation, in 12 cases neoplastic cells were totally absent and in the remaining 36 cases they did not exceed 10 % of cellularity. Mutational analyses were restricted to *EGFR* exons 19 and 21. NGS revealed a greater accuracy compared with Sanger

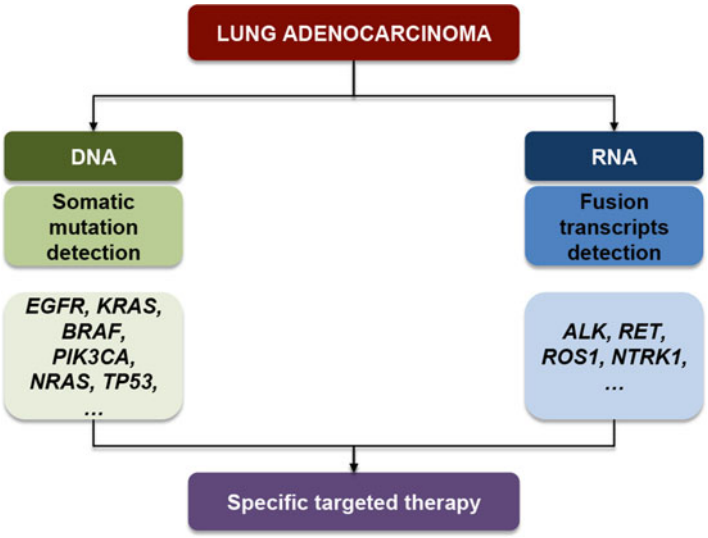


Fig. 5 The proposed NGS molecular diagnostic workflow for the management of lung adenocarcinoma cytological samples. Both DNA and RNA are extracted from the same specimen (with limited amount of material) and are concurrently processed for multiple gene testing

Table 2 Available articles about next-generation sequencing technology in lung adenocarcinoma cytological specimens

Article	Instrument	Method	Specimen
Buttitta et al. [35]	454 GS Junior System	Pyrosequencing	BAL and PE
Karnes et al. [43]	Illumina HiSeq 2000	Sequencing by synthesis	FNA
De Biase et al. [39]	454 GS Junior System	Pyrosequencing	FNA
Moskalev et al. [38]	454 GS Junior System	Pyrosequencing	PE and FNA
Scarpa et al. [42]	Ion Torrent	Ion semiconductor	FNA
Young et al. [41]	Illumina HiSeq 2000	Sequencing by synthesis	FNA

BAL bronchoalveolar lavage, *FNA* fine-needle aspiration, *PE* pleural effusion

sequencing testing these cytological specimens. Indeed, NGS was able to detect molecular alteration, corresponding to those observed in the matching resected tissues, in 32 out of 48 cases, included 5 cases without morphologic evidence of malignancy. Sanger sequencing, instead, only confirmed 5 cases, all belonging to the cytologically positive group. The detection by NGS of *EGFR* mutations in samples lacking malignant cells could be ascribed to the possible presence of DNA free molecules or microvesicles. On the other hand, negative NGS results in the specimens with morphologically confirmed neoplastic cells could be attributed to deficiencies of the sampling method or to cancer heterogeneity. Indeed, in a tumor

mass may coexist several cell clones harboring different molecular alterations, even in the same gene [33, 36, 37]. Thus, it is not a coincidence that in two cases Buttitta et al., in addition to the known molecular alterations, found through NGS analysis mutations that did not have been previously highlighted by Sanger sequencing in the surgical specimens [35].

Similar results were achieved by Moskalev and De Biase [38, 39]. Moskalev and colleagues assessed the mutational status of the *EGFR* (exons 18, 19, 20, and 21) and the *KRAS* (exons 2 and 3) genes in a series of 21 AdC samples including 4 pleural effusions and 3 FNAs. NGS identified *EGFR* mutation not only in the tumor cell rich (cell content >40 %) pleural effusion and FNA samples resulted mutated with Sanger sequencing too, but also in two FNA and in one pleural effusion specimens with low neoplastic cell rate (less than 10 %) previously labeled as wild type. Moreover, NGS detected a *KRAS* mutation in a pleural effusion sample with tumor cell content of 5 % that was negative for Sanger sequencing. The remaining cytological specimen with 35 % of neoplastic cells resulted wild type with both tested methods.

De Biase and colleagues defined a novel NGS protocol targeted to *EGFR* exons 18–21 suitable for the routine diagnosis of cytology/small biopsies samples and tested this protocol on 80 samples obtained from three referral medical centers in Italy. In six cases NGS identified exon 19 deletions or the L858R mutation not seen after Sanger sequencing, allowing the patient to be treated with TKIs. In one additional case the R831H mutation associated with treatment resistance was identified in an *EGFR* wild type tumor after Sanger sequencing.

These three studies agree in concluding that mutations can be reliably identified by NGS even in a minority (up to the 0.2 %) of DNA molecules [35, 38]. Of interest, the above mentioned PCR-based methods that should have superseded Sanger sequencing do not exceed a sensitivity of 1:100 in dilution experiments, whereas NGS reach a value of 1:10,000 [35, 40]. Thus, NGS achieves frontiers of sensitivity unconceivable before.

The other concern about cytological specimens in the AdC setting is the possibility to obtain adequate marker coverage. Indeed, just now multiple different molecular alterations need to be tested in a single cytological sample for selecting the appropriate therapy, and the number of determinations is destined to grow. The works of Young et al., Karnes et al., and Scarpa et al. were designed to explore this opportunity by applying NGS technology [41–43].

Young and colleagues analyzed a broad panel of genes comprehending 4,561 exons of 287 cancer-related genes and 47 introns of 19 genes in a series of 16 lung cancer FNA specimens [41]. The series included six AdC, five squamous cell carcinoma, three NSCLC not otherwise specified (NOS), and two SCLC cases. The NGS analysis required a small amount of DNA (50 ng) extracted from at least 15,000 cells (regardless of the relative amount of the neoplastic component) of each sample. As for AdC specimens, each case showed more than a single molecular alteration (mean = 5.6; range = 3–9). The most common affected genes were *TP53* (6/6 cases), *RB1* (2/6 cases), *FGF4* (2/6 cases), *FGF3* (2/6 cases), *FGF19* (2/6 cases), *EGFR* (2/6 cases), *CCND1* (2/6 cases), and *MCL1* (2/6 cases). *EMSY*,

STKI1, *SMAD4*, *NF1*, *NF2*, *SETD2*, *MYC*, *KEAP1*, *SMARCA4*, *SMARCB1*, *ASXL1*, *APC*, and *ATRX* were altered in single cases. NSCLC-NOS samples displayed a lower number of molecular alterations (three in a case and two in the remaining cases). Again *TP53* was the most frequently involved gene (in two out of three specimens). The other hit genes were *EGFR*, *KRAS*, *NOTCH2*, *APC*, and *NF1*.

In their studies Karnes and Scarpa tested the feasibility and reliability of two NGS panels targeting hot-spot regions of commonly mutated genes in cancer [42, 43]. Karnes analyzed five FNA smears (both trans-bronchial and trans-thoracic) of lung AdC and the matched histological samples, demonstrating that there were not significant differences among these kinds of specimens [43]. Indeed, the concordance of total reads and of single-nucleotide variants across specimens were both higher than 99 %. Moreover, the authors found that the total reads generated, the percentages of mapped (i.e., on target) and unique reads (i.e., read pairs with unique start coordinates), and the depth of sequencing coverage were virtually identical between cytological and histological samples.

Our group examined 504 mutational hotspots of the 22 genes included in the panel using barely 10 ng of DNA extracted from each of the 38 trans-thoracic AdC FNAs [42]. In this relatively large series of cytological specimens, we did not achieve library amplification only in two scraped slides. However, 9 out of 36 cases showed multiple molecular alterations and in 24 cases at least one mutated gene was observed. These included *EGFR*, *KRAS*, *PIK3CA*, *BRAF*, *TP53*, *PTEN*, *MET*, *SMAD4*, *FGFR3*, *STK11*, *MAP2K1*. Of interest, in this series *EGFR* and *KRAS* mutations resulted mutually exclusive.

5 Conclusions

Overall these findings underlie that the shortage of diagnostic material available in most cytological AdC specimens does not represent a limit in obtaining a sensitive, specific, and comprehensive molecular characterization of the tumor by using NGS.

Since the overwhelming superiority of NGS in comparison with Sanger sequencing is clear, a possible obstacle to its wide and routine application in the AdC cytological setting could be represented by its procedure that is relatively labor-intensive and therefore unpractical for the ad hoc analysis of individual specimens as soon as they arrive to the cytopathology laboratory. However, many samples can be analyzed at the same time, even for a considerable number of different genes. As a result, both costs and timing of NGS analysis are significantly lower in comparison to more “traditional” methods.

The application of NGS in routine cytopathology molecular diagnostics needs validation in larger series of cases. However, its performances in detecting a wide range of genetic alterations with an extremely high sensitivity and specificity can help to assess tumor-specific therapeutic susceptibility and individual prognosis. The upcoming challenge lies in the reliable identification of an ultimate AdC-specific multigene panel to significantly improve the care of lung cancer patients.

References

1. Siegel R, Ma J, Zou Z, Jemal A. Cancer statistics, 2014. CA Cancer J Clin. 2014;64:9–29.
2. Travis WD, World Health Organization, International Agency for Research on Cancer, International Association for the Study of Lung Cancer, International Academy of Pathology. Pathology and genetics of tumours of the lung, pleura, thymus and heart. Lyon: IARC Press; 2004. 344 pp.
3. Fassina A, Cappellesso R, Simonato F, Lanza C, Marzari A, et al. Fine needle aspiration of non-small cell lung cancer: current state and future perspective. Cytopathology. 2012;23:213–9.
4. Fassina A, Corradin M, Zardo D, Cappellesso R, Corbetti F, et al. Role and accuracy of rapid on-site evaluation of CT-guided fine needle aspiration cytology of lung nodules. Cytopathology. 2011;22:306–12.
5. Fischer AH, Cibas ES, Howell LP, Kurian EM, Laucirica R, et al. Role of cytology in the management of non-small-cell lung cancer. J Clin Oncol. 2011;29:3331–2. author reply 3332–3.
6. Baram D, Garcia RB, Richman PS. Impact of rapid on-site cytologic evaluation during transbronchial needle aspiration. Chest. 2005;128:869–75.
7. Santambrogio L, Nosotti M, Bellaviti N, Pavoni G, Radice F, et al. CT-guided fine-needle aspiration cytology of solitary pulmonary nodules: a prospective, randomized study of immediate cytologic evaluation. Chest. 1997;112:423–5.
8. Travis WD, Brambilla E, Noguchi M, Nicholson AG, Geisinger KR, et al. International association for the study of lung cancer/American thoracic society/European respiratory society international multidisciplinary classification of lung adenocarcinoma. J Thorac Oncol. 2011;6:244–85.
9. Gazziero A, Guzzardo V, Aldighieri E, Fassina A. Morphological quality and nucleic acid preservation in cytopathology. J Clin Pathol. 2009;62:429–34.
10. Sanz-Santos J, Serra P, Andreo F, Llatjos M, Castella E, et al. Contribution of cell blocks obtained through endobronchial ultrasound-guided transbronchial needle aspiration to the diagnosis of lung cancer. BMC Cancer. 2012;12:34.
11. Fassina A, Gazziero A, Zardo D, Corradin M, Aldighieri E, et al. Detection of EGFR and KRAS mutations on trans-thoracic needle aspiration of lung nodules by high resolution melting analysis. J Clin Pathol. 2009;62:1096–102.
12. Kris MG, Johnson BE, Berry LD, Kwiatkowski DJ, Iafrate AJ, et al. Using multiplexed assays of oncogenic drivers in lung cancers to select targeted drugs. JAMA. 2014;311:1998–2006.
13. Korpanty GJ, Graham DM, Vincent MD, Leighl NB. Biomarkers that currently affect clinical practice in lung cancer: EGFR, ALK, MET, ROS-1, and KRAS. Front Oncol. 2014;4:204.
14. Yu Y, He J. Molecular classification of non-small-cell lung cancer: diagnosis, individualized treatment, and prognosis. Front Med. 2013;7:157–71.
15. Sharma SV, Bell DW, Settleman J, Haber DA. Epidermal growth factor receptor mutations in lung cancer. Nat Rev Cancer. 2007;7:169–81.
16. Pennell NA. Integration of EGFR inhibitors and conventional chemotherapy in the treatment of non-small-cell lung cancer. Clin Lung Cancer. 2011;12:350–9.
17. Sasaki T, Rodig SJ, Chirieac LR, Janne PA. The biology and treatment of EML4-ALK non-small cell lung cancer. Eur J Cancer. 2010;46:1773–80.
18. Kwak EL, Bang YJ, Camidge DR, Shaw AT, Solomon B, et al. Anaplastic lymphoma kinase inhibition in non-small-cell lung cancer. N Engl J Med. 2010;363:1693–703.
19. Just PA, Cazes A, Audebourg A, Cessot A, Pallier K, et al. Histologic subtypes, immunohistochemistry, FISH or molecular screening for the accurate diagnosis of ALK-rearrangement in lung cancer: a comprehensive study of Caucasian non-smokers. Lung Cancer. 2012;76:309–15.
20. Cheng L, Alexander RE, Maclennan GT, Cummings OW, Montironi R, et al. Molecular pathology of lung cancer: key to personalized medicine. Mod Pathol. 2012;25:347–69.
21. Chin LP, Soo RA, Soong R, Ou SH. Targeting ROS1 with anaplastic lymphoma kinase inhibitors: a promising therapeutic strategy for a newly defined molecular subset of non-small-cell lung cancer. J Thorac Oncol. 2012;7:1625–30.

22. Davies KD, Le AT, Theodoro MF, Skokan MC, Aisner DL, et al. Identifying and targeting ROS1 gene fusions in non-small cell lung cancer. Clin Cancer Res. 2012;18:4570–9.
23. Sun Y, Ren Y, Fang Z, Li C, Fang R, et al. Lung adenocarcinoma from East Asian never-smokers is a disease largely defined by targetable oncogenic mutant kinases. J Clin Oncol. 2010;28:4616–20.
24. Janne PA, Shaw AT, Pereira JR, Jeannin G, Vansteenkiste J, et al. Selumetinib plus docetaxel for KRAS-mutant advanced non-small-cell lung cancer: a randomised, multicentre, placebo-controlled, phase 2 study. Lancet Oncol. 2013;14:38–47.
25. De Luca A, Normanno N. Predictive biomarkers to tyrosine kinase inhibitors for the epidermal growth factor receptor in non-small-cell lung cancer. Curr Drug Targets. 2010;11:851–64.
26. Sigel CS, Moreira AL, Travis WD, Zakowski MF, Thornton RH, et al. Subtyping of non-small cell lung carcinoma: a comparison of small biopsy and cytology specimens. J Thorac Oncol. 2011;6:1849–56.
27. Rekhtman N, Brandt SM, Sigel CS, Friedlander MA, Riely GJ, et al. Suitability of thoracic cytology for new therapeutic paradigms in non-small cell lung carcinoma: high accuracy of tumor subtyping and feasibility of EGFR and KRAS molecular testing. J Thorac Oncol. 2011;6:451–8.
28. Nomoto K, Tsuta K, Takano T, Fukui T, Fukui T, et al. Detection of EGFR mutations in archived cytologic specimens of non-small cell lung cancer using high-resolution melting analysis. Am J Clin Pathol. 2006;126:608–15.
29. Kanaji N, Bandoh S, Ishii T, Kushida Y, Haba R, et al. Detection of epidermal growth factor receptor mutations in a few cancer cells from transbronchial cytologic specimens by reverse transcriptase-polymerase chain reaction. Mol Diagn Ther. 2011;15:353–9.
30. van Eijk R, Licht J, Schrumpf M, Talebian Yazdi M, Ruano D, et al. Rapid KRAS, EGFR, BRAF and PIK3CA mutation analysis of fine needle aspirates from non-small-cell lung cancer using allele-specific qPCR. PLoS One. 2011;6:e17791.
31. Horiike A, Kimura H, Nishio K, Ohyanagi F, Satoh Y, et al. Detection of epidermal growth factor receptor mutation in transbronchial needle aspirates of non-small cell lung cancer. Chest. 2007;131:1628–34.
32. Kawahara A, Azuma K, Sumi A, Taira T, Nakashima K, et al. Identification of non-small-cell lung cancer with activating EGFR mutations in malignant effusion and cerebrospinal fluid: rapid and sensitive detection of exon 19 deletion E746-A750 and exon 21 L858R mutation by immunocytochemistry. Lung Cancer. 2011;74:35–40.
33. Mafficini A, Amato E, Fassan M, Simbolo M, Antonello D, et al. Reporting tumor molecular heterogeneity in histopathological diagnosis. PLoS One. 2014;9:e104979.
34. Simbolo M, Gottardi M, Corbo V, Fassan M, Mafficini A, et al. DNA qualification workflow for next generation sequencing of histopathological samples. PLoS One. 2013;8:e62692.
35. Buttitta F, Felicioni L, Del Grammastro M, Filice G, Di Lorito A, et al. Effective assessment of egfr mutation status in bronchoalveolar lavage and pleural fluids by next-generation sequencing. Clin Cancer Res. 2013;19:691–8.
36. Gerlinger M, Rowan AJ, Horswell S, Larkin J, Endesfelder D, et al. Intratumor heterogeneity and branched evolution revealed by multiregion sequencing. N Engl J Med. 2012;366:883–92.
37. Di Lorito A, Schmitt FC. (Cyto)pathology and sequencing: next (or last) generation? Diagn Cytopathol. 2012;40:459–61.
38. Moskalev EA, Stohr R, Rieker R, Hebele S, Fuchs F, et al. Increased detection rates of EGFR and KRAS mutations in NSCLC specimens with low tumour cell content by 454 deep sequencing. Virchows Arch. 2013;462:409–19.
39. de Biase D, Visani M, Malapelle U, Simonato F, Cesari V, et al. Next-generation sequencing of lung cancer EGFR exons 18–21 allows effective molecular diagnosis of small routine samples (cytology and biopsy). PLoS One. 2013;8:e83607.
40. Tuononen K, Maki-Nevala S, Sarhadi VK, Wirtanen A, Ronty M, et al. Comparison of targeted next-generation sequencing (NGS) and real-time PCR in the detection of EGFR, KRAS, and BRAF mutations on formalin-fixed, paraffin-embedded tumor material of non-small cell lung carcinoma-superiority of NGS. Genes Chromosomes Cancer. 2013;52:503–11.

41. Young G, Wang K, He J, Otto G, Hawryluk M, et al. Clinical next-generation sequencing successfully applied to fine-needle aspirations of pulmonary and pancreatic neoplasms. Cancer Cytopathol. 2013;121:688–94.
42. Scarpa A, Sikora K, Fassan M, Rachiglio AM, Cappellesso R, et al. Molecular typing of lung adenocarcinoma on cytological samples using a multigene next generation sequencing panel. PLoS One. 2013;8:e80478.
43. Karnes HE, Duncavage EJ, Bernadt CT. Targeted next-generation sequencing using fine-needle aspirates from adenocarcinomas of the lung. Cancer Cytopathol. 2014;122:104–13.

Whole-Genome/Exome Sequencing in Acute Leukemia: From Research to Clinics

Marc De Braekeleer, Etienne De Braekeleer, and Nathalie Douet-Guilbert

Abstract Acute leukemia is characterized by abnormal proliferation of hematopoietic cells. Development of new technologies, including whole-genome sequencing (WGS) and whole-exome sequencing (WES), now allows the deciphering of acute leukemia genomes in ever greater detail. WGS and WES have proven their capacity to identify novel, clinically relevant genetic abnormalities (driver mutations). Although these driver mutations occur in a large number of genes, their encoded proteins belong principally to a few classes. They also show preferential associations and mutual exclusions. Furthermore, the results indicate that most acute leukemias are a mosaic of multiple genomes and that their clonal architecture evolves during disease progression. However, many questions and difficulties remain. Indeed, the clinical application of WGS/WES will demand high levels of accuracy, sensitivity and specificity to align the genome and differentiate the significant findings among the huge amounts of data generated. WGS/WES remains expensive and the infrastructure, expertise, notably in bioinformatics, and time necessary to complete analysis are significant barriers to a routine use in the clinical setting. A major challenge would be to determine, among all the mutations identified, which ones are clinically relevant and really confer prognostic information. Therefore, one alternative could be to develop targeted resequencing of genes that have proven prognostic information. It is probable that the full determination by WGS/WES studies of the mutational landscape will lead to a more refined classification of acute leukemia. This could also lead to a more rational use of the chemotherapeutic drugs (personalized treatment) and even the development of new drugs.

M. De Braekeleer (✉) • N. Douet-Guilbert
Laboratoire d'Histologie, Embryologie et Cytogénétique, Faculté de Médecine et des Sciences de la Santé, Université de Brest, Brest, France

Institut National de la Santé et de la Recherche Médicale (INSERM), U1078, Brest, France

Service de Cytogénétique et Biologie de la Reproduction, Hôpital Morvan, CHRU Brest, 2, avenue Foch, 29609 Brest cedex, France
e-mail: marc.debraekeleer@univ-brest.fr

E. De Braekeleer
Division of Stem Cells and Cancer, German Cancer Research Center (DKFZ) & Heidelberg Institute for Stem Cell Technology and Experimental Medicine GmbH (HI-STEM), Heidelberg, Germany

W. Wu, H. Choudhry (eds.), *Next Generation Sequencing in Cancer Research, Volume 2*, DOI 10.1007/978-3-319-15811-2_22

1 Introduction

Acute leukemia is characterized by abnormal proliferation of hematopoietic cells. As in other tumors, the number and complexity of genetic aberrations tend to increase during disease evolution. As for other cancers that are mostly associated with gene mutations, copy number variations (deletions and/or amplifications), and loss of heterozygosity (LOH), acute leukemia is also characterized by the generation of fusion genes due to chromosomal translocations, inversions, or insertions [http://AtlasGeneticsOncology.org; http://cgap.nci.nih.gov/Chromosomes/Mitelman].

Since 1960, when the first specific chromosomal abnormality was identified in chronic myeloid leukemia, a large number of fusion genes due to chromosomal translocations have been identified [see, for example, [1–3]]. The number of fusion genes identified in acute leukemia is still increasing [see, for example, [4, 5]]. However, it was soon evident that other genetic abnormalities, such as gene mutations, were involved in leukemogenesis and/or progression of acute leukemia [see, for example, [6–8]].

In the meantime, technology has much evolved, now allowing the deciphering of cancer genomes in ever greater detail [9]. Several techniques, referred to as next-generation sequencing (NGS), have now been developed [10–12]. Among them, whole-exome sequencing (WES) was designed to selectively sequence the coding regions of the genome (about 1 % of the human genome) and whole-genome sequencing (WGS) was developed to cover the entire genome [12, 13].

In 2008, Ley et al. first applied WGS to an acute myeloid leukemia genome and its matched normal counterpart obtained from the same patient's skin [14]. Since this princeps report, several cases of acute leukemia (myeloid and lymphoid) were subjected to WES/WGS. We review the literature, discuss the benefits and difficulties, and examine how these technologies can be moved from the research laboratory to the clinical setting.

2 WES/WGS in Acute Myeloid Leukemia

2.1 WES/WGS in Cytogenetically Normal (CN) AML

As a normal karyotype is frequently observed in blast cells of AML patients, several attempts have been made to find new genetic abnormalities. Ley et al. (2008) performed WGS in a patient with CN-AML without maturation (FAB AML-M1) associated with a normal karyotype. Among the 181 single nucleotide variations (SNVs) predicted to alter gene function, 14 were validated as germ line and 152 as wild type (false positives). Eight SNVs were validated as somatic (acquired mutations), whereas two indels (small insertions or deletions) were detected in *FLT3* and *NPM1*, already known to be recurrently implicated in AML [14].

Because of improvements in sequencing techniques, they reevaluated this case with deeper sequence coverage [15]. Coverage of 99.6 % of the genome instead of

91.2 % allowed them to identify several nonsynonymous mutations that were missed in their initial sequencing analysis, including a somatic mutation in *DNMT3A*. Sequencing of all 24 exons of *DNMT3A* in 281 patients showed that 62 (22.1 %) had mutations. Mutations in *FLT3*, *NPM1*, and *IDH1* were enriched in samples with *DNMT3A* mutations. None of the 11 patients with structural variations involving 11q23 (*MLL*) had *DNMT3A* mutations [15].

Mardis et al. (2009) performed WGS in a patient with AML-M1 associated with a normal karyotype. They identified 12 acquired mutations within the coding sequences of genes, including *NRAS*, *NPM1*, and *IDH1*. Mutations in *IDH1* were also found in 16 of 188 AML patients (8.5 %), most of them having a normal karyotype or a trisomy 8 [16]. In a subsequent analysis on 358 AML patients associated with a normal karyotype, *IDH1* and *IDH2* mutations were identified in 14 % and 19 %, respectively [17].

Studying a case of AML with maturation that carry both wild type *NPM1*, *CEBPA*, *FLT3*, and *MLL* genes by WES, Grossmann et al. (2011) identified mutations in 11 genes, including *DNMT3A* and *BCOR*. Mutations in the *BCOR* gene were found in 13 of 81 AML patients carrying the same genotype as the index case (16 %). No *BCOR* mutation was found among 131 AML cases carrying various cytogenetic abnormalities such as tMLL, t(8;21), t(15;17), inv(16). More studies determined that *BCOR* mutations were mutually exclusive of *NPM1* mutations but associated with *DNMT3A* mutations [18].

WES was performed on five CN-AML with biallelic *CEBPA* gene mutations (biCEPBA) [19]. Tumor-specific nonsense and missense mutations were found in a total of 21 genes, among which *IKZF1*, *STAG2*, and *KRAS. DNMT3A* and *GATA2* mutations were identified in two of the five cases. More mutational screening detected *GATA2* mutations in 13 of 33 biCEBPA-positive AML patients (39.4 %) but none among 38 CN-AML patients with a monoallelic *CEBPA* mutation or in 89 CN-AML patients with wild-type *CEBPA* gene. Furthermore, the mutual exclusiveness of *GATA2* mutations and *FLT3-ITD* within biCEBPA-mutated patients suggested alternative mechanisms of leukemogenesis in these genetic subgroups [19].

2.2 WES/WGS in Specific AML Subtypes

Greif et al. (2011) performed WES of three patients diagnosed with *PML/RARA-positive* acute promyelocytic leukemia (APL—AML-M3) to identify somatic mutations. They identified three to six nonsynonymous coding mutations per patient. A mutation was found in *WT1* and *KRAS* in one patient each. Furthermore, no overlap between mutations in the three APL patients was found, indicating that the spectrum of mutations that may cooperate with *PML/RARA* might be large and diverse [20].

Welch et al. (2012) used WGS to compare 12 genomes from patients with APL associated with a t(15;17) to 12 genomes from patients with AML-M1 with normal cytogenetics [21]. Both AML subtypes had approximately the same number of overall mutations in their genomes. Nine recurrently mutated genes, including

FLT3, *NRAS*, and *WT1*, were found in both the AML-M1 and M3 genomes. This suggested that these mutations may cooperate with a variety of initiating events to produce the disease. Thirteen genes were recurrently mutated only in AML-M1 genomes. They included three genes members of the cohesin complex (*STAG2*, *SMC3*, and *SMC1A*), *NPM1*, *DNMT3A*, *IDH1*, *TET2*, *IDH2*, *RUNX1*, *ASXL1*, *PTPN11*, *DIS3*, and *KIT*, suggesting that they might be involved with AML initiation. It is worth noting that mutations in *NPM1*, *DNMT3A*, and *IDH1* occur only rarely in AML-M3 genomes. Furthermore, nonrandom associations of *NPM1*, *DNMT3A*, *IDH1*, and *FLT3* mutations were observed in AML-M1 cases [21].

The WES of nine paired samples of acute monocytic leukemia (AML-M5) cases allowed the identification of 58 somatic mutations (including *CEBPA*, *FLT3*, *GATA2*, *NRAS*, *NSD1*, *RUNX1*, and *WT1*) and 8 indels [22]. A mutation in *DNMT3A* was found in one case. Another patient in a second set of five AML-M5 was also found to have a *DNMT3A* mutation. Extending the search for *DNMT3A* mutations in a AML-M5 series revealed 23 patients out of 112 (20.5 %) to be carriers. *DNMT3A* mutations were also identified in 9 of 66 (13.6 %) AML-M4 but in none of 177 AML types M1, M2, and M3. Furthermore, *DNMT3A* mutations were preferentially associated with *NPM1* mutations [22].

Children with Down syndrome (DS) have a 10–50-fold higher incidence of leukemias than euploid children. This is particularly true for acute megakaryoblastic leukemia (AMKL), which is preceded by transient myeloproliferative disorder (TMD) [23]. Mutations in *GATA1* are always observed in cell expansions of TMD and AMKL associated with Down syndrome [24]. Nikolaev et al. (2013) performed WES on five TMD and three AMKL. They confirmed the presence of *GATA1* mutations in all samples. Additional mutations in genes affecting WNT, JAK-STAT, or MAPK/PI3K were identified in two of five TMD cases, including one that evolved to AMKL, and in all AMKL cases [25].

Yoshida et al. (2013) performed WGS in four patients with Down syndrome at the TMD and AMKL stages [26]. The mean number of validated somatic mutations was much higher in AMKL samples than in TAM samples. WES on 11 TAM and 10 AMKL samples detected *GATA1* mutations in all cases. *GATA1* was the only recurrent mutational target in TAM samples while an additional eight genes were recurrently mutated in AMKL samples, including *RAD21*, *STAG2*, *NRAS*, *CTCF*, *DCAF7*, *EZH2*, *KANSL1*, and *TP53. EZH2* is one of the most frequently mutated and deleted genes in childhood AMKL, as mutations or deletions were identified in 16 of 49 AMKL associated with Down syndrome (33 %) and in 3 of 19 AMKL not associated with Down syndrome (16 %) [26]. Interestingly, AMKL in children with or without Down syndrome are characterized by distinctive genetic features [27, 28].

Herold et al. (2014) performed WES in 34 patients with AML associated with an isolated trisomy 13. They identified mutations not only in genes already known to be involved in leukemia (*RUNX1*, *SRSF2*, *ASXL1*, *BCOR*) but also in *CEBPZ*, a novel recurrently mutated gene. The analysis revealed a striking clustering of lesions in a few genes, defining trisomy 13-associated AML as a genetically homogenous leukemia subgroup [29].

The larger series of adult de novo AML was analyzed in a collaborative effort to identify mutations involved in leukemogenesis [30]. Two hundred samples were studied using WGS (50 patients) or WES (150 patients). Samples with *MLL* fusion or *PML/RARA* fusion had a lower mean number of somatic mutations while samples containing a *RUNX1/RUNX1T1* fusion or a *TP53* mutation associated with a high-risk cytogenetic profile had a higher mean number of mutations. This suggests that AML initiating events require different numbers of cooperating mutations to achieve full leukemia picture. Twenty-three genes had a higher than expected mutation prevalence, having already been implicated in AML pathogenesis. Each sample contained evidence of a single founding clone, and more than half had one or more subclones derived from the founding clone, showing the heterogeneity. These mutated genes could be grouped in sets according to biologic function or pathways: transcription factor fusions (18 %), *NPM1* (27 %), tumor-suppressor genes (16 %), DNA-methylation genes (44 %), signaling genes (59 %), chromatin-modifying genes (30 %), myeloid transcription factor genes (22 %), cohesion complex genes (13 %), and spliceosome complex genes (14 %). This study also confirmed the strong association or the mutual exclusivity of some genes. For example, there was a strong association between mutations in *NPM1*, *FLT3*, and *DNMT3A* but a strong mutual exclusivity between *PML/RARA*, *MYH11/CBFB*, *MLL* fusions and mutations in *NPM1* and *DNMT3A* or between *RUNX1* and *TP53* mutations and *FLT3* and *NPM1* mutations [30].

2.3 WES/WGS in Secondary AML Following Myelodysplasia

In a study on 28 patients with myelodysplastic syndrome using WES, Yoshida et al. (2011) found that frequent spliceosome mutations were uniquely associated with myelodysplasia phenotypes [31]. A closer inspection of an updated list of mutations, including newly validated single-nucleotide variants, allowed Kon et al. (2013) to identify mutations in genes involved in the cohesin complex (recurrent mutations in *STAG2* and mutations in *STAG1* and *PDS5B* in single specimens) [32]. They examined 157 AML specimens for mutations in nine cohesin or cohesin-related genes using high-throughput sequencing. A total of 19 samples (12.1 %) were found to have a mutation or deletion in *STAG2*, *RAD21*, *SMC1A*, or *SMC3* in a mostly mutually exclusive manner. Cohesin mutations frequently coexisted with other mutations common in myeloid neoplasms and were significantly associated with mutations in *TET2*, *ASXL1*, and *EZH2* [32].

SETBP1 mutations were identified in two cases of refractory anemia with excess of blasts II (RAEB-2) among 20 patients with myeloid hemopathies subjected to WES by Makishima et al. [33]. Further studies showed that *SETBP1* mutations were present in 52 out of 727 cases (7.2 %), including 19 of 113 (16.8 %) secondary AML, but only one of 144 de novo AML. *CBL* mutations were significantly associated with *SETBP1* mutations while *FLT3* and *NPM1* mutations were exclusive of *SETBP1* mutations. This suggests that *CBL* and *SETBP1* mutations potentially cooperate in leukemia progression [33].

Li et al. (2011) performed WES in eight patients with secondary AML associated or not with cytogenetic abnormalities following myelodysplastic syndrome. They identified mutations in five genes, including *BCORL1*, *NRAS*, *IDH1*, *DNMT3A*, and RUNX1, in two patients and in *IDH2* and *TP53* in one patient each. Sequencing all exons of *BCORL1* in a set of 173 AML patients showed 10 patients (5.8 %) to carry a mutation [34].

Walter et al. (2012) performed WGS of seven patients who developed AML following a myelodysplastic syndrome. They compared the mutation burden both at the leukemic and the myelodysplastic stage. There were 17–32 somatic point mutations or indels per secondary AML genome in 168 genes among the seven samples. Two recurrently mutated genes, *RUNX1* and *UMODL1*, were detected in two samples each. In the samples from all seven subjects, the secondary AML genomes were oligoclonal. The preexisting myelodysplastic syndrome founding clone always persisted in secondary AML, although it was outcompeted by daughter subclones in some cases. With the acquisition of each new set of mutations, all the preexisting mutations were carried forward, resulting in subclones that contained increasing numbers of mutations during evolution. Mutations in new clones must confer a growth advantage for them to successfully compete with ancestral clones. The result is that these secondary AML samples are not monoclonal but are instead a mosaic of several genomes with unique sets of mutations; this mosaic is shaped by the acquisition of serial mutations and clonal diversification [35].

Fernandez-Mercado et al. (2013) performed WES on paired samples from one MDS case with del(7)(q21) before and after progression to AML to investigate the molecular events associated with disease progression in MDS. They identified 15 acquired nonsynonymous exonic variations, of which 9 were present at both the MDS and AML stages while 6, involving notably *SETBP1*, were exclusively present after AML transformation. Subsequent analyses revealed mutations in the *SETBP1* gene in 14 of 328 patients with myeloid hemopathies (4.3 %) [36].

Studying by WES three patients with AML that evolved from MDS, Pellagatti et al. (2014) identified a number of somatic SNVs that appeared during progression to AML. Several (*RLF*, *TET2*, *CECR2*) are recurrently observed in AML. It is interesting to note that *TP53* mutations were identified in both cases evolving from MDS associated with a deletion of the long arm of chromosome 5 [del(5q)] [37].

2.4 Sequential WES/WGS in AML

Ding et al. (2012) performed WGS in eight AML patients at diagnosis and relapse to investigate the genetic changes associated with relapse. The number of coding mutations was higher at relapse than at diagnosis in seven patients. A large number of somatic mutations were shared between the primary tumor and relapse samples while fewer mutations were specifically found at relapse only. These data demonstrate that AML cells routinely acquire a small number of additional mutations at relapse, and suggest that some of these mutations may contribute to clonal selection

and chemotherapy resistance. Several clusters of mutations (clones) were present at diagnosis in four patients, signing the clonal heterogeneity at diagnosis. Furthermore, these authors identified two evolution patterns during relapse. The unique clone present at diagnosis gains mutations and evolves into the relapse clone, or one or several subclones at diagnosis survive initial therapy, gain additional mutations and expand at relapse [38].

2.5 Targeted Resequencing of Genes in AML

Several groups have used a targeted resequencing approach to detect mutations in genes already known to be involved in leukemia and/or cancer. However, the nature and number of genes analyzed varies greatly between studies.

Duncavage et al. (2012) used targeted next-generation sequencing (exons and introns) for detecting translocations, somatic mutations and indels in 20 genes implicated in leukemia prognosis [39]. Spencer et al. (2013) used targeted next-generation sequencing-based panel for detecting somatic mutations in 27 genes that are frequently mutated in cancer [40].

Yamashita et al. (2010) performed large-scale resequencing of exons or exon–intron boundaries of 5,648 protein-coding genes in 19 AML samples. They identified nonsynoymous somatic mutations in 11 genes, including *DNMT3A*, *NRAS*, and *JAK3*. Sequencing the entire coding region of *JAK3* in 83 AML samples revealed *JAK3* sequence changes in 8 samples [41].

Van Vlierberghe et al. (2011) used an X-chromosome-targeted mutational analysis approach of T-cell acute lymphoblastic leukemia male patients. They found the *PHF6* gene to be recurrently involved through loss of function mutations or deletions [42]. DNA sequencing analysis of all coding exons of *PHF6* identified mutations in 10 of 353 AML patients (3 %). Sequencing all codons of the *IDH1*, *IDH2*, *TET2*, *ASXL1*, *FLT3*, *NPM1*, *CEBPA*, *WT1*, *KRAS*, and *NRAS* genes revealed additional mutations in *IDH2*, *ASXL1*, *CEBPA*, and *NRAS* in patients with mutated *PHF6* [43].

Patel et al. (2012) performed mutational analysis of the entire coding regions of *TET2*, *ASXL1*, *DNMT3A*, *PHF6*, *WT1*, *TP53*, *EZH2*, *RUNX1*, and *PTEN* and of coding exons with known somatic mutations of *FLT3*, *HRAS*, *KRAS*, *NRAS*, *KIT*, *IDH1*, and *IDH2* using bidirectional Sanger sequencing. They also performed pooled amplicon resequencing of *NPM1* and *CEBPA*. They identified frequently co-occurring mutations and mutations that were mutually exclusive in a cohort of 454 patients. Their results also suggested that mutational profiling could potentially be used for risk stratification and to inform prognostic and therapeutic decisions regarding patients with AML [44].

Dolnik et al. (2012) used a targeted resequencing approach to identify mutations in a set of 50 AML patients. First, they identified genomic regions affected by recurrent genomic gains and losses by array-based comparative genomic hybridization and single nucleotide polymorphism array-based analyses of 391 AML cases.

Then, they selected 1,000 candidate genes located in these regions and sequenced all the coding exons in 50 paired diagnosis and remission AML samples. This set included AML with a normal karyotype associated or not with known mutations (*CEBPA*, *NPM1*, *FLT3*, or *WT1*), AML with a complex karyotype or with specific chromosomal abnormalities (t(8;21) or inv(16)). They identified 120 tumor-specific missense or nonsense mutations and 60 indels in 73 genes, most of them in a nonrecurrent manner. Mutations in genes already known to be affected in AML (*GATA2*, *IDH1*, *KIT*, *KRAS*, *NRAS*) were found. Most importantly, mutations affecting at least one gene linked to epigenetic regulation of transcription (*TET2*, *TET1*, *DNMT3A*, *DNMT1*, *NSD1*, *EZH2*, and *MLL3*) were observed in 40 % of the patients. They also reported for the first time mutations in the *RAD21* gene, affecting about 4 % of the patients [45].

3 WES/WGS in Acute Lymphoblastic Leukemia

The majority of chromosomal abnormalities described in ALL are considered to be primary genetic changes driving leukemia initiation, with additional genetic events required for the development of overt ALL [46]. Furthermore, it has been demonstrated that the majority of ALL cases show changes in the patterns of structural genomic alterations from diagnosis to relapse [47]. Many acquired lesions at relapse are present at low levels at diagnosis, suggesting that genetically determined tumor heterogeneity is a key determinant of treatment failure and relapse [48, 49]. However, next-generation sequencing (targeted and genome wide) has revealed the presence of many additional genetic abnormalities and added a new dimension by identifying new genetic alterations.

3.1 T-Cell Acute Lymphoblastic Leukemia

T-cell ALL is an aggressive malignancy in which multiple genetic defects collaborate in the transformation of T-cell progenitors. Van Vlierberghe et al. (2010) performed an X chromosome-targeted exome sequencing in tumor DNA samples from 12 males with T-cell ALL. They identified somatic mutations and an insertion in the *PHF6* gene in three patients. Mutational analysis of *PHF6* in an extended panel of pediatric and adult T-cell ALL samples at diagnosis identified mutations in 38 % (16/42) of adult and 16 % (14/89) of pediatric samples [42].

Using WES on 67 T-cell ALL children and adults, De Keersmaecker et al. (2013) found that adults (age >15 years) had 2.5 times more somatic mutations than children (21.0 versus 8.2, $P<0.0001$); furthermore, there was a correlation between the age of the affected individual and the number of mutations. They identified 15 candidate driver genes, 8 already known and 7 new. Again, adult samples showed 2.7 times more mutations in candidate driver genes than children (1.9 versus 0.7,

$P = 0.0034$). Moreover, mutations in *CNOT3* and *PHF6* were mainly observed in adults, whereas *RPL10* mutations were almost exclusively found in children [50].

Tsoneva et al. (2013) performed WES of matched diagnosis, remission and relapse DNA samples from five pediatric patients with T-cell ALL. This analysis identified a mean mutation load of 13 somatic mutations per sample. Sixty somatic mutations were identified, 17 being present at diagnosis and relapse, 24 at relapse only, notably in the *NT5C2* gene, and 19 only at diagnosis. In some relapse samples, at least one somatic mutation present at diagnosis was still present with secondary mutations acquired at relapse. *NT5C2* mutation analysis of an extended panel of relapse T-cell ALL and B-cell ALL samples identified mutations in 19 % (20/103) of relapse T-cell ALL and 3 % (1/35) of relapse B-cell ALL. No *NT5C2* mutation was identified in patients at diagnosis [51]. This result supports that relapsed ALL can originate as derivate of ancestral subclones related to, but distinct from, the main leukemic population present at diagnosis.

Early T-cell precursor (ETP) acute lymphoblastic leukemia is a very aggressive leukemia. Zhang et al. (2012) performed WGS of 12 ETP ALL children. The majority of cases harbored alterations in three pathways: loss-of-function mutations in genes encoding regulators of hematopoietic development (*ETV6*, *GATA3*, *IKZF1*, *RUNX1*), activating mutations in cytokine receptor and Ras signaling (*NRAS*, *KRAS*, *FLT3*, *JAK1*, *JAK3*, and *IL7R*), and inactivating mutations targeting epigenetic regulators, most commonly components of the polycomb repressor complex 2 (*EZH2*, *SUZ12*, *EED*), *SETD2*, and *EP300*. The spectrum of mutations indicates that ETP ALL is distinct from non-ETP ALL [52]. Neumann et al. (2013) performed WES in five adult patients with ETP ALL. They identified mutations in genes already known to be involved in leukemogenesis (*ETV6*, *NOTCH1*, *JAK1*, and *NF1*) but also novel recurrent mutations in *FAT1*, *FAT3*, *DNM2*, *MLL2*, *BMI1*, and *DNMT3A*. Further studies revealed a high rate of *DNMT3A* mutations (16 %) in a cohort of 68 patients [53]. These results also suggested that ETP ALL represented a neoplasm of a less mature hematopoietic progenitor or stem cell, with arrest at a very early maturational stage that retained the capacity for myeloid differentiation. Furthermore, the mutation spectrum was different for pediatric and adult patients, pointing toward distinct molecular alterations in pediatric and adult ETP ALL.

3.2 B-Cell Acute Lymphoblastic Leukemia

The genome of *ETV6/RUNX1*-positive ALL has been well characterized at the copy number and cytogenetic level. Deletions affecting genes involved in B-lymphocyte development and differentiation, such as *CDKN2A*, *PAX5*, *RAG1/2*, and the wild-type copy of *ETV6*, are already well known. Lilljebjörn et al. (2012) performed WES of two cases of pediatric ALL carrying the *ETV6/RUNX1* (*TEL/AML1*) fusion gene. They identified 14 somatic mutations, none of them recurrent. None of the identified mutations was present in an extended collection of *ETV6/RUNX1*-positive ALL but 13 of them affected genes previously implicated in cancer, or speculated to be

important in carcinogenesis [54]. Papaemmanuil et al. (2014) performed WGS on leukemic samples of 51 cases of *ETV6/RUNX1*-positive ALL. They identified an average of 11 structural variations and 14 coding point mutations per case. Few recurrent coding-region mutations were observed but genomic rearrangements were frequent and appeared to be the predominant driver of *ETV6/RUNX1*-positive ALL [55].

Andersson et al. (2011) performed WGS on diagnostic leukemia blasts samples from 22 infants (less than 1-year-old) with *MLL* rearranged ALL. A mean of only two somatic structural variations and two single nucleotide variations affecting the coding region of genes were detected per case. Mutations leading to activation of signaling through the PI3K/RAS pathway were observed in 45 % of the cases (*KRAS*, *NRAS*, *NF1*, *PTPN11*, *PIK3R1*, and *ARHGAP32*). B cell differentiation was also altered (*PAX5* and *CDKN2A/B*). WES performed on 13 *MLL* acute leukemia (8 ALL and 5 AML) in older children (7–19 years of age) showed a mean of eight single nucleotide variations per case [56].

Philadelphia (*BCR/ABL1*)-like ALL is a novel subgroup of childhood ALL that shows a gene expression profile similar to that of the Philadelphia (Ph)-positive ALL and shares the same high-risk of relapse and poor outcome. Up to half of Ph-like ALL cases have a rearrangement of *CRLF2* (*IGH/CRLF2* or *P2RY8/CRLF2*). WGS of 15 Ph-like ALL cases, including 12 without *CRLF2* rearrangement, identified a strikingly diverse array of genetic alterations (notably *IKZF1*, *PAX5*, *CDKN2A/CDKN2B*, *JAK2*, *SH2B3*, and *IL7R*) activating cytokine receptor and tyrosine signaling in all cases [57].

Hypodiploid ALL may be subclassified by degree of aneuploidy into near haploid (NH-ALL, 23–31 chromosomes) and low hypodiploid (LH-ALL, 32–44 chromosomes) cases. NH-ALL is associated with particularly poor outcome, contrary to high hyperdiploidy (51–67 chromosomes) that has a favorable prognosis. WES of the "pseudo high hyperdiploid" cell line MHH-CALL-2, derived from a near haploid clone, showed homozygous nonsynonymous mutations in 63 genes, including CDKN2A/CDKN2B, *FANCA, NF1, TCF7L2*, *CARD11*, *EP400*, *KDM6B*, *KDM1A*, and *PRDM11* [58]. Interestingly, using WGS, only 8 of these 63 genes were also mutated, but heterozygously, in a set of high hyperdiploid ALL patients [58]. Holmfeld et al. (2013) performed a detailed genomic analysis of more than 120 hypodiploid ALL cases, including WGS or WES of more than 40 cases. NH-ALL was found to harbor alterations targeting receptor tyrosine kinase signaling and Ras signaling (71 %), including recurring novel alterations of *NF1* and IKZF3 (13 %). In contrast, LH-ALL had alterations in *TP53* (91.2 %), with the mutations present in the germ line in approximately half the cases, *IKZF2* (53 %) and *RB1* (41 %) [59].

Chang et al. (2013) performed WES in four patients with congenital ALL (diagnosis within the first month of life). One to three nonsynonymous somatic mutations were found in each tumor sample (including *FLT3* in two samples). Germ-line mutations in several genes known to be associated with cancer predisposition or involved in DNA repair were also identified in each sample [60].

Zhang et al. (2011) sequenced all exons and flanking splice site junctions of 120 candidate cancer genes in samples from 187 children and adolescents with high-risk B-cell ALL. Of the 680 putative novel variants, 179 were identified to be somatic mutations in 31 genes, 19 of which being recurrently mutated. They included genes

involved in four known cancer signaling pathways: B-cell development/differentiation, Ras signaling, JAK/STAT signaling, and the TP53/RB1 tumor suppressor pathway. Despite the low background mutational rate, multiple genes within the four individual signaling pathways were mutated in a subset of cases, suggesting a strong selection for mutations within these specific signaling pathways [61].

Although chemotherapy is very effective in obtaining remission in ALL (especially in children), relapse remains a real concern and even those patients with favorable cytogenetic abnormalities will eventually relapse. To identify novel sequence mutations in relapsed ALL, Mullighan et al. (2011) resequenced 300 genes in matched diagnosis-relapse samples from 23 children with B-cell ALL. They identified 52 somatic mutations in 32 genes in 20 cases [62]. Many deleterious mutations present at diagnosis were no longer evident at relapse. However, deletions/mutations of *IKZF1* were preserved at relapse or acquired as new lesions. Furthermore, analysis of an extended cohort of 71 diagnosis-relapse cases and 270 acute leukemia cases without relapse found that 18.3 % of relapse cases had sequence or deletion mutations of *CREBBP*. In contrast, *CREBBP* alterations in cases of childhood acute leukemia that did not relapse were rare [62]. Transcriptome sequencing of specimens from ten children with B-cell ALL was performed at diagnosis and relapse. Twenty missense mutations, including two mutations in the *NT5C2*, were specifically found at relapse but absent at diagnosis. Full-exon resequencing of *NT5C2* was completed in an additional 61 relapse specimens and five further somatic mutations were identified. Amplicon resequencing of DNA from diagnosis and relapse specimens identified two cases where a rare clone indeed existed at diagnosis. In the remaining five cases, no mutation could be detected at diagnosis [63]. These data suggested that the emergence of clones containing mutations in *NT5C2* is driven by powerful selective pressures presumably due to drug resistance [63].

Although each tumor type is characterized by a unique genomic landscape, several cellular pathways are mutated in multiple tumor types. They include signaling transcriptional regulation of development/differentiation, antigen receptor signaling, tyrosine kinase and Ras signaling, JAK/STAT signaling pathways [64, 65].

The rather low number of mutations suggests that ALL is more genetically stable than previously anticipated [54]. However, the nature and frequency of genetic lesions is subtype dependent and could also be age dependent, as shown for t(4;11)(q21;q23)-associated ALL [66]. For example, *MLL*-rearranged leukemia harbors very few additional structural or sequence alterations, in contrast to *BCR/ABL1* and *ETV6/RUNX1* subtypes that harbor more alterations. Such findings indicate the more aggressive nature of the *MLL* translocation in ALL, requiring fewer additional genetic alterations for induction of leukemogenesis than the *BCR/ABL1* and *ETV6/RUNX1* subtypes [67].

4 Research Lessons from WGS/WES

The application of various new high-throughput technologies over the past decade, including next-generation sequencing, to the study of acute leukemia has resulted in the identification of a host of genes that are recurrently mutated. It is now widely

anticipated that NGS will enable the in-depth characterization of the leukemia cell genome and further advance the fields of molecular pathology. Still, WGS/WES has raised many questions but few answers have been provided.

In the studies thus far published, a large number of somatic mutations have been identified. A major question addresses the issue of determining which mutations are important for leukemia initiation and progression. Indeed, with each cell division, there is a finite probability of somatic mutations secondary to errors in DNA replication. During their lifetime, hematopoietic stem cells, since they have self-renewal capacity, accumulate somatic mutations that are transferred to daughter hematopoietic stem cells. Thus, it is predicted that hematopoietic stem cells will accumulate mutations as a function of age [21]. As a consequence, the majority of mutations in acute leukemia represents random mutations that were acquired during normal aging of hematopoietic stem cells (called "background" or "passenger" mutations), the majority of these generally being irrelevant for AML pathogenesis [68].

Therefore, a current challenge is to identify those mutations present in a leukemia genome that contribute to leukemogenesis (called "driver" mutations) and to discriminate them from passenger mutations. This can be achieved by identifying those genes mutated in multiple samples. Some of the driver mutations are highly recurrent (>30 % of patients), but there seems to be a continuum of mutation frequency down to rare (<5 %) or even singleton mutations. Recurring mutations in a gene, or at an individual nucleotide position in the genome, are likely to be important for pathogenesis. However, nonrecurrent mutations that impact common biological pathways may also be important for leukemogenesis [69–71].

Because many genes, each of which being mutated infrequently, seem to contribute to leukemia development in only a small fraction of patients, large sample sets will have to be analyzed (https://tcga-data.nci.nih.gov/tcga/). It is likely that many more driver mutations are still to be discovered (http://cancer.sanger.ac.uk/cosmic/census).

WGS/WES studies have revealed that acute leukemia usually harbors few driver mutations. Many questions emerge from their results. Are there preferential associations and mutual exclusions among mutations? What are the gene classes targeted by these mutations? Is acute leukemia monoclonal or oligoclonal at diagnosis? How acute leukemia does evolve?

There is growing evidence that more than one hit is necessary to trigger acute leukemia [72]. A first step in leukemogenesis driven by a genetic alteration is likely to represent just a clonal expansion. However, this first hit could drive the fate of the following steps [73]. Indeed, two driver mutations may never occur together, being mutually exclusive, in the same cell because of redundancy (no growth advantage if both hits occur in the same pathway) or synthetic lethality (cell survival compromised because of counterselection of two hits). Association and cooperation can occur in all other cases [30, 74].

Although driver mutations occur in a large number of genes, their encoded proteins belong principally to a few classes [73]. Table 1 shows the main genes that are recurrently mutated and the major classes to which they belong. There is no doubt that the combination of mutations associated with a given leukemia is extremely large; although the pathways affected by these mutations may be limited. Mutational profiling of acute leukemia will undoubtedly provide insights into com-

Table 1 Distribution of the main genes recurrently mutated in acute leukemia

Signaling proteins	Transcription factors	Epigenetic regulators		Tumor suppressors	Spliceosome complex	Cohesin complex
		DNA methylation	Chromatin remodeling			
CBL	CEBPA	DNMT3A	ASXL1	CDKN2A	SF3A1	SMC1A
FLT3	ETV6	TET2	BCOR	CDKN2B	SF3B1	SMC3
JAK2	GATA1	IDH1	BCORL1	TP53	SRSF2	STAG2
KIT	GATA2	IDH2	CREBBP	WT1	U2AF1	
KRAS	NPM1		EZH2		ZRSR2	
NF1	RARA		IKZF1			
NRAS	RUNX1		IKZF2			
PTPN11			IKZF3			
SH2B3			MLL			
			PHF6			
			SUZ12			

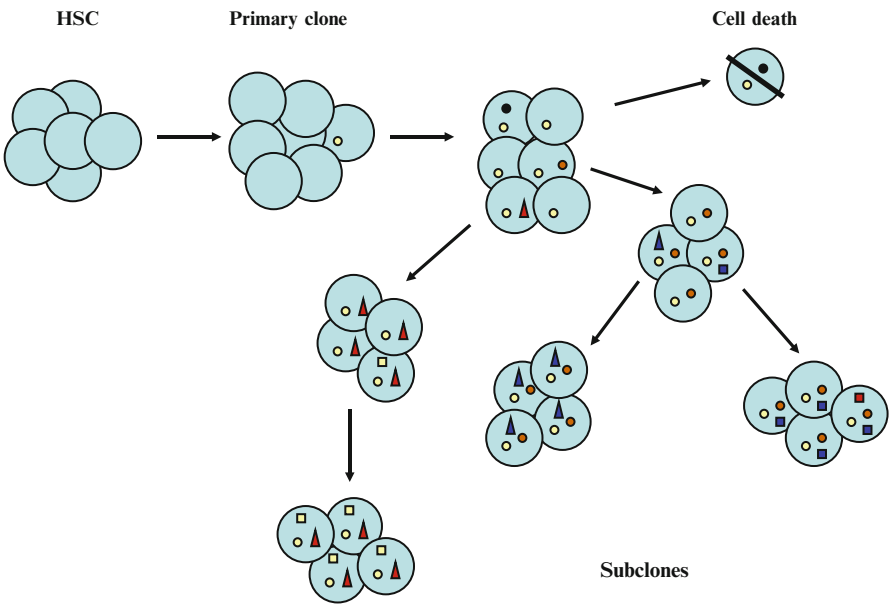


Fig. 1 Schematic picture of the clonal genetic architecture of acute leukemia

mon pathways involved in leukemia transformation and possibly enhance disease classification.

The results obtained by WGS/WES studies indicate that acute leukemia development may be a more complex process than previously thought. They revealed that most acute leukemias are oligoclonal. In fact, they are a mosaic of multiple genomes (Fig. 1). Indeed, once a driver mutation that gives an advantage has occurred in a hematopoietic stem cell, an expanding clone (primary or founding clone) can

develop. Additional cooperating driver mutation(s) can occur in some cells of the clone, leading to subclones in which some cells can also harbor new driver mutations, resulting in genetic heterogeneity [64, 73, 75, 76]. All passenger mutations that have already been accumulated in the hematopoietic stem cell are carried to the founding clone; additional passenger mutations occurring in that founding clone or in subclones are also captured, thus increasing the genetic heterogeneity of acute leukemia. Furthermore, the magnitude of clonal diversity could even be more important as WGS/WES can only detect clones that are present at 5 % or greater. It is likely that subclonal diversity varies continuously with the development and progression of disease. One or more subclones can gain selective advantage, compete with others for survival and outgrowth them [75].

Although a complete remission can be achieved by chemotherapy in most cases of acute leukemia, early or late relapse occurs in the majority of patients. WGS/WES studies have shown that the clonal architecture at relapse was different from that at diagnosis. Sequential studies at diagnosis and relapse demonstrated that the founding clone and/or subclone(s) present at diagnosis may gain a small number of additional mutations that could contribute to clonal selection and result in relapse. Furthermore, while some subclones may be eradicated by chemotherapy, one or more others could be selected, because they are resistant to treatment, and be the starting point of relapse [38, 71, 75].

5 How WGS/WES Can Move from Research to Clinics?

WGS/WES has allowed us to start deciphering the complexity of leukemia genomes. Now that it has proven its capacity to identify novel, clinically relevant genetic abnormalities, time has come to start moving from research laboratories to the hospital setting. However, many questions and difficulties remain. Indeed, its clinical application will demand high levels of accuracy, sensitivity, and specificity to align the genome and differentiate the significant findings among the huge amounts of data generated.

One difficulty is the ability of WGS/WES to detect some sequence variants, which is dependent upon the number of unique times a single nucleotide is sampled ("read-depth"). Even with a high number of reads (30–40), a number of sequence variants identified are in fact false positives, due notably to mapping or polymerase errors. Therefore, validation with another method is a must. Some rearrangements such as medium-sized indels could be difficult to be identified. The limit of sensitivity also remains to be determined for cases in which pathogenic genetic aberrations in leukemic cells are diluted by a larger population of normal cells [39].

WGS/WES remains expensive and the infrastructure, expertise, and time necessary to complete sequence analysis are significant barriers to its routine use in the clinical setting [11, 39, 77, 78]. There is no doubt that its cost will continue to decrease, therefore making it more accessible to academic and, possibly, nonacademic hospitals. However, the expertise in bioinformatics required to analyze sequencing data could remain a limiting factor in its expansion in the clinical setting.

Indeed, WGS/WES is highly dependent on software tools, which can handle large amounts of data, and faces interpretation challenges. Several software tools are used but they still need to be validated.

Another challenge will be to generate reports that can be used by hematologists. At present, most of the somatic mutations detected by WGS/WES are of uncertain biological and clinical significance. Correlation of these findings with leukemia evolution and prognosis will be necessary to provide clinicians with clinically useful information [77, 79].

All the studies thus far reported have used WGS or WES indistinctly. However, these two methods are interchangeable. In WGS, the entire genome is surveyed and structural variants, including deletions, amplifications, chromosomal translocations, and uniparental disomy can be identified. In WES, only 1–2 % of the genome containing coding genes is sequenced but relatively deep sequence coverage can be achieved. However, it does not detect mutations in regions outside of the exome, nor does it detect structural variants, such as chromosomal translocations with intronic breakpoints [12].

6 Clinical Applications of WGS/WES

When the leukemia genome of a patient is compared to his normal genome (usually obtained at remission or from skin DNA), hundreds of copy number alterations and single nucleotide variants are identified in each case. The vast majority of these variants are inherited. Some of these variants will be associated with unsuspected genetic diseases (notably mutations in genes associated with autosomal recessive disorders such as cystic fibrosis, hemochromatosis). Other inherited variants that contribute to cancer susceptibility will also be discovered using WGS/WES.

Two examples can illustrate this. Link et al. (2011) performed WGS in bone marrow and skin DNA from a patient with early-onset breast and ovarian cancer and therapy-related AML. Besides identifying several mutations in the leukemia genome, they also detected a 3 kb heterozygous deletion of *TP53*, encompassing exons 7–9, in the skin genome. Sequence analysis of leukemia DNA revealed a 17.6 Mb region of uniparental disomy on chromosome 17 that resulted in homozygous deletion of exons 7–9 of *TP53* in the leukemia genome [80]. Inherited mutations in *TP53* are known to be associated with the Li–Fraumeni syndrome, a hereditary cancer predisposition syndrome. Shah et al. (2013) identified a heterozygous germ-line *PAX5* variant that was found to segregate with disease in two unrelated kindred with autosomal dominant B-cell ALL [81].

With the development of WGS/WES analyses, it is likely that more germ-line variants contributing to cancer susceptibility will be discovered. This could lead to approaches for earlier cancer detection and even cancer prevention. However, several ethical issues need to be addressed, notably what should be done with this information, should it be communicated to patients and their relatives, and how should it be communicated. There is an urgent need to establish guidelines [60, 68, 70, 80].

Given its power to identify structural rearrangements, WGS could be used to help solving diagnostic difficulties. This is best illustrated by Welch et al. (2011) who analyzed leukemic blasts of a patient with acute promyelocytic leukemia lacking cytogenetic evidence of a rearrangement involving *RARA*. They found a 77-kb sequence from chromosome 15 to be inserted into the second intron of the *RARA* gene on chromosome 17. As a consequence, the medical care was modified and the patient then received all-*trans* retinoic acid (ATRA) treatment [82].

Because the phenotype and, at least in part, the clinical behavior of acute leukemia is the result of a combination of mutations, it is probable that the full determination by WGS/WES studies of the mutational landscape will lead to a more refined classification of AML and ALL. This could also lead to a more rational use of the chemotherapeutic drugs and even the development of new drugs. Indeed, as it is now evident that several pathways are affected in leukemogenesis, the strategy could be to target these pathways by specific drug regimens, which will lead to personalized medicine. However, a major challenge would be to target genetic alterations present in the founding clone and subclones, involving those minor subclones that could escape detection by WGS/WES, but be resistant to "conventional" chemotherapy and trigger relapse [78, 83, 84]. Another major challenge would be to determine, among all the mutations identified, which ones are clinically relevant and really confer prognostic information [85].

One alternative, at least in the near term, is to develop targeted resequencing of genes that have proven prognostic information [71, 77, 86]. This method offers greatly increased scalability, requires less technician labor, and is less expensive [39]. The potential of this approach has been highlighted in a study by Patel et al. (2012) in which a panel of 18 genes was screened using a high-throughput sequencing approach in a cohort of 398 patients with AML [44].

Rapidly advancing next-generation sequencing technology, including WGS/WES, bisulfite sequencing (to provide information about epigenetic modifications), transcriptome sequencing (to measure RNA expression), and microRNA profiling, may be required to comprehensively study cancer cells and will be, in the near future, part of a personalized approach of acute leukemia [12].

References

1. Gao J, Erickson P, Gardiner K, et al. Isolation of a yeast artificial chromosome spanning the 8;21 translocation breakpoint t(8;21)(q22;q22.3) in acute myelogenous leukemia. Proc Natl Acad Sci U S A. 1991;88:4882–6.
2. Golub TR, Barker GF, Bohlander SK, et al. Fusion of the TEL gene on 12p13 to the AML1 gene on 21q22 in acute lymphoblastic leukemia. Proc Natl Acad Sci U S A. 1995;92:4917–21.
3. de Thé H, Chomienne C, Lanotte M, et al. The t(15;17) translocation of acute promyelocytic leukaemia fuses the retinoic acid receptor alpha gene to a novel transcribed locus. Nature. 1990;347:558–61.
4. Meyer C, Hofmann J, Burmeister T, et al. The MLL recombinome of acute leukemias in 2013. Leukemia. 2013;27:2165–76.

5. De Braekeleer E, Auffret R, Douet-Guilbert N, et al. Recurrent translocation (10;17)(p15;q21) in acute poorly differentiated myeloid leukemia likely results in ZMYND11-MBTD1 fusion. Leuk Lymphoma. 2014;55:1189–90.
6. Pabst T, Mueller BU, Zhang P, et al. Dominant-negative mutations of CEBPA, encoding CCAAT/enhancer binding protein-alpha (C/EBPalpha), in acute myeloid leukemia. Nat Genet. 2001;27:263–70.
7. Verhaak RG, Goudswaard CS, van Putten W, et al. Mutations in nucleophosmin (NPM1) in acute myeloid leukemia (AML): association with other gene abnormalities and previously established gene expression signatures and their favorable prognostic significance. Blood. 2005;106:3747–54.
8. Thiede C, Steudel C, Mohr B, et al. Analysis of FLT3-activating mutations in 979 patients with acute myelogenous leukemia: association with FAB subtypes and identification of subgroups with poor prognosis. Blood. 2002;99:4326–35.
9. Hudson TJ, Anderson W, Artez A, et al. International network of cancer genome projects. Nature. 2010;464:993–8.
10. Metzker ML. Sequencing technologies—the next generation. Nat Rev Genet. 2010;11:31–46.
11. Ross JS, Cronin M. Whole cancer genome sequencing by next-generation methods. Am J Clin Pathol. 2011;136:527–39.
12. De Braekeleer E, Douet-Guilbert N, De Braekeleer M. Genetic diagnosis in malignant hemopathies: from cytogenetics to next-generation sequencing. Expert Rev Mol Diagn. 2014;14:127–9.
13. Choi M, Scholl UI, Ji W, et al. Genetic diagnosis by whole exome capture and massively parallel DNA sequencing. Proc Natl Acad Sci U S A. 2009;106:19096–101.
14. Ley TJ, Mardis ER, Ding L, et al. DNA sequencing of a cytogenetically normal acute myeloid leukaemia genome. Nature. 2008;456:66–72.
15. Ley TJ, Ding L, Walter MJ, et al. DNMT3A mutations in acute myeloid leukemia. N Engl J Med. 2010;363:2424–33.
16. Mardis ER, Ding L, Dooling DJ, et al. Recurring mutations found by sequencing an acute myeloid leukemia genome. N Engl J Med. 2009;361:1058–66.
17. Marcucci G, Maharry K, Wu YZ, et al. IDH1 and IDH2 gene mutations identify novel molecular subsets within de novo cytogenetically normal acute myeloid leukemia: a Cancer and Leukemia Group B study. J Clin Oncol. 2010;28:2348–55.
18. Grossmann V, Tiacci E, Holmes AB, et al. Whole-exome sequencing identifies somatic mutations of BCOR in acute myeloid leukemia with normal karyotype. Blood. 2011;118:6153–63.
19. Greif PA, Dufour A, Konstandin NP, et al. GATA2 zinc finger 1 mutations associated with biallelic CEBPA mutations define a unique genetic entity of acute myeloid leukemia. Blood. 2012;120:395–403.
20. Greif PA, Yaghmaie M, Konstandin NP, et al. Somatic mutations in acute promyelocytic leukemia (APL) identified by exome sequencing. Leukemia. 2011;25:1519–22.
21. Welch JS, Ley TJ, Link DC, et al. The origin and evolution of mutations in acute myeloid leukemia. Cell. 2012;150:264–78.
22. Yan XJ, Xu J, Gu ZH, et al. Exome sequencing identifies somatic mutations of DNA methyltransferase gene DNMT3A in acute monocytic leukemia. Nat Genet. 2011;43:309–15.
23. Al-Kasim F, Doyle JJ, Massey GV, et al. Incidence and treatment of potentially lethal diseases in transient leukemia of Down syndrome: Pediatric Oncology Group Study. J Pediatr Hematol Oncol. 2002;24:9–13.
24. Wechsler J, Greene M, McDevitt MA, et al. Acquired mutations in GATA1 in the megakaryoblastic leukemia of Down syndrome. Nat Genet. 2002;32:148–52.
25. Nikolaev SI, Santoni F, Vannier A, et al. Exome sequencing identifies putative drivers of progression of transient myeloproliferative disorder to AMKL in infants with Down syndrome. Blood. 2013;122:554–61.
26. Yoshida K, Toki T, Okuno Y, et al. The landscape of somatic mutations in Down syndrome-related myeloid disorders. Nat Genet. 2013;45:1293–9.

27. Gruber TA, Larson Gedman A, Zhang J, et al. An Inv(16)(p13.3q24.3)-encoded CBFA2T3-GLIS2 fusion protein defines an aggressive subtype of pediatric acute megakaryoblastic leukemia. Cancer Cell. 2012;22:683–97.
28. de Rooij JD, Hollink IH, Arentsen-Peters ST, et al. NUP98/JARID1A is a novel recurrent abnormality in pediatric acute megakaryoblastic leukemia with a distinct HOX gene expression pattern. Leukemia. 2013;27:2280–8.
29. Herold T, Metzeler KH, Vosberg S, et al. Isolated trisomy 13 defines a genetically homogenous AML subgroup with high frequency of mutations in spliceosome genes and poor prognosis. Blood. 2014. doi:10.1182/blood-2013-12-540716.
30. The Cancer Genome Atlas Research Network. Genomic and epigenomic landscapes of adult de novo acute myeloid leukemia. N Engl J Med. 2013;368:2059–74.
31. Yoshida K, Sanada M, Shiraishi Y, et al. Frequent pathway mutations of splicing machinery in myelodysplasia. Nature. 2011;478:64–9.
32. Kon A, Shih LY, Minamino M, et al. Recurrent mutations in multiple components of the cohesin complex in myeloid neoplasms. Nat Genet. 2013;45:1232–7.
33. Makishima H, Yoshida K, Nguyen N, et al. Somatic SETBP1 mutations in myeloid malignancies. Nat Genet. 2013;45:942–6.
34. Li M, Collins R, Jiao Y, et al. Somatic mutations in the transcriptional corepressor gene BCORL1 in adult acute myelogenous leukemia. Blood. 2011;118:5914–7.
35. Walter MJ, Shen D, Ding L, et al. Clonal architecture of secondary acute myeloid leukemia. N Engl J Med. 2012;366:1090–8.
36. Fernandez-Mercado M, Pellagatti A, Di GC, et al. Mutations in SETBP1 are recurrent in myelodysplastic syndromes and often coexist with cytogenetic markers associated with disease progression. Br J Haematol. 2013;163:235–9.
37. Pellagatti A, Fernandez-Mercado M, Di GC, et al. Whole-exome sequencing in del(5q) myelodysplastic syndromes in transformation to acute myeloid leukemia. Leukemia. 2014;28:1148–51.
38. Ding L, Ley TJ, Larson DE, et al. Clonal evolution in relapsed acute myeloid leukaemia revealed by whole-genome sequencing. Nature. 2012;481:506–10.
39. Duncavage EJ, Abel HJ, Szankasi P, et al. Targeted next generation sequencing of clinically significant gene mutations and translocations in leukemia. Mod Pathol. 2012;25:795–804.
40. Spencer DH, Abel HJ, Lockwood CM, et al. Detection of FLT3 internal tandem duplication in targeted, short-read-length, next-generation sequencing data. J Mol Diagn. 2013;15:81–93.
41. Yamashita Y, Yuan J, Suetake I, et al. Array-based genomic resequencing of human leukemia. Oncogene. 2010;29:3723–31.
42. Van Vlierberghe P, Palomero T, Khiabanian H, et al. PHF6 mutations in T-cell acute lymphoblastic leukemia. Nat Genet. 2010;42:338–42.
43. Van VP, Patel J, Abdel-Wahab O, et al. PHF6 mutations in adult acute myeloid leukemia. Leukemia. 2011;25:130–4.
44. Patel JP, Gonen M, Figueroa ME, et al. Prognostic relevance of integrated genetic profiling in acute myeloid leukemia. N Engl J Med. 2012;366:1079–89.
45. Dolnik A, Engelmann JC, Scharfenberger-Schmeer M, et al. Commonly altered genomic regions in acute myeloid leukemia are enriched for somatic mutations involved in chromatin remodeling and splicing. Blood. 2012;120:e83–92.
46. Schwab CJ, Chilton L, Morrison H, et al. Genes commonly deleted in childhood B-cell precursor acute lymphoblastic leukemia: association with cytogenetics and clinical features. Haematologica. 2013;98:1081–8.
47. Raimondi SC, Pui CH, Head DR, et al. Cytogenetically different leukemic clones at relapse of childhood acute lymphoblastic leukemia. Blood. 1993;82:576–80.
48. Mullighan CG, Phillips LA, Su X, et al. Genomic analysis of the clonal origins of relapsed acute lymphoblastic leukemia. Science. 2008;322:1377–80.
49. Yang JJ, Bhojwani D, Yang W, et al. Genome-wide copy number profiling reveals molecular evolution from diagnosis to relapse in childhood acute lymphoblastic leukemia. Blood. 2008;112:4178–83.

50. De Keersmaecker K, Atak ZK, Li N, et al. Exome sequencing identifies mutation in CNOT3 and ribosomal genes RPL5 and RPL10 in T-cell acute lymphoblastic leukemia. Nat Genet. 2013;45:186–90.
51. Tzoneva G, Perez-Garcia A, Carpenter Z, et al. Activating mutations in the NT5C2 nucleotidase gene drive chemotherapy resistance in relapsed ALL. Nat Med. 2013;19:368–71.
52. Zhang J, Ding L, Holmfeldt L, et al. The genetic basis of early T-cell precursor acute lymphoblastic leukaemia. Nature. 2012;481:157–63.
53. Neumann M, Heesch S, Schlee C, et al. Whole-exome sequencing in adult ETP-ALL reveals a high rate of DNMT3A mutations. Blood. 2013;121:4749–52.
54. Lilljebjorn H, Rissler M, Lassen C, et al. Whole-exome sequencing of pediatric acute lymphoblastic leukemia. Leukemia. 2012;26:1602–7.
55. Papaemmanuil E, Rapado I, Li Y, et al. RAG-mediated recombination is the predominant driver of oncogenic rearrangement in ETV6-RUNX1 acute lymphoblastic leukemia. Nat Genet. 2014;46:116–25.
56. Andersson AK, Ma J, Wang J, et al. Whole genome sequence analysis of 22 *MLL* rearranged infant acute lymphoblastic leukemias reveals remarkably few somatic mutations: a report from the St Jude Children's Research Hospital—Washington University Pediatric Cancer Genome Project. Blood. 2011;117(Suppl):69.
57. Roberts KG, Morin RD, Zhang J, et al. Genetic alterations activating kinase and cytokine receptor signaling in high-risk acute lymphoblastic leukemia. Cancer Cell. 2012;22:153–66.
58. Chen C, Bartenhagen C, Gombert M, et al. Next-generation-sequencing-based risk stratification and identification of new genes involved in structural and sequence variations in near haploid lymphoblastic leukemia. Genes Chromosomes Cancer. 2013;52:564–79.
59. Holmfeldt L, Wei L, Diaz-Flores E, et al. The genomic landscape of hypodiploid acute lymphoblastic leukemia. Nat Genet. 2013;45:242–52.
60. Chang VY, Basso G, Sakamoto KM, et al. Identification of somatic and germline mutations using whole exome sequencing of congenital acute lymphoblastic leukemia. BMC Cancer. 2013;13:55.
61. Zhang J, Mullighan CG, Harvey RC, et al. Key pathways are frequently mutated in high-risk childhood acute lymphoblastic leukemia: a report from the Children's Oncology Group. Blood. 2011;118:3080–7.
62. Mullighan CG, Zhang J, Kasper LH, et al. CREBBP mutations in relapsed acute lymphoblastic leukaemia. Nature. 2011;471:235–9.
63. Meyer JA, Wang J, Hogan LE, et al. Relapse-specific mutations in NT5C2 in childhood acute lymphoblastic leukemia. Nat Genet. 2013;45:290–4.
64. Mullighan CG. The molecular genetic makeup of acute lymphoblastic leukemia. Hematology Am Soc Hematol Educ Program. 2012;2012:389–96.
65. Mullighan CG. Genome sequencing of lymphoid malignancies. Blood. 2013;122:3899–907.
66. De Braekeleer E, Douet-Guilbert N, Le Bris MJ, et al. Gene expression profiling of adult t(4;11)(q21;q23)-associated acute lymphoblastic leukemia reveals a different signature from pediatric cases. Anticancer Res. 2012;32:3893–9.
67. Harrison CJ. Targeting signaling pathways in acute lymphoblastic leukemia: new insights. Hematology Am Soc Hematol Educ Program. 2013;2013:118–25.
68. Link DC. Molecular genetics of AML. Best Pract Res Clin Haematol. 2012;25:409–14.
69. Stratton MR, Campbell PJ, Futreal PA. The cancer genome. Nature. 2009;458:719–24.
70. Graubert TA, Mardis ER. Genomics of acute myeloid leukemia. Cancer J. 2011;17:487–91.
71. Sanders MA, Valk PJ. The evolving molecular genetic landscape in acute myeloid leukaemia. Curr Opin Hematol. 2013;20:79–85.
72. De Braekeleer M. Three steps mutation model of carcinogenesis. Med Hypotheses. 1984;14:363–71.
73. Murati A, Brecqueville M, Devillier R, et al. Myeloid malignancies: mutations, models and management. BMC Cancer. 2012;12:304.
74. Martelli MP, Sportoletti P, Tiacci E, et al. Mutational landscape of AML with normal cytogenetics: biological and clinical implications. Blood Rev. 2013;27:13–22.

75. Anderson K, Lutz C, van Delft FW, et al. Genetic variegation of clonal architecture and propagating cells in leukaemia. Nature. 2011;469:356–61.
76. Larsson CA, Cote G, Quintas-Cardama A. The changing mutational landscape of acute myeloid leukemia and myelodysplastic syndrome. Mol Cancer Res. 2013;11:815–27.
77. Rao AV, Smith BD. Are results of targeted gene sequencing ready to be used for clinical decision making for patients with acute myelogenous leukemia? Curr Hematol Malig Rep. 2013;8:149–55.
78. White BS, DiPersio JF. Genomic tools in acute myeloid leukemia: from the bench to the bedside. Cancer. 2014;120:1134–44.
79. Borate U, Absher D, Erba HP, et al. Potential of whole-genome sequencing for determining risk and personalizing therapy: focus on AML. Expert Rev Anticancer Ther. 2012;12:1289–97.
80. Link DC, Schuettpelz LG, Shen D, et al. Identification of a novel TP53 cancer susceptibility mutation through whole-genome sequencing of a patient with therapy-related AML. JAMA. 2011;305:1568–76.
81. Shah S, Schrader KA, Waanders E, et al. A recurrent germline PAX5 mutation confers susceptibility to pre-B cell acute lymphoblastic leukemia. Nat Genet. 2013;45:1226–31.
82. Welch JS, Westervelt P, Ding L, et al. Use of whole-genome sequencing to diagnose a cryptic fusion oncogene. JAMA. 2011;305:1577–84.
83. Walter MJ, Graubert TA, DiPersio JF, et al. Next-generation sequencing of cancer genomes: back to the future. Per Med. 2009;6:653.
84. Fisher R, Pusztai L, Swanton C. Cancer heterogeneity: implications for targeted therapeutics. Br J Cancer. 2013;108:479–85.
85. Grimwade D. The changing paradigm of prognostic factors in acute myeloid leukaemia. Best Pract Res Clin Haematol. 2012;25:419–25.
86. Rowe JM. The impact of mutational profiling on AML prognosis. Best Pract Res Clin Haematol. 2012;25:403–8.

Next-Generation Sequencing Applications in Head and Neck Oncology

Camile S. Farah, Maryam Jessri, Farzaneh Kordbacheh, Nigel C. Bennett, and Andrew Dalley

Abstract Head and neck cancer remains a major medical problem with significant morbidity, mortality and quality of life issues. Over the recent past there has been an increase in incidence, a shift in etiological factors, a growing proportion of tumours in younger cohorts, and a greater realisation of the heterogeneity of this group of tumours particularly within head and neck squamous cell carcinomas.

The arrival of high-throughput massively parallel sequencing technologies in diagnostic laboratories heralds an opportunity for uncovering driver mutations in head and neck cancer, understanding of disease stratification, personalisation of treatment strategies within the framework of genomic medicine, and discovery of potential druggable targets for disease-specific treatment.

Next-generation sequencing (NGS) is a powerful tool and has the potential to transform the reactive and treatment-based nature of cancer care, to actively predict the risk for disease and aim to prevent it. The underlying goal of NGS application is to achieve the concept of "genome-informed personalised medicine". An important factor in harnessing NGS technologies in personalised management of head and neck oncology lies in the feedback between scientists and clinicians involved in cancer care. A genuine diagnosis and appropriate aetiology-matched treatment is only possible if our decisions are based on both the genotype and phenotype of our patients.

C.S. Farah (✉)
UQ Centre for Clinical Research, The University of Queensland,
Herston, QLD 4029, Australia

The Australian Centre for Oral Oncology Research & Education, School of Dentistry,
University of Western Australia, 17 Monash Ave, Nedlands , WA 6009, Australia
e-mail: camile@oralmedpath.com.au

M. Jessri • F. Kordbacheh • N.C. Bennett • A. Dalley
UQ Centre for Clinical Research, The University of Queensland,
Herston, QLD 4029, Australia

W. Wu, H. Choudhry (eds.), *Next Generation Sequencing in Cancer Research, Volume 2*, DOI 10.1007/978-3-319-15811-2_23

1 Introduction

Cancer remains a leading cause of mortality and morbidity. In the United States, cancer is estimated to be the first cause of death of those between the ages of 40 and 79, with 1 in every 4 deaths predicted to be directly related to cancer [1]. Siegel et al. have estimated that in the United States in 2014, 73,240 new cases of head and neck cancer (including tongue, mouth, pharynx, other oral cavity, larynx and oesophagus) will be diagnosed, and 27,450 deaths would occur due to head and neck cancer (HNC) [1]. According to an Australian Institute of Health and Welfare report, there were 4,134 new cases of HNC in 2010, with an age-standardised mortality rate of 4.3 per 100,000 persons [2], where cancer of the lip, oral cavity and oropharynx represents 2.9 % of the total cancer load in Australia and 1.6 % of all cancer deaths [3].

Head and neck squamous cell carcinoma (HNSCC) comprises a great majority of HNC cases. Environmental risk factors, namely tobacco use and alcohol consumption have been suggested to contribute significantly to carcinogenesis of HNSCC [4]. HNSCC is traditionally considered to arise in older populations; however during recent years, a growing proportion of younger patients with poor prognosis and a distinctive clinical and histopathological pattern have been diagnosed with HNSCC [5, 6]. The increasing incidence of HNSCC in younger patients could be partially explained by a shift in social behaviours and the role of genetic factors contributing to carcinogenesis of these tumours [7]. As a temporal summation of immunological, biochemical and molecular changes which may or may not be triggered by environmental factors, HNSCC is a tumour with a highly heterogeneous nature [8].

Normal oral mucosa and oral potentially malignant lesions (OPML) which may precede malignancy in the oral cavity have been shown to have a low (less than 7 %) prevalence of infection with human papillomavirus (HPV)[9, 10], in comparison with the relatively higher rate of 35 % seen in oropharyngeal SCC [11]. Owing to differences in sampling, laboratory methods for detection of HPV, geographic location, ethnicity, sample size and most importantly inconsistencies in grouping of lesions in different anatomical regions of the upper aerodigestive tract, prevalence of HPV infection has been reported between 0 and 100 % in HNSCC [10, 12].

Regardless of the diagnostic method, genetic characteristics and prognosis of HPV-positive and negative HNSCC patients have been shown to significantly differ [13]. In addition, the heterogeneous nature of this cancer which is reflected in prognosis and recurrence rates [5] warrants the need for refinement of the current tumour classification system. A heterogeneous tumour with an ever changing profile necessitates clinicians to consider and evaluate each case with special attention to an individual's circumstances. Careful consideration of an individual's unique clinical, environmental and genetic profile prior to formulating a treatment plan is referred to as personalised medicine. Until recently, the genetics of a patient was only a contributing factor when the phenotype of an individual was affected. Progress made in our technical abilities coupled with improvements in our knowledge of the

human genome has facilitated involvement of the individual's genotype in clinical decision making. In light of these advancements, modification of cancer care solely based on patient clinical features no longer constitutes best practice.

A main contributor to the fast paced transformation of cancer care is the large quantities of data produced by high-throughput molecular profiling technologies. Next-generation sequencing (NGS) technologies, otherwise known as massively parallel sequencing platforms, produce a large amount of data in a very short period of time. To a great extent, these platforms have been used to answer fundamental research questions about the genetics and pathogenesis of various diseases [14]. First commercialised in 2005, NGS platforms have changed the face of cancer research as we know it for ever.

2 Applications of Next-Generation Sequencing in Head and Neck Oncology

The first two reports of application of NGS in studying HNSCC were published simultaneously in *Science* in August 2011 [15, 16]. To date, they remain the most comprehensive studies of the HNSCC exome through NGS. Stransky et al. sequenced the whole exome of 74 HNSCC tumours and their matched normal in addition to the whole exome of an oropharyngeal and a hypopharyngeal tumour to study the mutational profile of these cancers. They found an average of 130 mutations per tumour of which 25 % were synonymous. When spectrometric genotyping was used to query these mutations, 89.75 % (288 out of 321 mutations were validated). Although a widely varied number of base mutations was reported in this study, the general mutation rate and the nature of mutations in non-CpG sites were similar to that of other smoking-related malignancies such as small-cell lung cancer and lung adenocarcinoma [17, 18]. A higher number of mutations was reported in HPV negative tumours which was more apparent [15]. Similar findings were reported by Agrawal et al. from whole-exome sequencing of 32 HNSCC tumours with neoplastic cellularity of over 60 % and their matched non-neoplastic tissue [16]. Although the general coverage of the exome was lower than that of Stransky et al. [15], they also found HPV negative tumours to have higher mutation rates, and the difference was independent of patients' smoking status [16]. Although both studies found a higher number of mutations in smokers, unlike Stransky et al. [15], Agrawal did not find the mutation profile to be similar to that of smoking related tumours which show G:C>T:A transversion enrichment [16]. Despite their differences in the number of samples, platforms used and the general mutation trends reported, both studies found TP53 to harbour the highest number of mutations in HPV negative tumours. In addition both studies found frequent mutations in cyclin-dependent kinase inhibitor 2A (CDKN2A), phosphatidylinositol-4,5-bisphosphate 3-kinase catalytic subunit alpha (PIK3CA) and Notch homolog 1 translocation-associated (NOTCH1) to be a common finding in HNSCC tumours [15, 16].

The role and importance of TP53 mutations in carcinogenesis of HNSCC was well established prior to its validation through NGS [19–23], although to a lesser extent compared to TP53, the importance of CDKN2A and PIK3CA was also shown in HNSCC carcinogenesis prior to 2011 [24, 25]. Contrarily, although immunoexpression of NOTCH1 had been shown to correlate with cisplatin resistance in HNSCC [26], it was not until these two NGS studies of the HNSCC exome that NOTCH1 was considered a potential contributor to pathogenesis of these tumours [15, 16]. Brakenhoff emphasised this finding by an essay published in the cancer section of *Perspectives on Science* in September 2011[27]. Undoubtedly discovery of mutations in NOTCH1 that result in truncated proteins in HNSCC tumours remains the most discussed finding of parallel sequencing of HNSCC to date. Song et al. sequenced the whole coding region and intron–exon boundaries of TP53 and NOTCH1 in 13 HNSCC cell lines and 51 oral squamous cell carcinoma (OSCC) tumours [28] and found 43 % of the tumours to harbour non-synonymous mutations in NOTCH1 with only 7 % of the mutations showing a G:C>T:A transversion [28]. Although the number of mutations in TP53 was reported to be approximately the same as NOTCH1(non-synonymous SNP in 41 % of tumours), the G:C>T:A transversion rate was approximately threefold higher for TP53 [28].

Recently, NOTCH1 was sequenced in 37 HNSCC tumours and their matched lymphocytes; following the removal of putative germ-line variants or sample specific systematic error (defined as the variants present in both tumour and matched normal), a total of 5 NOTCH1 mutations were reported in 4 tumours [29]. Following a similar concept, the targeted exon of 51 highly actionable cancer-related genes from 37 primary HNSCC tumours and their matched lymphocytes were deep-sequenced [29]. Expectedly, with 13 mutations in 11 patients (29.7 % of the studied population), TP53 was the highest mutated gene in HNSCC samples; in addition NOTCH1 exomic alterations were reported in 8.1 % of the samples [29].

3 Translational Research

In studying and understanding NGS applications in treatment and care of HNSCC, there are two important observations to make: (a) there was a 5-year gap between commercialisation of NGS platforms and publication of the first [two] articles that used these platforms for studying the genetics of HNSCC and (b) since 2011, to our knowledge, application of NGS in analysis of the HNSCC genome has been limited to target enriched areas of the human exome. These two observations can be partly explained in light of the inherent limitations of NGS; NGS is costly and time consuming, requires highly skilled technicians, and the final result is difficult to interpret. To these limitations one should add the ethical dilemma of handling NGS information, small number of samples that qualify for inclusion in these studies, and the high attrition rate of samples due to technical difficulties faced during library preparation, sequencing and bioinformatics analysis. However, acknowledging these limitations should not distract from the potential of NGS to further the field of head and neck oncology.

Translational research can be defined as an engineering research question which aims to make findings beneficial and relevant to improving human health and well-being. This concept is particularly important in conducting research that is demanding of high level skills and scarce resources. Needless to note that this is inclusive of the feedback provided by clinicians and those who implement research findings in the "real world". This loop helps research to stay relevant and clinical practice to be as effective as possible.

By this definition, designing the future questions to be addressed by NGS seems to be influenced by the heterogeneous nature of HNSCC and the need to address both the biology and anatomy of this group of tumours. Translational research necessitates a focus on traditional clinical principles. Early diagnosis is the key to management of head and neck cancers [30]. The importance of improving diagnostic markers, screening of OPMLs and the discriminative factors that determine their malignant transformation are well justified in this regard. The heterogeneous nature of HNSCC has long been overlooked in classification of these tumours. The assumed polyclonal HNSCC means that the best course of action for one patient may prove harmful to another, and patients with similar clinical manifestations may have variable prognoses.

Head and neck oncology is currently at a turning point; although limitations of NGS are reflected within targeted sequencing of candidate genes as opposed to a more thorough approach to the cancer genome, we have reached a point where sequencing the whole exome is cost-effective. Our understanding of carcinogenesis of HNSCC has the potential to transform the idealised world of personalised medicine into a reality.

4 NGS in Personalised Medicine

Although intriguing, understanding the genetic mechanisms contributing to development and progression of disease is not the final answer to oncology questions. From a clinical viewpoint, the ultimate goal should be prevention and improvement of therapeutic outcomes for patients. At this stage, scarce resources, technical considerations and high cost have limited NGS to identification of highly mutated genes in different types of cancer, including HNSCC. Diagnostic laboratories still rely on low-throughput techniques such as Sanger sequencing, real-time PCR and fragment analysis for mutation detection and analysis. With improvement of knowledge of cancer through NGS, the number of actionable or druggable targets for each cancer increases. As Berger concluded, current routine low-throughput methods complicate the workflow for diagnosis of tumours with limited tissue [31]. Multiplexed mutational technologies such as Sequenom (CA, USA) and PCR based assays are considered an improvement; yet their throughput is negligible compared to that of NGS platforms. However despite their enormous promise, NGS platforms have certain limitations; the high sensitivity of these platforms entails the majority

of the reported mutations to be limited in only a small number of tumours or a small fraction of the tumour [32]. This, in addition to the dynamic nature of most cancers has brought forth the need for personalised use of NGS platforms in the care of HNSCC patients in order to manage the physiological effects of each genomic alteration [33].

As the most extreme result of accumulation of genetic alterations, it comes as no surprise that only the HNSCC exome has been subject to NGS investigation, and to our knowledge, genomic alterations of OPMLs have not been studied in any detail. However as discussed above, personalised medicine in cancer care is at its best if it can prevent onset of cancer and assist with early diagnosis. Information regarding progression of OPML to cancer would help clinicians tailor management of these lesions for each patient.

Personalised treatment of cancer through NGS studies of the cancer genome has already been adopted in non-small-cell lung cancer with epidermal growth factor receptor (EGFR) deletions. Evidence points to improvement of patients with mutations in EGFR by targeting the corresponding genetic alteration [34]. However systematic pharmacogenomic analysis of a whole genome from a Malay male volunteer for targeted therapy remains the best example of personalised care using NGS [35]. A total of 3,375 genes involved in the metabolism and transformation of 6,707 drugs/therapeutic agents were mapped and their potential effect on transport, metabolism and drug end targets were included [35].

Although there are examples of application of NGS in individualising treatment of cancer patients, their true functionality in clinical oncology is still progressing slowly. Nonetheless, NGS platforms are revolutionising genome analysis and an important improvement incorporated into their design is the attempt to make them more user-friendly and facilitate data analysis. We are nearing a new era in head and neck oncology when designing personalised treatments for each patient based on their comprehensive molecular profile is becoming more clinically feasible. However identification of clinically relevant or "driver" mutations in a heterogeneous tumour with high mutation rates is challenging. Overlying multiple complementary data sets is amongst the approaches taken to overcome this problem. Before NGS platforms can find their place in routine patient care, issues such as overall cost, clinical utility, relevance, bioinformatics analysis and timeline requirements need to be resolved.

Although the massive throughput of these platforms is a potential benefit to personalising care, in real world scenarios it could act as a hindrance. Extensive processing of NGS generated data is required to make biological and clinical sense, as the computational skill needed for deciphering the raw data is beyond the scope of clinicians and molecular biology researchers. In addition, despite the precision of these platforms and complicated algorithms involved in processing NGS data, the detected mutations need to be further validated.

5 Validation of Targets

Mutations, be they (single nucleotide polymorphisms, insertions, deletions, copy number variations, gene fusions, chromosomal rearrangements), whether germ line or somatic, found by NGS technologies in cancer samples should be compared to a clinical control sample originating from the same patient where possible. In clinical cancer pharmacogenetics trials, germ-line DNA samples are typically collected from blood or buccal mucosa [36]. For HNC, patient control samples (reference DNA) must be collected via blood as buccal mucosa samples are unsuitable due to field cancerization [37]. If patient control samples are not available, a reference sequence such as 1000 Genomes Project, dbSNP, HapMap Project, NHLBI Grand Opportunity Exome Sequencing Project, GATK Resource Bundle, NCBI RefSeq, NCBI ClinVar is needed [38].

Considering that thousands of variants can be discovered from a relatively small panel of targeted genes [39], confirmation of all is not clinically plausible nor financially realistic. Confirmation also depends on the specific NGS assay (single vs multiple gene panel) and perceived clinical relevance of specific variations highlighted in NGS data. Regardless of whether the genetic foundation of disease is monogenic (e.g. cystic fibrosis) or polygenic (e.g. cancer), at present the gold standard for variant confirmation is done using Sanger sequencing. This approach negates the significant number of false-positives currently generated by NGS [40]. An alternative NGS platform can be utilised for confirmation but financial constraints will most-likely shape this decision. In future, Sanger sequencing or alternative platform confirmation processes may be eliminated as NGS technologies improve, error rates decrease and as quality control and proficiency testing systems are implemented and made mandatory. Improvement of current variant-calling algorithms and pipelines are also critical in this arena. A variant, highlighted as being a potential cancer driving variant (pCDV), is more likely to be correctly identified if called by multiple algorithms than any one caller [41]. Using multiple variant caller algorithms is a broadly used approach which attempts to mitigate false negative and/or false positive errors [42, 43]. Although optimal combinations of variant calling algorithms with respect to specific variant types are not currently known, they would make a welcomed addition to the field [41].

Nucleic acid sequence variations can potentially be targeted or labelled as a disease biomarker. Variation exists in the genome (sequence variant, structural alteration), the transcriptome (splice variant, relative expression) and the methylome (variable methylation signature). Irrespective of their nature, biomarkers need to be readily detectable (with robust technology and validated methods), correlate with a specific tumour or clinical response (have clinical validity and functional relevance), and be exploitable for improved patient survival (have clinical utility via therapeutics) [44]. It has been surmised that sequence variations in cancer can be loosely categorised as either “drivers” (i.e. cancer causing) or “passengers” (i.e. secondary mutations caused by genome instability) [45]. Nonetheless, disease progression and treatment regimens may alter the roles of individual variants [46].

Further sub-classification of "drivers" into a deleterious impact scale (i.e. most important target/s), combined with confidence of clinically relevant interpretation, and application of known druggable targets confound the true value of NGS data. Variant confirmation is important and complex but it does not exist in isolation. For reporting purposes all observed pCDVs in processed data should be declared, but only some observed pCDVs will progress to the confirmation phase while others will be eliminated. Justifying the selection process is difficult when considering the potential clinical impact on the cancer patient. Nonetheless, as confidence within variant-calling algorithms improves interpretation, and increased knowledge of variant impact improves target allocation, coupled with expansion of the pool of therapeutic utilities, difficulties associated with this process will most likely subside.

6 Cancer Druggable Genome

The druggable genome is defined as the altered genes or gene products that can interact with molecules containing therapeutic properties [47, 48]. Genomic alterations include gene deletions and amplifications that change the abundance of the gene and its downstream products, alternative splicing or translocations that can create novel proteins, and mutations such as single base changes that may modify protein activity [49, 50]. Passenger mutations, which are the product of carcinogens and genomic instability do not appear to be involved in cancer progression, while driver mutations are known to contribute to tumour development or progression [51–53]. The distinction between driver and passenger mutations can change throughout the course of disease [54]. Driver mutations and their associated cellular pathways may have significant diagnostic, prognostic, or therapeutic implications and are branded as "actionable". A subset of actionable events may also be "druggable" which designates them as valuable targets for therapeutic development [53, 55]

In the past decade, massively parallel sequencing has enabled unbiased cancer genome sequencing in order to search and screen for new cancer genes at an unprecedented rate and scale. NGS technologies have been widely implemented for de novo whole genome, exome and transcriptome sequencing for assessment of DNA copy number, re-arrangements, loss of heterozygosity, allele specific amplification, methylation, transcription, aberrant splicing and RNA editing at rates that are dramatically faster and more cost-effective than traditional methods including Sanger-based sequencing [44, 46, 56].

To date, studies of driver mutations in cancer genes with protein altering mutations have yielded partial cancer genome data sets such as The Cancer Genome Atlas (TCGA, http://cancergenome.nih.gov) and the International Cancer Genome Consortium (ICGC, http://www.icgc.org) that are available for several types of cancers for more than 7,500 genes [56, 57]. The availability of these large cancer data sets has led to the validation of many new driver mutations, clinically useful

biomarkers and drug developments. Most of these known cancer genes were found through primary cytogenetic analyses; however, introducing systematic sequencing of cancer genomes has revealed new cancer genes including BRAF, EGFR, ERBB2, PIK3CA, PPP2R1A and JAK2 [58–63].

NGS has begun to replace existing Sanger sequencing, microarrays and PCR-based assays for cancer genome research. Not only has applying NGS in cancer genomic research revealed new cancer genes, but has also provided further insight of intra-tumour heterogeneity [54]. Recently a number of targeted therapies have become available for various cancers including melanoma, breast and lung cancer.

In 2011, Wagle et al. used a targeted resequencing approach to identify a previously unknown mechanism of resistance to the BRAF inhibitor, vemurafinib, in melanoma [64]. Analysis of 138 cancer genes in tumour samples from a single patient before and after relapse revealed the presence of a p. Cys121Ser mutation in MEK1 kinase in the relapse sample only. This was the first report of an activating mutation occurring downstream of BRAF kinase, adding to a growing list of other known mechanisms of acquired resistance to BRAF inhibition [65–67].

New studies of whole-genome analysis of breast cancer patients has resulted in new findings in copy-number variations, new descriptions of driver and other mutations, and elevated mutation rates in treatment-resistant tumours [68]. In 2012, Banerji et al. showed recurrent somatic mutations in five genes (PIK3CA, TP53, AKT1, GATA3 and MAP3K1), and new recurrent mutations and deletions were discovered for CBFB and RUNX1 [69]. Other studies have revealed that only 36 % of gene mutations are detected as transcribed [70] and that many mutations encode truncated proteins [71]. It would seem unlikely that these novel findings would have been discovered using conventional sequencing or genotyping approaches.

7 Druggable Targets for Head and Neck Cancer

Despite the numerous and growing number of studies on cancer biomarkers, only a few reports are available for head and neck cancer biomarkers and therapies. HNSCC has been considered an environmental tumour mostly caused by tobacco and alcohol consumption and more recently human papilloma virus (HPV) infection in Western populations [4]. However, a growing proportion of younger low-risk patients with a poor prognosis and a distinctive clinical and histopathological pattern [5] [6] in addition to genetic and prognostic differences between HPV-positive and -negative HNSCC [13] as well as the heterogeneous nature of this cancer, suggests a critical role for genetic alterations contributing to the carcinogenesis of HNSCC tumours [7].

Prior to utilisation of NGS in HNSCC, studies on cellular signalling pathway alterations (i.e. TP53 and CDK2NA) [72] and chromosomal abnormalities (amplification of region 11q13, cyclin D1 gene and region 7p11, EGFR gene) [73] had revealed a few HNSCC biomarkers. In 2011, the first reports of applying NGS to HNSCC were published simultaneously [15, 16], and both confirmed the previously

known HNSCC genome alterations such as mutations in TP53, CDKN2A, PIK3CA, PTEN and HRAS as well as introducing a new gene, NOTCH1 (an important tumour suppressor gene), which has been shown to be the second most common gene involved in HNSCC [74]. Although NOTCH1 has been shown to be important in skin squamous cell carcinoma through functional studies [75], it had not been identified by classic (Sanger) sequencing techniques due to its large size (34 coding exons).

Taking a more targeted approach, Mahjabeen et al. sequenced 17 exons of XRCC1 in 300 head and neck cancer cases and 150 matched normal controls and found two silent mutations in 45 % and two missense mutations in 55 % of cases, accounting for a total mutation frequency of 87 %. Both silent mutations were distributed equally among males and females and smokers vs. non-smokers [76]. In another study, sequencing all exons and adjacent introns of RAD51C revealed five distinct heterozygous sequence alterations in 5.8 % of HNSCC cases [77]. Moreover, Laborde et al. performed whole-transcriptome sequencing of ten matched tumour and cancer-free tissue samples from patients with previously untreated oropharyngeal carcinoma. Their results showed elevated levels of expression for two gene targets (CHEK2 and ATR) in p53 DNA damage repair pathway in HPV-negative current smoker patients compared with past smokers or non-smokers [78].

So far, among all the known frequently mutated genes in head and neck cancers, EGFR has been a good candidate for developing cancer therapies. Over-expression and mutation of EGFR has been associated with a variety of human tumours including breast, lung, colorectal, ovary and prostate [79–81]. EGFR is a transmembrane receptor belonging to a family of four related proteins, human epidermal receptor (HER) family of growth factor receptors (HER2, HER3 and HER4). Formation of either EGFR-EGFR homodimers or heterodimers (i.e. EGFR-HER2) trigger a series of intracellular pathways that may result in cancer cell proliferation, blocking apoptosis, activation of invasion and metastasis, and stimulating tumour-induced neovascularisation. [82, 83]. Two classes of anti-EGFR drugs including anti-EGFR monoclonal antibodies and EGFR tyrosine kinase inhibitors have been developed since 1980.

Anti-EGFR monoclonal antibodies, such as cetuximab and panitumumab, recognise EGFR exclusively and bind to its extracellular domain in the inactive configuration, compete for receptor binding and block ligand-induced EGFR tyrosine kinase activation. Small-molecule EGFR tyrosine kinase inhibitors, such as erlotinib and gefitinib, compete reversibly with ATP to bind to the intracellular catalytic domain of EGFR tyrosine kinase and, thus, inhibit EGFR autophosphorylation and downstream signalling. Moreover, small-molecule EGFR tyrosine kinase inhibitors can block different growth factor receptor tyrosine kinases, including other members of the EGFR family, or the vascular endothelial growth factor receptor. In February 2006, FDA approved cetuximab to be used in combination with radiotherapy to treat HNSCC patients with locally advanced, unresectable tumours. It was also approved as mono-therapy for metastatic disease in patients who have not had a response to chemotherapy. In March 2006, the EMEA approved cetuximab in combination with radiotherapy for the treatment of locally advanced disease.

Head and neck cancers and their treatment can result in cosmetic deformity with impairment of vital functions such as breathing, taste, swallowing, speech, hearing and smell. Patients are often diagnosed at advanced stages (stage IV) with serious lymph node involvement. Therefore, the diagnosis of HNC at pre- and early cancerous stages has become vitally important. To date, the optimal treatment for HNSCC patients involves a multidisciplinary approach including coordination of surgery, chemotherapy, radiation therapy and systemic therapies such as anti-EGFR molecules. However, the current targeted treatments have benefitted only a small subgroup of cancer patients due to intrinsic (primary) and extrinsic (secondary) drug resistance with limited drug efficacy [84] which decreases with long-term follow-up [85]. Therefore, identification of new molecular targets and novel therapies, as well as selecting patients for existing commercial drugs calls for new methods of identification and clinical validation of biomarkers.

Binding of EGFR to its ligands triggers two main signalling pathways: PI3K-Akt and MAPK/ERK. These downstream signalling pathways can be switched on by mutations in intermediate molecules. Mellinghoff et al. showed that only 10–20 % of glioblastoma patients are responsive to EGFR kinase inhibitor treatments which is associated with co-expression of EGFR vIII and PTEN [80]. In addition, activating mutations in KRAS results in EGFR-independent signalling pathway activation and has been seen in almost 15–30 % and 40–45 % of patients diagnosed with small-cell lung cancer and colorectal cancer, respectively, all with a history of tobacco use [86, 87]. These patients showed resistance or limited efficacy to cetuximab and panitumumab therapy [88, 89]. MET amplification also leads to EGFR-independent activation of PI3K-AKT pathway through activation of HER3-dependent pathway. It has been shown that inhibition of MET signalling can restore the sensitivity of lung cancer cell lines to gefitinib [90].

Some actionable mutations are altered in several common cancers, and it is reasonable to assume that a therapy effective for one may indeed be useful for another. BRAF V600E is a gene marker for vemurafenib therapy in melanoma [91], which could be used for similar mutations detected in ovarian cancer [92]. On the other hand, a specific genetic abnormality may not confer the same sensitivity to an agent across all cancers. As a case in point, trastuzumab has been shown to benefit breast and gastric cancer patients with HER2-amplification, but not those with ovarian or endometrial cancer [93–95].

In recent years, NGS has been bridging the gap between molecular screening and clinical applications with 96.1 % accuracy in comparison to Sanger sequencing. In addition, it can reveal gene alterations at very low allelic frequencies [96]. Not only can NGS identify new altered genes for new biomarker development [15, 16], but by revealing specific gene alterations it may help to identify patients whose cancers are either sensitive or resistant to a certain therapy [97]. Moreover, tumour-specific alterations could be detected in patient plasma. Put together, NGS can be used to personalise disease monitoring in clinical practice.

8 Possible Future Therapies

To date, cetuximab is the only molecular targeted therapy approved by FDA for HNSCC patients. However, many factors have been identified to be associated with resistance to EGFR targeted therapies. As a case in point, EGFRvIII is a constitutively active ligand independent EGFR that has been observed in about 40 % of HNSCC [98] and has been shown to lower the effect of cetuximab and cisplatin on tumour responsiveness [98, 99].

Alternative EGFR activation pathways such as HER2, HER3, Auroa (a family of protein kinases that play a critical role in the mitotic process) and MET have been implicated in cetuximab resistance in head and neck cancers [100–102]. Regarding EGFR targeted therapies, although, panitumumab is currently being investigated in a Phase III trial (SPECTRUM) and positive outcomes may replace it with cetuximab in a few years, clinical development of TKIs have been less promising as gefitinib is no longer in active development for HNSCC and erlotinib is not being further studied in a Phase III trial. As a biological approach, targeting other regulatory pathways including angiogenesis using FDA approved agents such as bevacizumab, which is currently under Phase III investigation in HNSCC, might find use as monotherapy or combined targeted chemotherapy. Therefore, development of prognostic tests, discovering new biomarkers, elucidation of mechanisms of resistance to targeted therapies and associated drug side effects are matters that need further research.

Not only are different genetic aberrations involved in head and neck tumours, but also various environmental factors including clinical and epidemiological effects such as HPV infection and tobacco and alcohol exposure have been associated with HNSCC tumours. Moreover, HNSCC is a high heterogeneous tumour and therefore individual management should be based on both patient and tumour characteristics. In the last decade, NGS has allowed the identification of numerous genes in a time- and cost-effective manner. Additionally, the availability of NGS investigated genomic alterations collected by the International Cancer Genome Consortium (ICGC) (https://dcc.icgc.org) [103] and The Cancer Genome Atlas (TCGA) (http://cancergenome.nih.gov) projects in addition to Catalog of Somatic Mutations in Cancer (COSMIC) database (Welcome trust Sanger Institute) (http://cancer.sanger.ac.uk/cancergenome/projects/cosmic) provide the most comprehensive source of somatic mutations in cancer. However, like all other techniques, NGS has been linked to both challenges and promises. The challenges involve the scarcity of samples, collection of high quality test samples (e.g. FFPE and contamination with other cell types) [54] as well as their matching negative/normal controls, disease heterogeneity and the complexity and influence of the epigenome. The importance of data interpretation which involves both computational analysis and making biological and clinical sense of an enormous amount of produced data cannot be neglected. Furthermore, not all mutations are driver mutations and the importance of loss of tumour suppressors as well as germ-line mutations and compensating pathways cannot be overlooked. In addition to utilising NGS as a rapid gene

screening tool, the forthcoming third generation of NGS platforms promise the ability of utilising a single tumour cell for such purposes, further extending the reach of NGS in clinical practice.

In addition to DNA sequencing including selective exome- and whole-genome sequencing, transcriptome sequencing allows the identification of unique genes and pathways being expressed as a result of patient exposure to environmental factors as well as revealing critically important genetic mutations and rearrangements reflected at the transcriptome level and can therefore be used as diagnostic and prognostic indicators. As an example, transcriptional profiling using mRNA sequencing from ten oropharyngeal cancer patients with SCC and matching normal samples, showed *TP53* mutation and *CHEK2* and *ATR* increased expression in HPV negative current smokers compared to past or non-smokers [78]. We are currently undertaking whole-transcriptome sequencing of a variety of pre-malignant and malignant oral lesions in order to identify new classes of druggable target genes as well as developing diagnostic tools for early head and neck cancer detection.

All together, the promises for NGS are tremendous. The underlying goal of NGS application is to achieve the concept of "genome-informed personalised medicine" where a family-based disease history, geographically and epigenetically variant gene expression pathways, and drug resistance and toxicity can be considered in formulating a patient management plan (see Fig. 1).

9 Ethical Considerations

Transition in application of NGS from a research tool to clinical settings necessitates further ethical considerations. The foremost question in application of NGS in clinical settings is based on patient selection. With new advances in massively parallel sequencing platforms and the attempts to facilitate the use and analysis of data, many research groups choose to use these platforms and as a result the volume of the produced data is overwhelming for clinicians. Clinicians of the future will be treating patients who are aware of the possibilities and options that NGS platforms offer, and they should decide when to benefit from these platforms in formulating a treatment plan. Furthermore, the decision on what percentage of the genome should be studied by NGS should be based on clinical need rather than the availability and the ever decreasing cost of whole-exome sequencing.

The second issue would be handling the data and the likely findings. In research settings when a cohort of volunteers are sequenced to answer a question, providing the participants with the research findings is both uncontroversial and a welcome practice [104]. Contrary to research data, the feedback on the findings of individualised sequencing for clinical purposes is more complex. This is more apparent in whole-exome/genome sequencing which provides clinicians with an unabridged set of data in regard to their patients. Still a highly debated issue, the safest approach should be taken applying the same standards used in MRI or CT studies where only

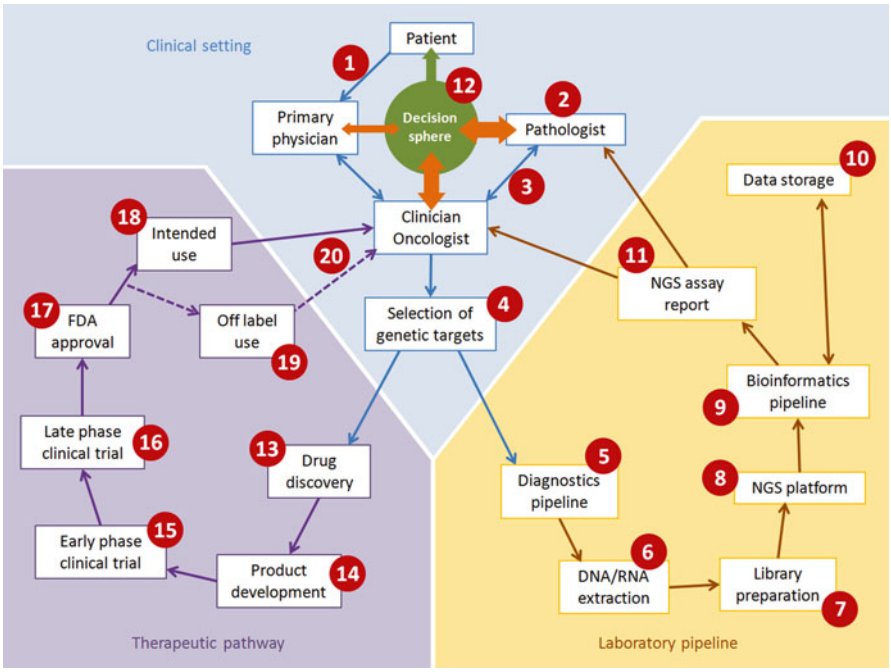


Fig. 1 Genome-informed personalised medicine in head and neck oncology: role of NGS pipeline in informing clinical decision making and therapeutic druggable pathway determination. *Clinical setting*: (1) Patient presents with primary tumour; (2) Primary biopsy assessed by pathologist; (3) Pathologist and oncologist consider possible treatment options; (4) Oncologist and/or pathologist select genetic targets for analysis. *Laboratory setting*: (5) Sample from primary biopsy sent to pathology/diagnostics laboratory; (6) Extraction process specific to nucleic acid targets, good laboratory practices and regulatory guidelines; (7) Library preparation guided by nature of extracted material and selected targets, good laboratory practices and regulatory guidelines; (8) NGS platform considerations related to target size and required coverage, sequence coverage relevant to platform data capabilities, error rates, read lengths; (9) Bioinformatics pipeline shaped by data source, data quality, specific targets, clinical purpose of test, good laboratory practices and regulatory guidelines; (10) NGS data storage requires large volumes of HDD space, data duplicated as a minimum, duplicated data stored at different sites, regular monitoring of storage systems and data integrity; (11) NGS report contains all non-synonomous variants whether known to be clinically relevant or not, citations of supporting peer-reviewed literature, analysis pipeline methods, statements explaining limitations of methods and clinical interpretations; (12) Physicians making therapeutic regimen decisions must consider the NGS report in combination with all other clinical and diagnostic information. *Drug development setting*: (13) Genetic analysis results in druggable target gene discovery; (14) Development of targeted drug; (15) Drug passes the laboratory test settings on in vitro and animal models; (16) Evaluation of safety and efficiency of the drug in late phase clinical trials; (17) Drug receives FDA approval for a certain treatment condition/disease; (18) The application of the drug for the same FDA approved condition/disease; (19) The administration of the drug to the patient with similar or relevant condition/disease to the original FDA approved agreement; (20) The oncologist recommends the drug as a possible treatment for either intended or off label use

reliable findings with significant relevance are shared with research participants [105, 106]. In clinical settings, however, the decision lies with clinicians and the appropriate ethical boards.

Data management and sharing is another important issue to be considered. Research groups tend to share their information and upload databases for public access. Although all personal information is normally removed from the data, joining the dots and identifying participants could potentially infringe the privacy of research participants. Moreover, despite precise regulations and rules that govern NGS data, scientists tend to share information through informal routes or deposition in cloud-based databases which can compromise participant privacy [107].

In general, fundamental ethical requirements not only for research but also the clinical aspect of NGS applications in cancer care should be re-evaluated. Obtaining "informed" consent, regulation of patients' access to their data, treatment plans and genomic profile are amongst the issues that need to be resolved before more pervasive application of NGS technologies occur in clinical oncology.

10 Concerns and Conclusions

The new generation of NGS platforms accommodate research and clinical needs with their higher accuracy, time-efficiency and cost-effectiveness. This assists with the ongoing attempt to profile different types of cancers best reflected in projects such as The Cancer Genome Atlas (http://cancergenome.nih.gov) and the International Cancer Genome Consortium (http://icgc.org). This is a new era for both scientists and clinicians: we have the ability to produce more than one billion base pairs in a 4-day run. At the same time interpretation of NGS data in order to make this relevant to clinical decision making requires sophisticated bioinformatics methods. Bioinformatics approaches still struggle with inevitable challenges such as stromal contamination of tumour samples and tumour heterogeneity. Needless to mention, high cost, demanding laboratory and computational skills, complexity of data output and scarcity of samples with high integrity nucleic acid currently limit the practical use of NGS in clinical scenarios.

Despite these limitations it is expected that in the near future, NGS will transform from a research focused technique to a more clinically oriented one. With more clinicians embracing the concepts of evidence-based and minimal intervention medicine [108], there is a need for removing the language barrier between research and clinical settings. The final product of bioinformatic analysis of NGS data is not straightforward to understand, and clinicians will need some assistance before they can use these findings in clinical practice (see Fig. 1).

In the world of oral health care, head and neck oncology is the main beneficiary of NGS findings by real world patients. Applying these findings in routine clinical practice requires overcoming a number of obstacles [46]. Clinicians need to understand the differences between "driver" and "passenger" mutations and acknowledge that identified mutations are not necessarily relevant to clinical manifestations.

In addition, identification of a mutation does not make it preventable nor definitely treatable as silencing mutations is a complex task which may even prove futile [27]. Silencing or altering mutations may interfere with crucial cellular pathways as the majority of cancer-related mutations are components of important cellular pathways [109]. Finally, any attempts to modify the human genome would put the patient at risk of resistance and toxicity [110].

We have learnt valuable lessons from the first forays into studying HNSCC by NGS; the most significant of which is that we are only beginning to understand the complex landscape of this heterogeneous group of malignancies. Although treatment regimens for this group of tumours vary based on anatomical location, the inevitable molecular differences introduced by lesion site is often still overlooked. In addition, racial and geographical differences in the human genome [111] necessitate extreme attention to detail in designing NGS studies particularly in regard to normal controls. Removing the common mutations between the tumour and matched normal sample, is helpful in removing false positive data but at the same time removes the chance of discovery of germ-line mutations.

NGS is a powerful tool and has the potential to transform the reactive and treatment-based nature of cancer care, to actively predict the risk for disease and aim to prevent it. To advance personalised care in head and neck oncology, there is a need for NGS platforms to continue to decrease in cost, while concurrently solving their technical shortcomings while there is an ongoing effort to connect the generated data to meaningful clinical findings. An important factor in harnessing NGS technologies in management of HNSCC lies in the feedback between scientists and clinicians involved in cancer care (see Fig. 1). A genuine diagnosis and appropriate aetiology-matched treatment are only possible if our decisions are based on both the genotype and phenotype of our patients.

Conflict of Interest Statement The authors have no conflict of interest to declare in relation to the work presented in this chapter. The authors are undertaking next-generation sequencing of oral cancer and oral potentially malignant lesions utilising SOLiD™ and Ion™ technologies, funded by grants held by author Camile S Farah awarded by the Queensland Government Smart Futures Co-Investment Fund and Cancer Australia, in collaboration with Life Technologies and Agilent Technologies.

References

1. Siegel R, Ma J, Zou Z, Jemal A. Cancer statistics, 2014. CA Cancer J Clin. 2014;64(1):9–29.
2. Australian Institute of Health and Welfare. Australian Cancer Incidence and Mortality (ACIM). Head and neck for Australia. Canberra ACT: Australian Institute of Health and Welfare; 2014.
3. Farah C, Simanovic B, Dost F. Oral cancer in Australia 1982-2008: a growing need for opportunistic screening and prevention. Aust Dent J. 2014;59:349–59.
4. Hashibe M, Brennan P, Benhamou S, Castellsague X, Chen C, Curado MP, Maso LD, Daudt AW, Fabianova E, Wünsch-Filho V, et al. Alcohol drinking in never users of tobacco, cigarette

smoking in never drinkers, and the risk of head and neck cancer: pooled analysis in the international head and neck cancer epidemiology consortium. J Natl Cancer Inst. 2007;99(10):777–89.
5. Chung CH, Parker JS, Karaca G, Wu J, Funkhouser WK, Moore D, Butterfoss D, Xiang D, Zanation A, Yin X, et al. Molecular classification of head and neck squamous cell carcinomas using patterns of gene expression. Cancer Cell. 2004;5(5):489–500.
6. Dahlstrom KR, Little JA, Zafereo ME, Lung M, Wei Q, Sturgis EM. Squamous cell carcinoma of the head and neck in never smoker-never drinkers: a descriptive epidemiologic study. Head Neck. 2008;30(1):75–84.
7. Liang XH, Lewis J, Foote R, Smith D, Kademani D. Prevalence and significance of human papillomavirus in oral tongue cancer: the mayo clinic experience. J Oral Maxillofac Surg. 2008;66(9):1875–80.
8. Baxi S, Fury M, Ganly I, Rao S, Pfister DG. Ten years of progress in head and neck cancers. J Natl Compr Canc Netw. 2012;10(7):806–10.
9. Tezal M. Interaction between chronic inflammation and oral HPV infection in the etiology of head and neck cancers. Int J Otolaryngol. 2012;2012:1–9.
10. Kansy K, Thiele O, Freier K. The role of human papillomavirus in oral squamous cell carcinoma: myth and reality. Oral Maxillofac Surg. 2012;18(2):165–72. Epub 16 Dec 2012.
11. Tang AL, Owen JH, Hauff SJ, Park JJ, Papagerakis S, Bradford CR, Carey TE, Prince ME. Head and neck cancer stem cells: the effect of HPV–an in vitro and mouse Study. Otolaryngol Head Neck Surg. 2013;149(2):252–60.
12. Feller L, Lemmer J. Oral leukoplakia as it relates to HPV infection: a review. Int J Dent. 2012;2012:1–7.
13. Chung CH, Gillison ML. Human papillomavirus in head and neck cancer: its role in pathogenesis and clinical implications. Clin Cancer Res. 2009;15(22):6758–62.
14. Feldman AL, Dogan A, Smith DI, Law ME, Ansell SM, Johnson SH, Porcher JC, Özsan N, Wieben ED, Eckloff BW, et al. Discovery of recurrent t(6;7)(p25.3;q32.3) translocations in ALK-negative anaplastic large cell lymphomas by massively parallel genomic sequencing. Blood. 2011;117(3):915–9.
15. Stransky N, Egloff AM, Tward AD, Kostic AD, Cibulskis K, Sivachenko A, Kryukov GV, Lawrence MS, Sougnez C, McKenna A, et al. The mutational landscape of head and neck squamous cell carcinoma. Science. 2011;333(6046):1157–60.
16. Agrawal N, Frederick MJ, Pickering CR, Bettegowda C, Chang K, Li RJ, Fakhry C, Xie TX, Zhang J, Wang J, et al. Exome sequencing of head and neck squamous cell carcinoma reveals inactivating mutations in NOTCH1. Science. 2011;333(6046):1154–7.
17. Lee W, Jiang Z, Liu J, Haverty PM, Guan Y, Stinson J, Yue P, Zhang Y, Pant KP, Bhatt D, et al. The mutation spectrum revealed by paired genome sequences from a lung cancer patient. Nature. 2010;465(7297):473–7.
18. Pleasance ED, Stephens PJ, O'Meara S, McBride DJ, Meynert A, Jones D, Lin ML, Beare D, Lau KW, Greenman C, et al. A small-cell lung cancer genome with complex signatures of tobacco exposure. Nature. 2010;463(7278):184–90.
19. Field JK, Spandidos DA, Malliri A, Gosney JR, Yiagnisis M, Stell PM. Elevated P53 expression correlates with a history of heavy smoking in squamous cell carcinoma of the head and neck. Br J Cancer. 1991;64(3):573–7.
20. Ögmundsdóttir HM, Hilmarsdóttir H, Björnsson J, Holbrook WP. Longitudinal study of TP53 mutations in eight patients with potentially malignant oral mucosal disorders. J Oral Pathol Med. 2009;38(9):716–21.
21. Tokman B, Gultekin SE, Sezer C, Alpar R. The expression of p53, p16 proteins and prevalence of apoptosis in oral squamous cell carcinoma. Correlation with mode of invasion grading system. Saudi Med J. 2004;25(12):1922–30.
22. Snyder LA, Bertone ER, Jakowski RM, Dooner MS, Jennings-Ritchie J, Moore AS. p53 expression and environmental tobacco smoke exposure in feline oral squamous cell carcinoma. Vet Pathol. 2004;41(3):209–14.

23. Simionescu C, Margaritescu C, Georgescu CV, Surpaţeanu M. HPV and p53 expression in dysplastic lesions and squamous carcinomas of the oral mucosa. Rom J Morphol Embryol. 2005;46(2):155–9.
24. O'Regan EM, Toner ME, Finn SP, Fan CY, Ring M, Hagmar B, Timon C, Smyth P, Cahill S, Flavin R, et al. p16INK4A genetic and epigenetic profiles differ in relation to age and site in head and neck squamous cell carcinomas. Hum Pathol. 2008;39(3):452–8.
25. Qiu W, Schonleben F, Li X, Ho DJ, Close LG, Manolidis S, Bennett BP, Su GH. PIK3CA mutations in head and neck squamous cell carcinoma. Clin Cancer Res. 2006;12(5):1441–6.
26. Gu F, Ma Y, Zhang Z, Zhao J, Kobayashi H, Zhang L, Fu L. Expression of Stat3 and Notch1 is associated with cisplatin resistance in head and neck squamous cell carcinoma. Oncol Rep. 2010;23(3):671–6.
27. Brakenhoff RH. Another NOTCH for cancer. Science. 2011;333(6046):1102–3.
28. Song X, Xia R, Li J, Long Z, Ren H, Chen W, Mao L. Common and complex notch1 mutations in Chinese oral squamous cell carcinoma. Clin Cancer Res. 2014;20(3):701–10.
29. Gaykalova DA, Mambo E, Choudhary A, Houghton J, Buddavarapu K, Sanford T, Darden W, Adai A, Hadd A, Latham G, et al. Novel insight into mutational landscape of head and neck squamous cell carcinoma. PLoS One. 2014;9(3):e93102.
30. Kulasinghe A, Perry C, Jovanovic L, Nelson C, Punyadeera C (2014) Circulating Tumour Cells in Metastatic Head and Neck Cancers. Int J Cancer. 2014 Aug 1. doi: 10.1002/ijc.29108.
31. Berger MF. Harnessing massively parallel DNA sequencing for the personalization of cancer management. Pers Med. 2013;10(2):183–90.
32. Mardis ER. Applying next-generation sequencing to pancreatic cancer treatment. Nat Rev Gastroenterol Hepatol. 2012;9(8):477–86.
33. Jones SJM, Laskin J, Li YY, Griffith OL, An J, Bilenky M, Butterfield YS, Cezard T, Chuah E, Corbett R, et al. Evolution of an adenocarcinoma in response to selection by targeted kinase inhibitors. Genome Biol. 2010;11(8):R82.
34. Guo G, Gui Y, Gao S, Tang A, Hu X, Huang Y, Jia W, Li Z, He M, Sun L, et al. Frequent mutations of genes encoding ubiquitin-mediated proteolysis pathway components in clear cell renal cell carcinoma. Nat Genet. 2012;44(1):17–9.
35. Salleh MZ, Teh LK, Lee LS, Ismet RI, Patowary A, Joshi K, Pasha A, Ahmed AZ, Janor RM, Hamzah AS, et al. Systematic pharmacogenomics analysis of a Malay whole genome: proof of concept for personalized medicine. PLoS One. 2013;8(8):e71554.
36. McWhinney SR, McLeod HL. Using germline genotype in cancer pharmacogenetic studies. Pharmacogenomics. 2009;10(3):489–93.
37. Slaughter DP, Southwick HW, Smejkal W. Field cancerization in oral stratified squamous epithelium; clinical implications of multicentric origin. Cancer. 1953;6(5):963–8.
38. Brownstein CA, Beggs AH, Homer N, Merriman B, Yu TW, Flannery KC, Dechene ET, Towne MC, Savage SK, Price EN, et al. An international effort towards developing standards for best practices in analysis, interpretation and reporting of clinical genome sequencing results in the CLARITY Challenge. Genome Biol. 2014;15(3):R53.
39. Salto-Tellez M, Gonzalez de Castro D. Next generation sequencing: a change of paradigm in molecular diagnostic validation. J Pathol. 2014;234(1):5–10.
40. Clark MJ, Chen R, Lam HY, Karczewski KJ, Chen R, Euskirchen G, Butte AJ, Snyder M. Performance comparison of exome DNA sequencing technologies. Nat Biotechnol. 2011;29(10):908–14.
41. Ding L, Wendl MC, McMichael JF, Raphael BJ. Expanding the computational toolbox for mining cancer genomes. Nat Rev Genet. 2014;15(8):556–70.
42. Kim SY, Speed TP. Comparing somatic mutation-callers: beyond Venn diagrams. BMC bioinformatics. 2013;14:189.
43. Goode DL, Hunter SM, Doyle MA, Ma T, Rowley SM, Choong D, Ryland GL, Campbell IG. A simple consensus approach improves somatic mutation prediction accuracy. Genome med. 2013;5(9):90.
44. Simon R, Roychowdhury S. Implementing personalized cancer genomics in clinical trials. Nat Rev Drug Discov. 2013;12(5):358–69.

45. Jessri M, Farah CS. Next generation sequencing and its application in deciphering head and neck cancer. Oral Oncol. 2014;50(4):247–53.
46. Jessri M, Farah CS. Harnessing massively parallel sequencing in personalized head and neck oncology. J Dent Res. 2014;93(5):437–44.
47. Hopkins AL, Groom CR. The druggable genome. Nat Rev Drug Discov. 2002;1(9):727–30.
48. Russ AP, Lampel S. The druggable genome: an update. Drug Discov Today. 2005;10(23–24):1607–10.
49. Stratton M. Evolution of the cancer genome. J Med Genet. 2011;48:S43.
50. Wong KM, Hudson TJ, McPherson JD. Unraveling the genetics of cancer: genome sequencing and beyond. Annu Rev Genomics Hum Genet. 2011;12:407–30.
51. Pleasance ED, Cheetham RK, Stephens PJ, McBride DJ, Humphray SJ, Greenman CD, Varela I, Lin ML, Ordonez GR, Bignell GR, et al. A comprehensive catalogue of somatic mutations from a human cancer genome. Nature. 2010;463(7278):191–U173.
52. Hanahan D, Weinberg RA. Hallmarks of cancer: the next generation. Cell. 2011;144(5):646–74.
53. Dancey JE, Bedard PL, Onetto N, Hudson TJ. The genetic basis for cancer treatment decisions. Cell. 2012;148(3):409–20.
54. Desmedt C, Voet T, Sotiriou C, Campbell PJ. Next-generation sequencing in breast cancer: first take home messages. Curr Opin Oncol. 2012;24(6):597–604.
55. Mwenifumbo JC, Marra MA. Cancer genome-sequencing study design. Nat Rev Genet. 2013;14(5):321–32.
56. Mills GB. An emerging toolkit for targeted cancer therapies. Genome Res. 2012;22(2):177–82.
57. Hudson TJ, Anderson W, Aretz A, Barker AD, Bell C, Bernabe RR, Bhan MK, Calvo F, Eerola I, Gerhard DS, et al. International network of cancer genome projects. Nature. 2010;464(7291):993–8.
58. Davies H, Bignell GR, Cox C, Stephens P, Edkins S, Clegg S, Teague J, Woffendin H, Garnett MJ, Bottomley W, et al. Mutations of the BRAF gene in human cancer. Nature. 2002;417(6892):949–54.
59. Paez JG, Janne PA, Lee JC, Tracy S, Greulich H, Gabriel S, Herman P, Kaye FJ, Lindeman N, Boggon TJ, et al. EGFR mutations in lung cancer: correlation with clinical response to gefitinib therapy. Science. 2004;304(5676):1497–500.
60. Samuels Y, Wang ZH, Bardelli A, Silliman N, Ptak J, Szabo S, Yan H, Gazdar A, Powell DM, Riggins GJ, et al. High frequency of mutations of the PIK3CA gene in human cancers. Science. 2004;304(5670):554.
61. Stephens P, Hunter C, Bignell G, Edkins S, Davies H, Teague J, Stevens C, O'Meara S, Smith R, Parker A, et al. Intragenic ERBB2 kinase mutations in tumours. Nature. 2004;431(7008):525–6.
62. Levine RL, Wadleigh M, Cools J, Ebert BL, Wernig G, Huntly BJP, Boggon TJ, Wlodarska L, Clark JJ, Moore S, et al. Activating mutation in the tyrosine kinase JAK2 in polycythemia vera, essential thrombocythemia, and myeloid metaplasia with myelofibrosis. Cancer Cell. 2005;7(4):387–97.
63. Jones S, Wang TL, Shih IM, Mao TL, Nakayama K, Roden R, Glas R, Slamon D, Diaz LA, Vogelstein B, et al. Frequent mutations of chromatin remodeling gene arid1a in ovarian clear cell carcinoma. Science. 2010;330(6001):228–31.
64. Wagle N, Emery C, Berger MF, Davis MJ, Sawyer A, Pochanard P, Kehoe SM, Johannessen CM, MacConaill LE, Hahn WC, et al. Dissecting therapeutic resistance to RAF inhibition in melanoma by tumor genomic profiling. J Clin Oncol. 2011;29(22):3085–96.
65. Johannessen CM, Boehm JS, Kim SY, Thomas SR, Wardwell L, Johnson LA, Emery CM, Stransky N, Cogdill AP, Barretina J, et al. COT drives resistance to RAF inhibition through MAP kinase pathway reactivation. Nature. 2010;468(7326):968–U370.
66. Nazarian R, Shi HB, Wang Q, Kong XJ, Koya RC, Lee H, Chen ZG, Lee MK, Attar N, Sazegar H, et al. Melanomas acquire resistance toB-RAF(V600E) inhibition by RTK or N-RAS upregulation. Nature. 2010;468(7326):973–U377.

67. Villanueva J, Vultur A, Lee JT, Somasundaram R, Fukunaga-Kalabis M, Cipolla AK, Wubbenhorst B, Xu XW, Gimotty PA, Kee D, et al. Acquired resistance to BRAF inhibitors mediated by a RAF kinase switch in melanoma can be overcome by cotargeting MEK and IGF-1R/PI3K. Cancer Cell. 2010;18(6):683–95.
68. Kaur H, Mao SH, Shah S, Gorski DH, Krawetz SA, Sloane BF, Mattingly RR. Next-generation sequencing: a powerful tool for the discovery of molecular markers in breast ductal carcinoma in situ. Expert Rev Mol Diagn. 2013;13(2):151–65.
69. Banerji S, Cibulskis K, Rangel-Escareno C, Brown KK, Carter SL, Frederick AM, Lawrence MS, Sivachenko AY, Sougnez C, Zou LH, et al. Sequence analysis of mutations and translocations across breast cancer subtypes. Nature. 2012;486(7403):405–9.
70. Shah SP, Roth A, Goya R, Oloumi A, Ha G, Zhao YJ, Turashvili G, Ding JR, Tse K, Haffari G, et al. The clonal and mutational evolution spectrum of primary triple-negative breast cancers. Nature. 2012;486(7403):395–9.
71. Stephens PJ, Tarpey PS, Davies H, Van Loo P, Greenman C, Wedge DC, Nik-Zainal S, Martin S, Varela I, Bignell GR, et al. The landscape of cancer genes and mutational processes in breast cancer. Nature. 2012;486(7403):400–4.
72. Leemans CR, Braakhuis BJM, Brakenhoff RH. The molecular biology of head and neck cancer. Nat Rev Cancer. 2011;11(1):9–22.
73. Berenson JR, Yang J, Mickel RA. Frequent amplification of the Bcl-1 locus in head and neck squamous-cell carcinomas. Oncogene. 1989;4(9):1111–6.
74. Loyo M, Li RJ, Bettegowda C, Pickering CR, Frederick MJ, Myers JN, Agrawal N. Lessons learned from next-generation sequencing in head and neck cancer. Head Neck. 2013;35(3):454–63.
75. Dotto GP. Notch tumor suppressor function. Oncogene. 2008;27(38):5115–23.
76. Mahjabeen I, Masood N, Baig RM, Sabir M, Inayat U, Malik FA, Kayani MA. Novel mutations of OGG1 base excision repair pathway gene in laryngeal cancer patients. Fam Cancer. 2012;11(4):587–93.
77. Scheckenbach K, Baldus SE, Balz V, Freund M, Pakropa P, Sproll C, Schafer KL, Wagenmann M, Schipper J, Hanenberg H. RAD51C–a new human cancer susceptibility gene for sporadic squamous cell carcinoma of the head and neck (HNSCC). Oral Oncol. 2014;50(3):196–9.
78. Laborde RR, Wang VW, Smith TM, Olson NE, Olsen SM, Garcia JJ, Olsen KD, Moore EJ, Kasperbauer JL, Tombers NM, et al. Transcriptional profiling by sequencing of oropharyngeal cancer. Mayo Clin Proc. 2012;87(3):226–32.
79. Mendelsohn J, Baselga J. The EGF receptor family as targets for cancer therapy. Oncogene. 2000;19(56):6550–65.
80. Mellinghoff IK, Wang MY, Vivanco I, Haas-Kogan DA, Zhu SJ, Dia EQ, Lu KV, Yoshimoto K, Huang JHY, Chute DJ, et al. Molecular determinants of the response of glioblastomas to EGFR kinase inhibitors. N Engl J Med. 2005;353(19):2012–24.
81. Yarden Y. The EGFR family and its ligands in human cancer: signalling mechanisms and therapeutic opportunities. Eur J Cancer. 2001;37:S3–8.
82. Citri A, Yarden Y. EGF-ERBB signalling: towards the systems level. Nat Rev Mol Cell Biol. 2006;7(7):505–16.
83. Hynes NE, Lane HA. ERBB receptors and cancer: the complexity of targeted inhibitors (vol 5, pg 341, 2005). Nat Rev Cancer. 2005;5(7):341–54.
84. Ziogas DE, Katsios CS, Tzaphlidou M, Roukos DH. Targeted therapy: overcoming drug resistance with clinical cancer genome. Expert Rev Anticancer Ther. 2012;12(7):861–4.
85. Ross JS, Torres-Mora J, Wagle N, Jennings TA, Jones DM. Biomarker-based prediction of response to therapy for colorectal cancer: current perspective. Am J Clin Pathol. 2010;134(3):478–90.
86. Rodenhuis S, Slebos RJC, Boot AJM, Evers SG, Mooi WJ, Wagenaar SS, Vanbodegom PC, Bos JL. Incidence and possible clinical significance of K-Ras oncogene activation in adenocarcinoma of the human lung. Cancer Res. 1988;48(20):5738–41.

87. Ahrendt SA, Decker PA, Alawi EA, Zhu YR, Sanchez-Cespedes M, Yang SC, Haasler GB, Kajdacsy-Balla A, Demeure MJ, Sidransky D. Cigarette smoking is strongly associated with mutation of the K-ras gene in patients with primary adenocarcinoma of the lung. Cancer. 2001;92(6):1525–30.
88. Pao W, Wang TY, Riely GJ, Miller VA, Pan QL, Ladanyi M, Zakowski MF, Heelan RT, Kris MG, Varmus HE. KRAS mutations and primary resistance of lung adenocarcinomas to gefitinib or erlotinib. PLoS Med. 2005;2(1):57–61.
89. De Roock W, Piessevaux H, De Schutter J, Janssens M, De Hertogh G, Personeni N, Biesmans B, Van Laethem JL, Peeters M, Humblet Y, et al. KRAS wild-type state predicts survival and is associated to early radiological response in metastatic colorectal cancer treated with cetuximab. Ann Oncol. 2008;19(3):508–15.
90. Engelman JA, Zejnullahu K, Mitsudomi T, Song YC, Hyland C, Park JO, Lindeman N, Gale CM, Zhao XJ, Christensen J, et al. MET amplification leads to gefitinib resistance in lung cancer by activating ERBB3 signaling. Science. 2007;316(5827):1039–43.
91. Chapman PB, Hauschild A, Robert C, Haanen JB, Ascierto P, Larkin J, Dummer R, Garbe C, Testori A, Maio M, et al. Improved survival with vemurafenib in melanoma with BRAF V600E mutation. N Engl J Med. 2011;364(26):2507–16.
92. Sieben NLG, Macropoulos P, Roemen GM, Kolkman-Uljee SM, Fleuren GJ, Houmadi R, Diss T, Warren B, Al Adnani M, de Goeij AP, et al. In ovarian neoplasms, BRAF, but not KRAS, mutations are restricted to low-grade serous tumours. J Pathol. 2004;202(3):336–40.
93. Bang YJ, Van Cutsem E, Feyereislova A, Investigators TT. Trastuzumab in combination with chemotherapy versus chemotherapy alone for treatment of HER2-positive advanced gastric or gastro-oesophageal junction cancer (TOGA): a phase 3, open-label, randomised controlled trial (vol 376, pg 687, 2010). Lancet. 2010;376(9749):1302.
94. Bookman MA, Darcy KM, Clarke-Pearson D, Boothby RA, Horowitz IR. Evaluation of monoclonal humanized anti-HER2 antibody, trastuzumab, in patients with recurrent or refractory ovarian or primary peritoneal carcinoma with overexpression of HER2: a phase II trial of the Gynecologic Oncology Group. J Clin Oncol. 2003;21(2):283–90.
95. Fleming GF, Sill MW, Darcy KM, McMeekin DS, Thigpen JT, Adler LM, Berek JS, Chapman JA, DiSilvestro PA, Horowitz IR, et al. Phase II trial of trastuzumab in women with advanced or recurrent, HER2-positive endometrial carcinoma: a Gynecologic Oncology Group study. Gynecol Oncol. 2010;116(1):15–20.
96. Fang WJ, Radovich M, Zhou AW, Zheng YL, Zhao P, Huang WY, Mao CY, Zheng Y, Jia YK, Zheng SS. "Druggable" alterations detected by Ion Torrent in metastasis colorectal cancer patients. Oncol Lett. 2014;7(6):1761–6.
97. Ciardiello F, Tortora G. Drug therapy: EGFR antagonists in cancer treatment. N Engl J Med. 2008;358(11):1160–74.
98. Sok JC, Coppelli FM, Thomas SM, Lango MN, Xi SC, Hunt JL, Freilino ML, Graner MW, Wikstrand CJ, Bigner DD, et al. Mutant epidermal growth factor receptor (EGFRvIII) contributes to head and neck cancer growth and resistance to EGFR targeting. Clin Cancer Res. 2006;12(17):5064–73.
99. Munger K, Werness BA, Dyson N, Phelps WC, Harlow E, Howley PM. Complex-formation of human papillomavirus-E7 proteins with the retinoblastoma tumor suppressor gene-product. Embo J. 1989;8(13):4099–105.
100. Chong CR, Janne PA. The quest to overcome resistance to EGFR-targeted therapies in cancer. Nat Med. 2013;19(11):1389–400.
101. Hoellein A, Pickhard A, von Keitz F, Schoeffmann S, Piontek G, Rudelius M, Baumgart A, Wagenpfeil S, Peschel C, Dechow T, et al. Aurora kinase inhibition overcomes cetuximab resistance in squamous cell cancer of the head and neck. Oncotarget. 2011;2(8):599–609.
102. Wheeler DL, Huang S, Kruser TJ, Nechrebecki MM, Armstrong EA, Benavente S, Gondi V, Hsu KT, Harari PM. Mechanisms of acquired resistance to cetuximab: role of HER (ErbB) family members. Oncogene. 2008;27(28):3944–56.

103. Zhang JJ, Baran J, Cros A, Guberman JM, Haider S, Hsu J, Liang Y, Rivkin E, Wang JX, Whitty B, et al. International cancer genome consortium data portal—a one-stop shop for cancer genomics data. Database-Oxford. 2011;2011:bar026.
104. Kaye J, Boddington P, De Vries J, Hawkins N, Melham K. Ethical implications of the use of whole genome methods in medical research. Eur J Hum Genet. 2010;18(4):398–403.
105. Wolf SM, Crock BN, Van Ness B, Lawrenz F, Kahn JP, Beskow LM, Cho MK, Christman MF, Green RC, Hall R, et al. Managing incidental findings and research results in genomic research involving biobanks and archived data sets. Genet Med. 2012;14(4):361–84.
106. Clayton EW. Incidental findings in genetics research using archived DNA. J Law Med Ethics. 2008;36(2):286–91.
107. Zonta MA, Monteiro J, Santos Jr G, Pignatari ACC. Oral infection by the human papilloma virus in women with cervical lesions at a prison in São Paulo Brazil. Braz J Otorhinolaryngol. 2012;78(2):66–72.
108. Farah CS, Bhatia N, John K, Lee BW. Minimum intervention dentistry in oral medicine. Aust Dent J. 2013;58(Suppl1):85–94.
109. Fong PC, Boss DS, Yap TA, Tutt A, Wu P, Mergui-Roelvink M, Mortimer P, Swaisland H, Lau A, O'Connor MJ, et al. Inhibition of poly(ADP-ribose) polymerase in tumors from BRCA mutation carriers. N Engl J Med. 2009;361(2):123–34.
110. Garcia I, Kuska R, Somerman MJ. Expanding the foundation for personalized medicine: implications and challenges for dentistry. J Dent Res. 2013;92(7 Suppl):S3–10.
111. Rosenberg NA, Pritchard JK, Weber JL, Cann HM, Kidd KK, Zhivotovsky LA, Feldman MW. Genetic structure of human populations. Science. 2002;298(5602):2381–5.

CIC Mutation as Signature Alteration in Oligodendroglioma

Shiekh Tanveer Ahmad, Wei Wu, and Jennifer A. Chan

Abstract Oligodendroglioma (ODG) is a type brain tumor that predominantly affects young adults and that is characterized by unique clinical, morphological, genetic, and molecular features. Molecular characterization of ODGs has identified concurrent 1p and 19q whole-arm chromosomal losses and *IDH* mutations as signature alterations in ODG. More recently, mutations in the gene *CIC* on chromosome 19q13.2 have been found to be present in the majority of ODGs. *CIC* mutations occur nearly exclusively in the context of 1p/19q co-deletion, and it is likely that the *CIC* mutation on the remaining 19q allele is integral to the disease pathogenesis. In contrast to ODGs, where >70 % of cases harbor CIC mutations, *CIC* mutations are found at a low frequency across diverse tumor types. To date, little is known how CIC mutation contributes to development of ODGs or other cancers. Most of literature on CIC has been based on studies in *Drosophila*, where CIC has spatiotemporal effects in regulating RTK signaling for normal embryonic development. In this chapter, we provide a brief introduction to oligodendrogliomas, review CIC's role as a transcriptional repressor and functions in development, and discuss the potential role(s) of CIC mutation in the pathogenesis of human cancer.

Abbreviations

CIC	Capicua
CP	Cortical plate
EGFR	Epidermal growth factor receptor
ETS	E twenty-six
ETV	ETS translocation variant

S.T. Ahmad • W. Wu
Department of Pathology and Laboratory Medicine, Southern Alberta Cancer Research Institute, University of Calgary, Calgary, AB, Canada T2N 4N1

J.A. Chan (✉)
Department of Pathology and Laboratory Medicine, Southern Alberta Cancer Research Institute, University of Calgary, Room HRIC 2A-08, 3330 Hospital Drive NW, Calgary, AB, Canada T2N 4N1
e-mail: jawchan@ucalgary.ca

W. Wu, H. Choudhry (eds.), *Next Generation Sequencing in Cancer Research, Volume 2*, DOI 10.1007/978-3-319-15811-2_24

HMG	High mobility group
IDH	Isocitrate dehydrogenase
IZ	Intermediate zone
MMP	Metallomatrix protein
NSC	Neural stem cell
ODG	Oligodendroglioma
OPC	Oligodendrocyte precursor cell
PDGFR	Platelet derived growth factor receptor
PEA3	Polyoma enhancer activator 3
PTEN	Phosphatase and tensin homolog
RTK	Receptor tyrosine kinase
SVZ	Subventricular zone
VZ	Ventricular zone

1 Introduction

Oligodendrogliomas (ODGs) are clinically, pathologically, and genetically distinctive gliomas that are composed of neoplastic cells resembling oligodendrocytes. Histologically, the tumors are composed of densely packed cells that have round, regular nuclei and clearing of the cytoplasm and are associated with fine branching vasculature [1]. In the past, the diagnosis of ODG has suffered from inter-observer variation, as the diagnosis was solely based on histology. Over the past two decades, however, several molecular alterations have been discovered that, together, characterize this distinctive tumor type. The first of these was the observation that whole-arm 1p and 19q chromosomal co-deletions are a signature alteration in ODGs [2, 3]. As one of the first known diagnostic, prognostic, and predictive genetic markers in neuro-oncology, 1p/19q testing quickly became a diagnostic standard-of-care that continues to be used clinically today. With that discovery, the hunt for potential causative ODG gene(s) on 1p and 19q began. A decade later, it was found that virtually all ODGs also harbor gain-of-function mutations in isocitrate dehydrogenase (*IDH*)—in either *IDH1* on chromosome 2 or *IDH2* on chromosome 15 [4]. But IDH mutations, although characteristic, were not specific to ODG, and the suspected ODG genes remained elusive. Recently, Capicua (*CIC*) on chromosome 19q13 has been identified as a gene mutated in the majority of ODGs [5, 6]. CIC mutations are unique to ODGs, and *CIC* mutation in the setting of concurrent 1p/19q loss and *IDH1/2* mutation constitutes the prototypical genetic signature of ODG [5–8].

The challenge now is to move from the genetic characterization of ODG to understanding how these genes contribute to the initiation and progression of ODG. CIC is a known transcriptional repressor whose default repressor activities are normally relieved upon RAS/MAPK signaling. We postulate that loss of CIC function de-represses critical transcriptional programs either to bias neural progenitor cells to an oligodendrocyte precursor cell (OPC)-like cell fate and/or to promote aberrant proliferation of OPCs. In a background of *IDH* mutation, inactivation of

CIC may deregulate cellular responses specifically controlling OPC differentiation and proliferation, contributing to ODG genesis. In this chapter we summarize the current knowledge about CIC as a transcriptional repressor, touch on its roles during development and normal tissue maintenance, and focus on its associations with human disease, in particular with oligodendrogliomas.

2 Clinical and Molecular Characteristics of Oligodendroglioma

2.1 Clinical Overview of Oligodendroglioma

ODG is a type of malignant brain tumor that occurs most frequently in the frontal and temporal lobes, and is composed of oligodendroglial-like cells [3, 9, 10]. ODGs represent ~10 % of gliomas [9]. Tragically, new ODG diagnoses peak in young adulthood, burdening patients with seizures, cognitive difficulties, headaches, and personality changes among other problems [9, 11]. Furthermore, although slowly growing and more responsive to therapy than other gliomas, ODGs infiltrate brain tissue diffusely, progress in malignancy, and are eventually fatal [9, 11, 12]. Unlike other gliomas such as astrocytomas and ependymomas, oligodendrogliomas are chemosensitive and often progress in a slow and predictable manner [13]. While median survival for ODGs treated with surgery and radiotherapy is 4–7 years, and those treated with surgery, radiotherapy, and chemotherapy is 13–15 years [12, 14], tumor recurrences are common and many patients die of their disease. To improve outcome for ODG patients, more work is needed to understand the molecular and cellular events behind ODG genesis and progression.

2.2 Molecular Characterization of ODGs

IDH mutation: IDH mutations most commonly affect *IDH1* but are occasionally seen in *IDH2*. The most common *IDH1* mutations in glioma (>95 %) result in an amino acid substitution at arginine 132 (R132), which resides in the enzyme's active site. *IDH1* and *IDH2* mutations affect a single amino acid (either amino acid 132 of Idh1 or the analogous amino acid 172 of Idh2) [4, 15], creating gain-of-function alleles. Mutation of both of the IDHs impart the ability to produce 2-hydroxyglutarate (2-HG), a potential oncometabolite [16] that may promote neoplasia by altering the epigenetic landscape and activating hypoxic responses [15]. In gliomas, a distinctive CpG island methylator phenotype (G-CIMP) is seen in *IDH*-mutant tumors [17], and in vitro experiments indicate that *IDH* mutations are sufficient to establish the methylator phenotype and alter cellular differentiation [18, 19]. *IDH* mutations may also alter cellular differentiation via modification of histone methylation patterns. Although present in >85 % of ODGs, *IDH* mutations are not unique to ODG.

Other types of diffuse gliomas (e.g., WHO grade II and III diffuse astrocytomas) as well as some non-nervous system tumors such as chondrosarcoma, myelodysplastic syndrome, acute myelogenous leukemia, and cholangiocarcinoma also harbor mutations in *IDH1/2* [20].

1p/19q co-deletion: The combined whole-arm losses of 1p and 19q are perhaps the most distinctive of molecular alterations in ODG. Such 1p/19q co-deletion is mediated by an unbalanced translocation of 19p to 1q [21, 22]—most likely the result of a centrosomal or pericentrosomal translocation of chromosomes 1 and 19 results in two derivative chromosomes, der(1,19)(p10;q10) and der(1,19)(q10;p10), after which the derivative chromosome with the short arm of chromosome 1 and the long arm of chromosome 19 is lost. A possible explanation for this translocation is the strong homology of the centromeric regions of chromosomes 1 and 19. With respect to 1p loss, it is the loss of the entire short arm of chromosome 1 that is defining for ODG, as partial 1p deletions that are not associated with 19q loss may occur in other glioma types such as glioblastoma [23]. Of note, some ODGs possess polysomy of 1q and 19p in the context of relative 1p/19q co-deletion [24]. Such co-polysomy is independently associated with shorter overall survival in 1p/19q co-deleted ODGs, irrespective of tumor grade.

1p/19q co-deleted ODGs are associated with the constellation of positive prognostic markers including methylation of the *MGMT* promoter, *IDH* mutations and G-CIMP. ODGs with 1p/19q co-deletion have also been shown to be enriched with a proneural gene expression signature [25]. Although MGMT promoter methylation, IDH mutations, and G-CIMP are also present in diffuse astrocytomas and in glioblastomas that arise from lower-grade astrocytomas, important differences in the molecular signature of ODGs and lower-grade diffuse astrocytomas is 1p/19q co-deletion in the former and ATRX loss in the latter [26, 27]. The mutual exclusivity of these events underscores the distinct molecular characteristics of ODGs.

CIC mutations: Although 1p/19q loss had been known as a signature alteration in ODGs for nearly two decades, the causative gene(s) on 1p or 19q remained unknown until more recently. Next-generation sequencing enabled the discovery that *CIC* located on chromosome 19q is mutated in most ODGs [5, 6]. These *CIC* mutations are present nearly exclusively together with *IDH* mutation and single copy 1p/19q loss [5, 6]. Furthermore, the type and distribution of *CIC* mutations includes frameshifting insertions/deletions and frequent truncations that are distributed across the gene, albeit with some increased frequency in exons 5 and 20 [5, 6]. Though such hemizygous *CIC* mutations on the retained 19q allele are thought to be functionally important in ODG, the mechanism of action is as yet undetermined. Nevertheless, considering the patterns of mutations and copy number alterations, *CIC* is likely a tumor suppressor gene [28].

The prototypical genetic signature of ODG is now recognized as a trifecta of *IDH* mutation, 1p/19q chromosomal co-deletion, and *CIC* mutation [3, 5–8, 25]. The majority of classic ODGs carry the constellation of 1p/19q loss, *IDH* and *CIC* mutation is now well documented. Beyond these changes, few recurrent genetic alterations are known. Mutations of *FUBP1* on chr1p have also been identified in

some ODG, but these are much less frequent than *CIC* mutations and all occur in ODGs with *CIC* mutations [6, 8]. *FUBP1* mutations may be a later change related to anaplastic progression rather than ODG initiation. *PTEN* loss and *CDKN2A/B* loss are also seen in some cases, and are more frequently present in higher grade ODG than lower grade ODG. Another recognized event associated with higher grade ODG is silencing of *RB1* by promoter methylation [9, 29–31]. It is worth mentioning that some tumors can resemble oligodendroglioma histologically but behave very differently from typical ODG. Lesions that bear some ODG-like histology include glioblastoma with oligodendroglial component and oligoastrocytoma. These tumors have a different mutational spectrum than classic ODG and are more frequently characterized by *EGFR* amplification and/or *TP53* mutation [30, 32] rather than the constellation of 1p/19q loss, *CIC* mutation, and *IDH* mutation.

As we know that the *CIC* mutations occur concurrently with *IDH* mutations in ODGs, mutant *IDH* may be a prerequisite that alters the cell's epigenetic state such that Cic's tumor suppressive role is unmasked. Carefully designed functional studies taking into account genomic context (such as 1p/19q loss and *IDH* mutations) are necessary to delineate the role of *CIC* in the pathogenesis of ODG.

3 CIC Structure and Function

3.1 CIC is a Default Transcriptional Repressor Downstream of RAS/MAPK Signaling

CIC was discovered in several developmental contexts in *Drosophila* [33]. Activation of specific receptor tyrosine kinase/Ras/Raf/MAPK pathways relieves repression of target genes that are normally suppressed by CIC, leading to the transcription of genes that are involved in important developmental processes. In *Drosophila* these processes include patterning and differentiation in wing veins, eye imaginal discs, and head and tail regions [33–37] (the latter inspiring the gene name "capicua" meaning head-and-tail in Catalan). Torso signaling is relayed through the MAPK (mitogen-activated protein kinase) signal transduction pathway in which signals are propagated through sequential phosphorylation of Ras, Raf, DSOR, and Rolled (Drosophila homologs of mammalian Ras, Raf, Mek, and Erk, respectively) to result in activation of Tailless (*tll*) and Huckebein (*hkb*). Loss-of-function mutations in CIC lead to increased expression of *tll* and *hkb* as well as alteration of tissue patterning, suggesting its role as a default transcriptional repressor regulating genes downstream of RTK/MAPK signaling in flies [36, 38]. *Drosophila* CIC is known to act downstream of both Torso and the epidermal growth factor receptor (EGFR), both of which promulgate their signals through the Ras–Raf–MAPK set of intermediate signaling proteins. Thus, CIC acts as a common transcriptional repressor, normally silencing the targets of these two RTKs.

As in the case with Drosophila, RTK/MAPK signaling in other systems including humans, is a core pathway that regulates diverse cellular processes such as

growth, proliferation, apoptosis, migration, metabolism and differentiation [39]. Aberrations in RTK signaling may lead to a variety of human diseases, including cancer. Binding of a growth factor to its cognate RTK initiates activation of the canonical RTK pathway during which the auto-phosphorylated and active RTK signals through a complex of Grb/Gab/Shc/Shp to convert inactive Ras-GDP to active Ras-GTP (Fig. 1). From Ras activation, the signal transduction cascade proceeds largely (albeit not exclusively) through phosphorylation and activation of Raf, Mek, and Erk kinases. Erk kinases subsequently phosphorylate several cytoplasmic and nuclear substrates. The nuclear factors that are downstream of RTK/MAPK pathway flux have key roles in dictating the ultimate changes in gene expression and, hence, eliciting of the appropriate biological responses [39]. Studies in human cells suggest that CIC regulates RTK-dependent responses that are important in to decide the fate of a cell whether to proliferate or differentiate, providing a plausible link to cancer [6, 40–43]. In human cells, among the best characterized downstream effectors of MAPK signaling that have been found to be regulated by CIC are transcription factors of the ETS superfamily such as ETV1, ETV4, and ETV5 [39, 44, 45].

Subsequent work has delineated several structural elements and other requirements for positive RTK/MAPK signaling to relieve CIC repression of target genes. In the following sections we highlight some important domains in CIC and then discuss biological processes that may be impacted by CIC dysregulation in the context of human disease.

3.2 Structural Elements in CIC Important for Function

Human CIC is present on chromosome 19q13.2 and consists of 20 exons. It is member of a SOX-related HMG subfamily [39], and is conserved across evolution, from flies and worms to mammals. There are two known major isoforms of CIC recognized in flies as well as in mammals, a short form (CIC-S) and a long form (CIC-L). The CIC-S is ~160 kD and CIC-L is ~250 kD and both differ in their N-terminal regions [5, 42] (Fig. 2). At the N-terminal CIC-L isoform has an extended segment (N1) with a highly conserved domain of unknown cellular function [39]. There is a nuclear localization signal domain present in human CIC that is not conserved and is involved in nuclear transport of CIC (Fig. 2). Recently the CIC-L form has been shown to localize to nucleus while as CIC-S form in cytoplasm, and the two isoforms may have distinct functions [46]. CIC-L is the predominant form expressed in the brain and in ODG.

CIC contains two highly conserved domains—the HMG-box that binds to DNA, and a C-terminal Groucho-like (Gro-L) domain that may be important in mediating protein–protein interactions [33, 40, 41, 47] (Fig. 2). Human CIC cDNA exhibits 92 % identity with the mouse gene, with 100 % identity in the HMG DNA binding domain. Through the HMG-box, CIC binds the octameric sequence T(G/C) AATG(G/A)A in target enhancers and promoters, leading to transcriptional repression of the target genes [36, 38, 40, 42, 45]. This default repression is relieved upon induction of RTK signaling to allow transcription of RTK-dependent target genes.

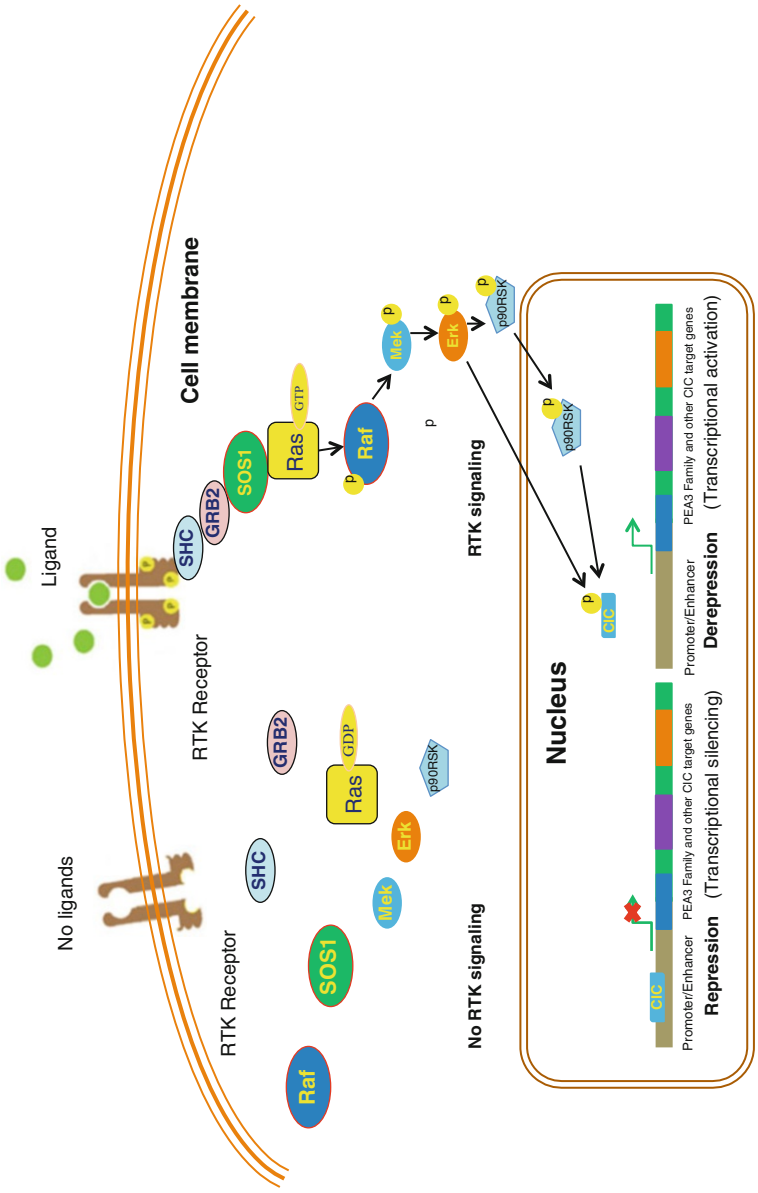


Fig. 1 Structural features of human CIC proteins. Two main isoforms, short (CIC-S) and long (CIC-L), are present in each species. Functional domains that have been identified in humans are indicated, and their sequence conservation is depicted with *colored boxes*. Note that the nuclear localization signal (NLS) that has been identified in human CIC [43] is not conserved. All CIC domains (except the HMG-box) appear to be unique to CIC proteins. *Numbers* indicate amino acid positions

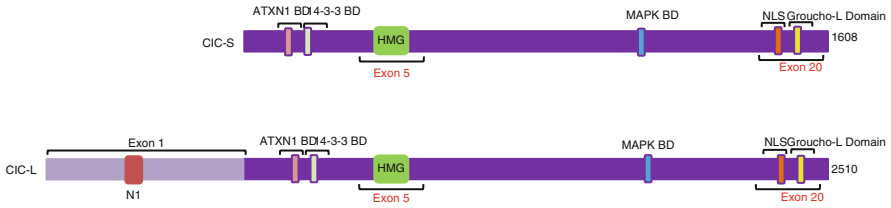


Fig. 2 Schematic of general mechanism of CIC regulation by RTK–Ras–MAPK signaling. In the absence of active RTK signaling, the CIC behaves as transcriptional repressor of its target genes by binding to their regulatory elements. But in the presence of RTK signaling, the ligand binds to its cognate receptor, the later gets auto-phosphorylated. The auto-phosphorylated and active RTK then signals through a complex of Grb/Gab/Shc/Shp to convert inactive Ras-GDP to active Ras-GTP. The signal transduction cascade proceeds through phosphorylation and activation of Raf1, Mek, and MAPK. MAP kinase subsequently phosphorylates several cytoplasmic and nuclear substrates. With respect to CIC, conserved (pS/T)P Erk phosphorylation sites are present in the C-terminal region and adjacent to the HMG box, which have been found to affect nuclear localization and binding to partner transcriptional regulators, thus relieving repression of target genes that include the PEA3 subfamily of ETS transcription factors such as *ETV1*, *ETV4*, and *ETV5* [43]

Aside from binding to DNA, the interaction of Cic with other proteins, possibly through its C1 motif, also appears important [34, 40, 47]. Cic requires co-repressors such as Groucho (GRO), ATXN1, ATAXN1L that have been implicated in its repressive activities during the development. Maintenance of repression of Cic's target genes tll and hkb in *Drosophila* embryonic development requires the presence of co-repressor GRO in the enhancer region of these genes to create a state of transcriptional inactivation, but in vivo studies have failed to show physical interaction between GRO and Cic [38, 48, 49]. In contrast, studied in mammalian cells have indicated that the Cic requires physical interaction to form repressor complexes with co-repressors ATAXN1 and its related factor ATAXN1-Like protein [42].

CIC has conserved (pS/T)P Erk phospho-acceptor sites in the C-terminal region and region adjacent to the HMG box, which affect nuclear localization and binding to partner transcriptional regulators in response to RTK activation [43]. In mammals, activation of EGFR signaling is directly proportionate to CIC degradation [50, 51]. EGFR activation phosphorylates CIC at several sites directly through MAPK/Erk as well as ribosomal protein S6 kinase II (p90RSK), which is itself a downstream target of activated MAPK [56].

Following MAPK activation, CIC is phosphorylated and downregulated either via degradation or export to cytoplasm [33, 34, 37–39, 47, 52], thus relieving repression of target genes (Fig. 1). MAPK-dependent phosphorylation can prevent CIC binding to importin-α4 (or KPNA3), a nuclear import protein. CIC phosphorylation by p90RSK can also promote its binding to 14-3-3 regulatory proteins [43, 53]. Importantly, this interaction alters CIC binding to its target consensus DNA binding sequence, ultimately resulting in transcriptional upregulation of CIC targets, such as the PEA3 subfamily of ETS transcription factors including *ETV1*, *ETV4*, and *ETV5*

(Fig. 1) [43]. Thus, the possible mechanisms by which Cic activity is repressed include phosphorylation-dependent changes in subcellular localization, protein stability/degradation, and alteration in DNA binding properties or interactions with other proteins. Of note, although canonical RTK signaling largely has effects through the Ras/MAPK cascade, other Ras effector pathways are recognized, including PI3K and Ral pathways. The phosphorylation of CIC and the transcriptional responses of *ETV* genes, however, appear to be predominantly a result of MAPK-mediated phosphorylation, and not activation of other effector pathways [43]. Recently studies have also suggested that CIC may be an important modulator and integrator of RTK signals, functionally interpreting different intensities and/or duration of signals as a result of competition between various MAPK substrates [54, 55].

4 CIC from the Perspective of Cancer: From Default Repressor to Tumor Suppressor

Analyses of protein coding genes for types of mutations and mutational frequencies in cancer has led to the identification of CIC as a cancer driver gene with likely tumor suppressor functions [28, 56]. Compiling data from the Cancer Genome Atlas (TCGA), International Cancer Genome Consortium (ICGC) and other online cancer genomic databases, we see that CIC alterations (including mutations, deletions, and copy number variations) are present in diverse cancer types, but are most highly represented among gliomas (Fig. 3). Notably, although the frequency of CIC mutation in the figure suggests its mutation in only about 1/5 of gliomas in general, as discussed above, CIC alterations are highly specific to oligodendrogliomas in particular, where they are found the in the majority of cases.

In considering CIC's potential role in pathogenesis of cancer and particularly oligodendrogliomas, it is worthwhile briefly reviewing our knowledge of CIC's biologic functions in normal developmental contexts. Most of our current understanding of CIC's biologic functions stems from work in *Drosophila*; however, CIC and the pathways that regulate it are highly conserved from flies to mammals. In *Drosophila*, Cic is important in embryonic patterning, as its loss results in abnormal boundaries for the head and tail regions at the embryonic poles. Patterning functions are also evident in wing development where abnormal wing vein pattern and vein cell determination downstream of EGFR signaling are effects of Cic loss [33, 38, 47, 57, 58]. Cell fate and differentiation functions have been discerned in neuroblasts of the embryonic neuroectoderm and in the dorsoventral specification of follicle cells during oogenesis [34, 38, 59]. Finally, roles in regulating cell proliferation are evident in the drosophila eye and intestine, where mutations that disrupt Cic function increase proliferation without affecting cell size, differentiation, or patterning of the eye [52], and cause proliferation of ectopic intestinal stem cells [60]. All of these processes—tissue patterning, cell fate determination, and regulation of proliferation—are dysregulated in the pathogenesis of cancer.

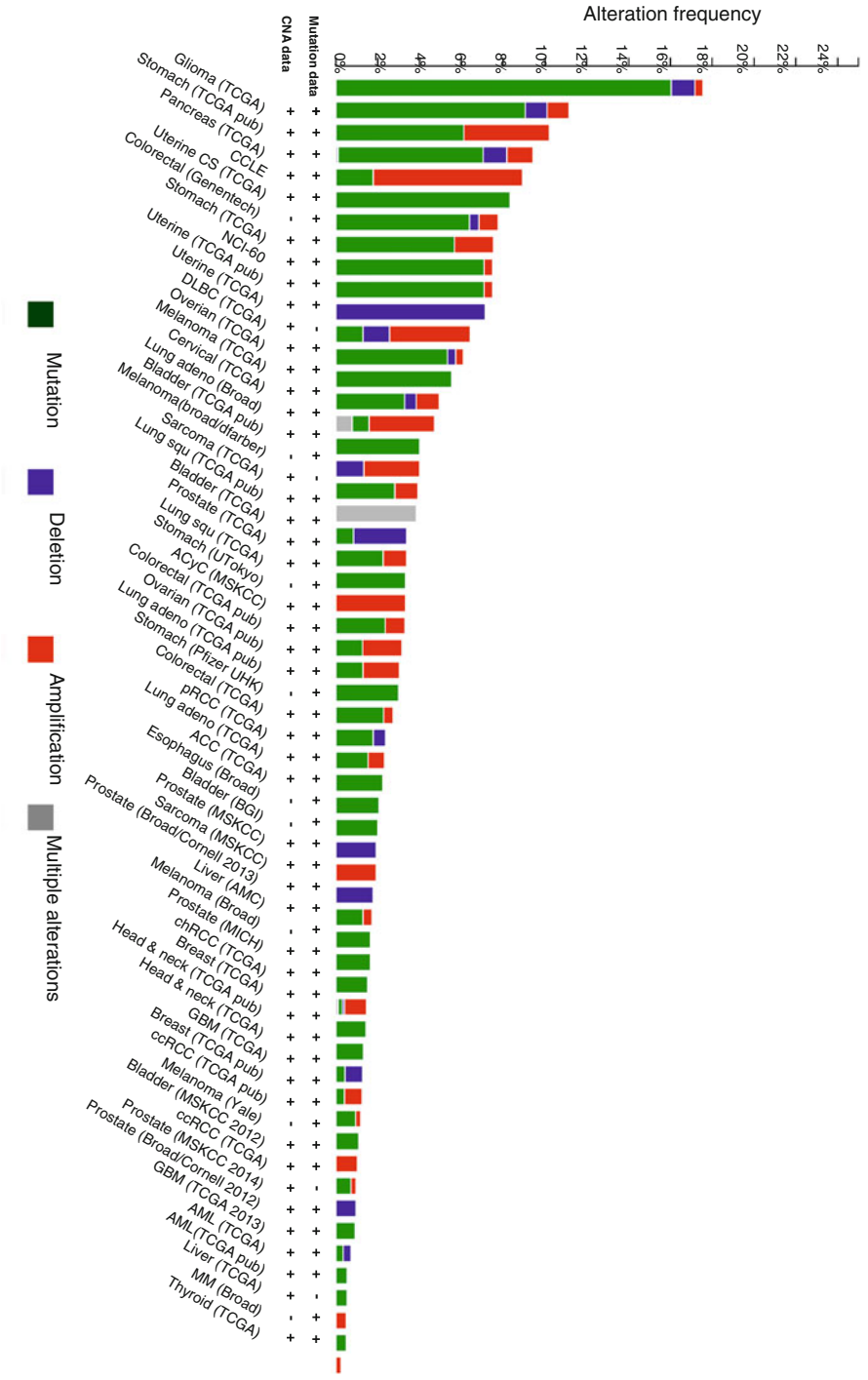


Fig. 3 Distribution of CIC alteration across different cancers types. The type of alteration include mutations, deletions and amplification of CIC are shown as *bars* across different cancer types and the database sources. The *height of the bars* represents the relative percentage frequency of each alteration across each cancer

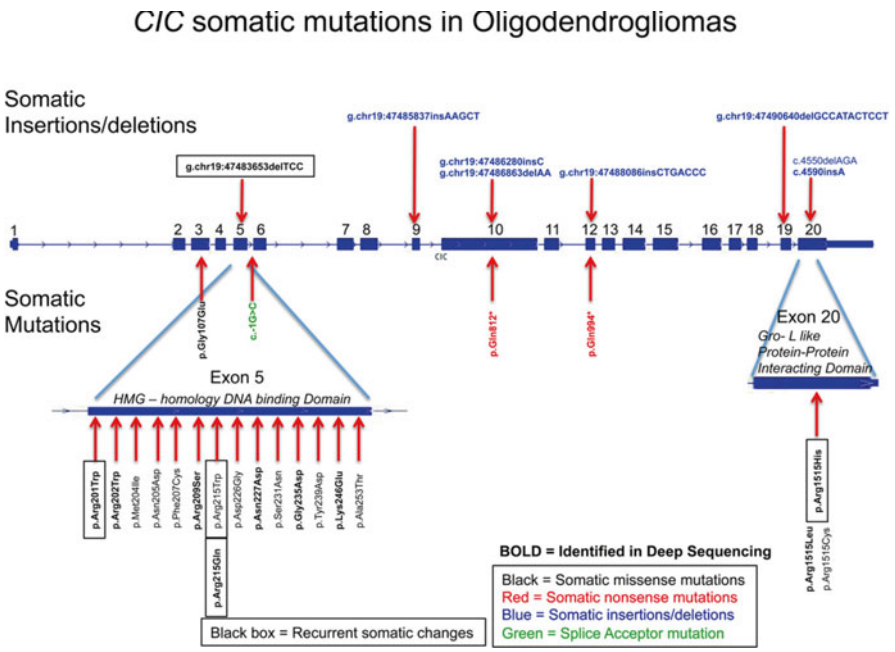


Fig. 4 Schematic of distribution of recurrent mutations across the human CIC gene. Most frequent mutations in 1p/19q co-deleted oligodendroglioma (IDH1/2 mutated) are found within the highly conserved DNA-interacting HMG domain (exon 5) and protein–protein interacting GRO-L homology domain (exon 20) [5]. Somatic mutations are identified with *arrows* and amino acid changes. (Reproduced with prior permission from publisher)

In ODGs, mutations are present throughout CIC gene, but are most highly concentrated in exon 5 encoding the HMG-box, and exon 20 in the Gro-L domain (Fig. 4). With respect to potential effects of CIC loss of function, some effects may be due to loss of its repressor activity; it is possible that loss of Cic-DNA binding or loss of Cic interactions with normal partner proteins in ODG due to mutation or other alterations may lead to constitutive de-repression of cancer-promoting genes. PEA3 Ets transcription factors such as Etv4 and 5 are candidates in this respect, as there are several lines of evidence that point to dysregulation of PEA3 Ets proteins as oncogenic in other cancer types [61, 62]. Evidence that PEA genes are important mediators of oncogenic effects related to CIC come from the findings in a subset of Ewing-like sarcomas in which a t(4;19) chromosomal translocation results in the creation of a *CIC-DUX4* fusion protein that results in transcriptional activation of ETV1, ETV4, and ETV5 instead of transcriptional repression [40, 63, 64]. Kawamura-Saito et al. demonstrated binding of the HMG box of *CIC* to a DNA sequence within the promoter of PEA genes *ETV1* and *ETV5* and further revealed that fusion of DUX4 to CIC sequence provides strong transcriptional activity, resulting in mostly upregulated gene expression, with minimal downregulated genes [40]. Aside from the few recognized target genes such as *ETV*s, however, the

transcriptional repressive repertoire of CIC that may be responsible for various aspects of oncogenicity is not well known. In addition, there may be some CIC functions that are independent of its repressor activity that are relevant to cancer, including gliomas.

Extracelllular matrix remodeling: CIC–PEA3 transcriptional circuits appear to affect extracellular matrix (ECM) remodeling in both development and cancer. In the mammals, *Cic* interacts with Ataxin1, the causative gene in the neurodegenerative disease Spinocerebellar Ataxia type 1 [42]. Molecular and genetic analyses have identified a crucial role for *CIC–ATXN1* and *CIC–ATAXNL1* complexes in mediating direct transcriptional repression of PEA3 genes during lung development [45]. Mutant mice that lack *Cic* or *Atxn1* and *Ataxnl1* activities present several defects, including abnormal alveolarization in developing lungs and de-repression of PEA3 subfamily genes, particularly of ETV4. Increased activity of *Etv4* in the mutant mice upregulates expression of the matrix metalloproteinase 9 (*Mmp9*) gene which is known for their role in ECM remodeling and lung alveolarization [45]. The oncogenic activities of PEA3 transcription factors are found in several types of tumor, such as Ewing sarcoma, melanoma or prostate cancer, lead to the upregulation of MMP family genes and other targets involved in ECM remodeling—which contributes to invasive and metastatic behavior [65, 66].

Proliferation/cell cycle control: CIC is known to affect proliferation and cell cycle control in model organisms during development [67]. During the development of *Drosophila* eye, Cic restricts the rate of proliferation in response to EGFR/Ras signaling, and the functional loss of Cic bypasses the requirement for EGFR/Ras activity in proliferation thus hinting at its role in cell cycle regulation [52]. Cic in *Drosophila* has also been shown to synergize with RBF1 (*Drosophila* functional homologue of the human RB tumor suppressor protein) to restrict Cyclin E expression to maintain G1 arrest and not allow it to cross the threshold necessary for the cell to enter into S phase. Moreover, the proliferative effects of Cic loss of function mutation has been linked to regulation to reactive oxygen species levels [67]. With respect to oligodendroglioma biology, the question of whether and how Cic might regulate proliferation in neural stem cells or OPCs is an area of active investigation.

Cell metabolism: Interestingly, a recent study reported for the first time non-nuclear function of *CIC-S*, and indicated it to be involved in cell metabolism [46]. This study found human *CIC-S* localized in the cytoplasm in proximity to mitochondria where it seems to interact with the enzyme ATP-citrate lyase (*ACLY*). This enzyme converts the cytosolic citrate into oxaloacetate and Acetyl Co-A in an ATP dependent manner. As we know that, there is a correlation between the levels of ACLY/pACLY and tumorigenicity in gliomas [68], increased levels of ACLY and pACLY have been found to be associated with proliferation in other types of aggressive cancers [69–71]. In glioma cells, increased pACLY is observed to be associated with increased clonogenicity and cell migration [68] while as other studies indicate, inhibition or reduction of ACLY suppresses cell proliferation [72, 73]. Consistent with these findings, Dr. Marco Marra and colleagues found mutant *CIC-S* to be

involved in reducing ACLY/pACLY levels in brain tumor cell lines and in human tumor samples and correlated this reduction with reduced clonogenicity in vitro [46]. This study has provided a hint that the *CIC* may be involved in diverse biological functions that may ultimately converge into tumorigenic pathways upon its loss than simply being a transcriptional repressor.

Response to hypoxia: Recently, Udpa and colleagues in a population-based study using whole-genome sequencing to identify genes involved in high-altitude adaptation identified genetic regions with significant loss of diversity, including a region on chromosome 19 that contains 8 genes, including CIC, LIPE, and PAFAH1B3 [74]. They evaluated the roles of these genes in hypoxia tolerance by using loss of function approach in *Drosophila*. Most importantly the knockdown of Cic resulted in increased tolerance and survival of flies in hypoxic environments [74]. Further studies could determine whether Cic loss-of-function might similarly aid tumor cells in surviving the hypoxic intratumoral environment.

A role for alteration of cell fate in neural progenitors? Although there is no current evidence that Cic loss of function can alter cell fate choices in neural stem or progenitors to promote the formation of oligodendrogliomas, that Cic could have such a developmental and oncogenic role is an intriguing possibility. Events during cerebral cortical development may represent a temporal window for ODG genesis. Although most of the ODGs are diagnosed during the early to mid-adulthood, the initial transformative events likely occur years before clinical presentation. Considering their indolent growth, younger demographic, and substantial tumor size at diagnosis, we postulate that there is a temporal window of susceptibility for ODG development that occurs while the brain is not yet fully mature.

During brain development, the tissue adjacent to the ventricles (the ventricular zone, VZ) constitutes a transient zone of proliferating neural precursors. The biology of these neural precursors has been reviewed extensively elsewhere [75] and is briefly summarized here. Early nestin$^+$ neural stem cells called neuroepithelial cells first expand by symmetric cell divisions to increase the progenitor pool. These neuroepithelial cells later either transform into radial glial cells or give rise to intermediate neural progenitors. Radial glia continue to express nestin, but the transition is marked by increased expression of glial markers (e.g., GFAP, vimentin, BLBP, GLAST, S100b) and a change in ultrastructure (e.g., glycogen granules, vascular end feet). Radial glia have potential to give rise to all lineages, i.e., neurons, astrocytes, and oligodendrocytes in a temporally dependent manner. In the mouse, OPC specification and proliferation in the forebrain begins in utero, with the bulk of oligodendrocyte production occurring in the first 2 postnatal weeks, before it declines in adulthood [76, 77]. Although most radial glia terminally differentiate into astrocytes and oligodendrocytes [75], some persist in adulthood as specialized neural stem cells,[78–81]. In humans, neuronogenesis and gliogenesis similarly unfolds over an extended period from mid-gestation through the first two decades before declining [76, 77]. It is possible that during this period of pre-adult glial specification, proliferation, and differentiation, acquired mutations in *IDH* and *CIC* initiate molecular and cellular events that lead to ODG formation.

Several lines of evidence highlight similarities between ODG and OPCs, and either implicate OPCs as a cell-of-origin for ODG, or suggest that OPC-like lineage restriction is a prerequisite step in the formation of ODGs from earlier progenitors. Oligodendrocytes arise from actively proliferating OPCs that are found in the stem-cell rich SVZ and white matter of brain of mammalian brain [82, 83]. An OPC fate is specified by the transcription factors Ascl1 and Olig2, which are also highly expressed in ODGs [84–88]. Moreover, both OPCs and ODGs express platelet-derived growth factor (PDGF), PDGF receptor (PDGFR), and neural/glial antigen 2 (NG2), which control OPC differentiation [3, 89]. In a transgenic model in which EGFR is activated in glial cells, ODG-like tumors arise from $NG2^+$ OPCs [90]. $NG2^+$ cells in both human ODG and this mouse model of also show increased tumorigenicity compared to $NG2^-$ cells, all suggesting that the biology of OPC specification or regulation of OPC proliferation are intimately linked to the genesis of ODGs.

Our work and that of others further supports a unique role for RAS/ERK signaling as a determinant of OPC specification and modulator of OPC proliferation [91, 92]. In addition, our findings and others' suggest that ETV5 is an important downstream mediator of RAS/ERK-induced glial/OPC specification and proliferation in the mammalian brain [91, 92]. Thus, in the brain, loss of CIC-mediated transcriptional repression on ETV5 or other targets may deregulate cellular responses specifically controlling OPC differentiation and proliferation, leading to ODG initiation. Studies investigating the biological functions of CIC in the context of normal brain development thus have potential to provide mechanistic insights regarding the importance of CIC loss in ODG genesis.

5 Conclusion and Future Perspectives

A growing body of evidence suggests that CIC is a tumor suppressor and its loss is one of the potential drivers of cancer development and progression in distinctive subset of human malignancies. The recent discovery of *CIC* and *IDH* as recurrently mutated genes in ODG brings new hope for understanding the origin and therapeutic vulnerabilities of ODG. At present, there are no targeted therapies for ODG, and the relative dearth of representative in vivo models for ODG has been a limitation for basic investigations into ODG biology and for translational and therapeutic drug discovery studies [93]. Investigating further the functions of *CIC* in the setting of wild-type and mutant IDH in neural progenitors would begin to address these needs and build the foundational knowledge necessary for the development of more targeted therapies. Finally, although we focus on ODG, the Ras/ERK signaling pathway is one of the most commonly dysregulated pathways in human cancer. The knowledge gained from the work on *CIC* as downstream repressor of RTK pathway may extend well beyond ODG and neural progenitors, and could ultimately be relevant to a broad range of cancers in the brain and beyond.

References

1. Cohen N, Weller RO. Who classification of tumours of the central nervous system. New York: Wiley Online Library; 2007.
2. Cairncross JG, et al. Specific genetic predictors of chemotherapeutic response and survival in patients with anaplastic oligodendrogliomas. J Natl Cancer Inst. 1998;90(19):1473–9.
3. Cairncross G, Jenkins R. Gliomas with 1p/19q codeletion: aka oligodendroglioma. Cancer J. 2008;14(6):352–7.
4. Yan H, et al. IDH1 and IDH2 mutations in gliomas. N Engl J Med. 2009;360(8):765–73.
5. Yip S, et al. Concurrent CIC mutations, IDH mutations, and 1p/19q loss distinguish oligodendrogliomas from other cancers. J Pathol. 2012;226(1):7–16.
6. Bettegowda C, et al. Mutations in CIC and FUBP1 contribute to human oligodendroglioma. Science. 2011;333(6048):1453–5.
7. Jiao Y, et al. Frequent ATRX, CIC, and FUBP1 mutations refine the classification of malignant gliomas. Oncotarget. 2012;3(7):709–22.
8. Sahm F, et al. CIC and FUBP1 mutations in oligodendrogliomas, oligoastrocytomas and astrocytomas. Acta Neuropathol. 2012;123(6):853–60.
9. Reifenberger G, et al. *Oligodendroglioma*. In: Louis DN et al., editors. In *WHO Classification of Tumours of the Central Nervous System*. Lyon: International Agency for Research on Cancer (IARC); 2007. p. 54–62.
10. Zlatescu MC, et al. Tumor location and growth pattern correlate with genetic signature in oligodendroglial neoplasms. Cancer Res. 2001;61(18):6713–5.
11. Lwin Z, Gan HK, Mason WP. Low-grade oligodendroglioma: current treatments and future hopes. Expert Rev Anticancer Ther. 2009;9(11):1651–61.
12. Van den Bent MJ, et al. Oligodendroglioma. Crit Rev Oncol Hematol. 2008;66(3):262–72.
13. Mason W, Louis DN, Cairncross JG. Chemosensitive gliomas in adults: which ones and why? J Clin Oncol. 1997;15(12):3423–6.
14. Cairncross G, et al. Phase III trial of chemoradiotherapy for anaplastic oligodendroglioma: long-term results of RTOG 9402. J Clin Oncol. 2013;31(3):337–43.
15. Guo C, et al. Isocitrate dehydrogenase mutations in gliomas: mechanisms, biomarkers and therapeutic target. Curr Opin Neurol. 2011;24(6):648–52.
16. Dang L, et al. Cancer-associated IDH1 mutations produce 2-hydroxyglutarate. Nature. 2009;462(7274):739–44.
17. Noushmehr H, et al. Identification of a CpG island methylator phenotype that defines a distinct subgroup of glioma. Cancer Cell. 2010;17(5):510–22.
18. Turcan S, et al. IDH1 mutation is sufficient to establish the glioma hypermethylator phenotype. Nature. 2012;483(7390):479–83.
19. Duncan CG, et al. A heterozygous IDH1R132H/WT mutation induces genome-wide alterations in DNA methylation. Genome Res. 2012;22(12):2339–55.
20. Cairns RA, Mak TW. Oncogenic isocitrate dehydrogenase mutations: mechanisms, models, and clinical opportunities. Cancer Discov. 2013;3(7):730–41.
21. Jenkins RB, et al. A t (1; 19)(q10; p10) mediates the combined deletions of 1p and 19q and predicts a better prognosis of patients with oligodendroglioma. Cancer Res. 2006;66(20): 9852–61.
22. Griffin CA, et al. Identification of der (1; 19)(q10; p10) in five oligodendrogliomas suggests mechanism of concurrent 1p and 19q loss. J Neuropathol Exp Neurol. 2006;65(10):988–94.
23. Idbaih A, et al. Two types of chromosome 1p losses with opposite significance in gliomas. Ann Neurol. 2005;58(3):483–7.
24. Ren X, et al. Co-polysomy of chromosome 1q and 19p predicts worse prognosis in 1p/19q codeleted oligodendroglial tumors: FISH analysis of 148 consecutive cases. Neuro Oncol. 2013;15(9):1244–50.
25. Huse JT, Phillips HS, Brennan CW. Molecular subclassification of diffuse gliomas: seeing order in the chaos. Glia. 2011;59(8):1190–9.

26. Okamoto Y, et al. Population-based study on incidence, survival rates, and genetic alterations of low-grade diffuse astrocytomas and oligodendrogliomas. Acta Neuropathol. 2004;108(1):49–56.
27. Kannan K, et al. Whole-exome sequencing identifies ATRX mutation as a key molecular determinant in lower-grade glioma. Oncotarget. 2012;3(10):1194–203.
28. Davoli T, et al. Cumulative haploinsufficiency and triplosensitivity drive aneuploidy patterns and shape the cancer genome. Cell. 2013;155(4):948–62.
29. Mizoguchi M, et al. Molecular characteristics of glioblastoma with 1p/19q co-deletion. Brain Tumor Pathol. 2012;29(3):148–53.
30. Miller CR, et al. Significance of necrosis in grading of oligodendroglial neoplasms: a clinicopathologic and genetic study of newly diagnosed high-grade gliomas. J Clin Oncol. 2006;24(34):5419–26.
31. Kuo LT, et al. Correlation among pathology, genetic and epigenetic profiles, and clinical outcome in oligodendroglial tumors. Int J Cancer. 2009;124(12):2872–9.
32. Appin CL, et al. Glioblastoma with oligodendroglioma component (GBM-O): molecular genetic and clinical characteristics. Brain Pathol. 2013;23(4):454–61.
33. Jimenez G, et al. Relief of gene repression by torso RTK signaling: role of capicua in Drosophila terminal and dorsoventral patterning. Genes Dev. 2000;14(2):224–31.
34. Goff DJ, Nilson LA, Morisato D. Establishment of dorsal-ventral polarity of the Drosophila egg requires capicua action in ovarian follicle cells. Development. 2001;128(22):4553–62.
35. Atkey MR, et al. Capicua regulates follicle cell fate in the Drosophila ovary through repression of mirror. Development. 2006;133(11):2115–23.
36. Lohr U, et al. Antagonistic action of Bicoid and the repressor Capicua determines the spatial limits of Drosophila head gene expression domains. Proc Natl Acad Sci U S A. 2009; 106(51):21695–700.
37. Roch F, Jimenez G, Casanova J. EGFR signalling inhibits Capicua-dependent repression during specification of Drosophila wing veins. Development. 2002;129(4):993–1002.
38. Ajuria L, et al. Capicua DNA-binding sites are general response elements for RTK signaling in Drosophila. Development. 2011;138(5):915–24.
39. Jimenez G, Shvartsman SY, Paroush Z. The Capicua repressor–a general sensor of RTK signaling in development and disease. J Cell Sci. 2012;125(Pt 6):1383–91.
40. Kawamura-Saito M, et al. Fusion between CIC and DUX4 up-regulates PEA3 family genes in Ewing-like sarcomas with t(4;19)(q35;q13) translocation. Hum Mol Genet. 2006;15(13): 2125–37.
41. Lee CJ, et al. CIC, a member of a novel subfamily of the HMG-box superfamily, is transiently expressed in developing granule neurons. Brain Res Mol Brain Res. 2002;106(1–2):151–6.
42. Lam YC, et al. ATAXIN-1 interacts with the repressor Capicua in its native complex to cause SCA1 neuropathology. Cell. 2006;127(7):1335–47.
43. Dissanayake K, et al. ERK/p90(RSK)/14-3-3 signalling has an impact on expression of PEA3 Ets transcription factors via the transcriptional repressor capicua. Biochem J. 2011;433(3):515–25.
44. O'Neill EM, et al. The activities of two Ets-related transcription factors required for Drosophila eye development are modulated by the Ras/MAPK pathway. Cell. 1994;78(1):137–47.
45. Lee Y, et al. ATXN1 protein family and CIC regulate extracellular matrix remodeling and lung alveolarization. Dev Cell. 2011;21(4):746–57.
46. Chittaranjan S, et al. Mutations in CIC and IDH1 cooperatively regulate 2-hydroxyglutarate levels and cell clonogenicity. Oncotarget. 2014;5(17):7960.
47. Astigarraga S, et al. A MAPK docking site is critical for downregulation of Capicua by Torso and EGFR RTK signaling. EMBO J. 2007;26(3):668–77.
48. Cinnamon E, Paroush Z. Context-dependent regulation of Groucho/TLE-mediated repression. Curr Opin Genet Dev. 2008;18(5):435–40.
49. Paroush Z, Wainwright SM, Ish-Horowicz D. Torso signalling regulates terminal patterning in Drosophila by antagonising Groucho-mediated repression. Development. 1997;124(19): 3827–34.
50. Olsen JV, et al. Global, in vivo, and site-specific phosphorylation dynamics in signaling networks. Cell. 2006;127(3):635–48.

51. Fryer JD, et al. Exercise and genetic rescue of SCA1 via the transcriptional repressor Capicua. Science. 2011;334(6056):690–3.
52. Tseng AS, et al. Capicua regulates cell proliferation downstream of the receptor tyrosine kinase/ras signaling pathway. Curr Biol. 2007;17(8):728–33.
53. Morrison DK. The 14-3-3 proteins: integrators of diverse signaling cues that impact cell fate and cancer development. Trends Cell Biol. 2009;19(1):16–23.
54. Kim Y, et al. MAPK substrate competition integrates patterning signals in the *drosophila* embryo. Curr Biol. 2010;20(5):446–51.
55. Kim Y, et al. Gene regulation by MAPK substrate competition. Dev Cell. 2011;20(6):880–7.
56. Vogelstein B et al. (2013) Cancer genome landscapes. Science 339(6127): 1546–58
57. Cinnamon E, et al. Capicua integrates input from two maternal systems in Drosophila terminal patterning. EMBO J. 2004;23(23):4571–82.
58. Heras d l JM, Casanova J. Spatially distinct downregulation of Capicua repression and tailless activation by the Torso RTK pathway in the Drosophila embryo. Mech Dev. 2006;123(6): 481–6.
59. Weiss JB, et al. Dorsoventral patterning in the Drosophila central nervous system: the intermediate neuroblasts defective homeobox gene specifies intermediate column identity. Genes Dev. 1998;12(22):3591–602.
60. Jiang H, et al. EGFR/Ras/MAPK signaling mediates adult midgut epithelial homeostasis and regeneration in Drosophila. Cell Stem Cell. 2011;8(1):84–95.
61. Kurpios NA, et al. Function of PEA3 Ets transcription factors in mammary gland development and oncogenesis. J Mammary Gland Biol Neoplasia. 2003;8(2):177–90.
62. Kar A, Gutierrez-Hartmann A. Molecular mechanisms of ETS transcription factor-mediated tumorigenesis. Crit Rev Biochem Mol Biol. 2013;48(6):522–43.
63. Italiano A, et al. High prevalence of CIC fusion with double-homeobox (DUX4) transcription factors in EWSR1-negative undifferentiated small blue round cell sarcomas. Genes Chromosomes Cancer. 2012;51(3):207–18.
64. Graham C, et al. The CIC-DUX4 fusion transcript is present in a subgroup of pediatric primitive round cell sarcomas. Hum Pathol. 2012;43(2):180–9.
65. de Launoit Y, et al. The Ets transcription factors of the PEA3 group: transcriptional regulators in metastasis. -Reviews on. Biochim Biophys Acta. 2006;1766(1):79–87.
66. Kessenbrock K, Plaks V, Werb Z. Matrix metalloproteinases: regulators of the tumor microenvironment. Cell. 2010;141(1):52–67.
67. Krivy K, Bradley-Gill M-R, Moon N-S. Capicua regulates proliferation and survival of RB-deficient cells in Drosophila. Biol Open. 2012;2(2):183–90. BIO20123277.
68. Beckner ME, et al. Identification of ATP citrate lyase as a positive regulator of glycolytic function in glioblastomas. Int J Cancer. 2010;126(10):2282–95.
69. Zaidi N, Swinnen JV, Smans K. ATP-citrate lyase: a key player in cancer metabolism. Cancer Res. 2012;72(15):3709–14.
70. Migita T, et al. ATP citrate lyase: activation and therapeutic implications in non-small cell lung cancer. Cancer Res. 2008;68(20):8547–54.
71. Turyn J, et al. Increased activity of glycerol 3-phosphate dehydrogenase and other lipogenic enzymes in human bladder cancer. Horm Metab Res. 2003;35(10):565–9.
72. Hanai J-i, et al. Inhibition of lung cancer growth: ATP citrate lyase knockdown and statin treatment leads to dual blockade of mitogen-activated protein Kinase (MAPK) and Phosphatidylinositol-3-kinase (PI3K)/AKT pathways. J Cell Physiol. 2012;227(4):1709–20.
73. Bauer DE, et al. ATP citrate lyase is an important component of cell growth and transformation. Oncogene. 2005;24(41):6314–22.
74. Udpa N, et al. Whole genome sequencing of Ethiopian highlanders reveals conserved hypoxia tolerance genes. Genome Biol. 2014;15(2):R36.
75. Kriegstein A, Alvarez-Buylla A. The glial nature of embryonic and adult neural stem cells. Annu Rev Neurosci. 2009;32:149–84.
76. Jakovcevski I, et al. Oligodendrocyte development and the onset of myelination in the human fetal brain. Front Neuroanat. 2009;3:5.

77. Yakovlev PI, Lecours AR. The myelinogenetic cycles in regional maturation of the brain. In: Minkowsky A, editor. Regional development of the brain in early life. Oxford: Blackwell; 1967. p. 3–70.
78. Noctor SC, et al. Dividing precursor cells of the embryonic cortical ventricular zone have morphological and molecular characteristics of radial glia. J Neurosci. 2002;22(8):3161–73.
79. Anthony TE, et al. Radial glia serve as neuronal progenitors in all regions of the central nervous system. Neuron. 2004;41(6):881–90.
80. Strand AD, et al. Conservation of regional gene expression in mouse and human brain. PLoS Genet. 2007;3(4):e59.
81. Rakheja D, et al. The emerging role of d-2-hydroxyglutarate as an oncometabolite in hematolymphoid and central nervous system neoplasms. Front Oncol. 2013;3:169.
82. Geha S, et al. NG2+/Olig2+cells are the major cycle-related cell population of the adult human normal brain. Brain Pathol. 2010;20(2):399–411.
83. Menn B, et al. Origin of oligodendrocytes in the subventricular zone of the adult brain. J Neurosci. 2006;26(30):7907–18.
84. Rousseau A, et al. Expression of oligodendroglial and astrocytic lineage markers in diffuse gliomas: use of YKL-40, ApoE, ASCL1, and NKX2-2. J Neuropathol Exp Neurol. 2006;65(12):1149–56.
85. Ligon KL, et al. The oligodendroglial lineage marker OLIG2 is universally expressed in diffuse gliomas. J Neuropathol Exp Neurol. 2004;63(5):499–509.
86. Lu QR, et al. Oligodendrocyte lineage genes (OLIG) as molecular markers for human glial brain tumors. Proc Natl Acad Sci U S A. 2001;98(19):10851–6.
87. Casarosa S, Fode C, Guillemot F. Mash1 regulates neurogenesis in the ventral telencephalon. Development. 1999;126(3):525–34.
88. Zhou Q, Wang S, Anderson DJ. Identification of a novel family of oligodendrocyte lineage-specific basic helix-loop-helix transcription factors. Neuron. 2000;25(2):331–43.
89. Shoshan Y, et al. Expression of oligodendrocyte progenitor cell antigens by gliomas: implications for the histogenesis of brain tumors. Proc Natl Acad Sci U S A. 1999;96(18):10361–6.
90. Persson AI, et al. Non-stem cell origin for oligodendroglioma. Cancer Cell. 2010;18(6): 669–82.
91. Li X, et al. MEK is a key regulator of gliogenesis in the developing brain. Neuron. 2012;75(6):1035–50.
92. Li S, et al. RAS/ERK signaling controls proneural genetic programs in cortical development and gliomagenesis. J Neurosci. 2014;34(6):2169–90.
93. Kelly JJ, et al. Oligodendroglioma cell lines containing t(1;19)(q10;p10). Neuro Oncol. 2010;12(7):745–55.

Isocitrate Dehydrogenase (IDH) Mutation in Gliomas

Charles Chesnelong

Abstract First identified in 2006 in a colorectal cancer sequencing effort, Isocitrate Dehydrogenase (IDH) mutations were later reported in secondary glioblastomas by Parsons et al. leading the way for further studies which have revealed the presence of mutations in either IDH1 or IDH2 in over 70 % of grade II–III gliomas and secondary glioblastomas. In the clinic, IDH1 and IDH2 mutations are important prognostic factors associated with prolonged survival and enhanced radio- and chemo-sensitivity. At the benchside, IDH mutations are a major focus of glioma research. Significant progresses have been made elucidating the roles of IDH mutations in tumorigenesis. IDH mutations were shown to confer a neomorphic enzymatic activity: the reduction of α-Ketoglutarate (αKG) to 2-hydroxyglutarate (2HG). 2HG was further shown to be the main mediator of the oncogenic effects of IDH mutation leading to epigenetic alterations, extracellular matrix remodeling, and hypoxia-inducible factor 1α (HIF1a) degradation. However, many aspects remain unclear such as the potential influence of IDH mutations on cancer cell metabolism and whether IDH mutations, despite increasingly well-characterized oncogenic mechanisms, may also trigger pro-survival effects. Elucidating the roles of mutant IDH enzymes in tumorigenesis will significantly improve our understanding of glioma biology and will lead to novel therapeutic strategies that should aim to disrupt the oncogenic properties of IDH mutations while promoting properties that may contribute to the slower growth, enhanced sensitivity to conventional therapies and overall longer survival characteristic of IDH mutant gliomas.

1 Introduction

Gliomas are the most common form of brain tumor. The World Health Organization (WHO) classifies gliomas into three subtypes according to tumor cell morphology. They consist of astrocytomas, oligodendrogliomas, and mixed oligoastrocytomas.

C. Chesnelong (✉)
South Alberta Cancer Research Institute (SACRI), Cumming Medical School, University of Calgary, Calgary, AB, Canada
e-mail: cchesnel@ucalgary.ca

W. Wu, H. Choudhry (eds.), *Next Generation Sequencing in Cancer Research, Volume 2*, DOI 10.1007/978-3-319-15811-2_25

Histological grading further stratifies those subtypes. Indolent gliomas are considered grade I–II, while gliomas featuring nuclear atypia, dense cellularity, and elevated mitotic activity are classified as grade III. Grade III are invasive tumors that often progress to grade IV tumors (Glioblastomas: GBM) characterized by microvascular proliferation and necrosis. Importantly, primary GBM arising de novo have to be distinguished from secondary GBM arising from the malignant progression of lower grade tumors since these two subtypes appear to have a very distinct evolution based on their respective genetic alterations.

Although the WHO classification system has proven considerably useful in the clinic, a striking clinical heterogeneity is still found within WHO subtypes and especially within GBMs. The development of high-throughput genomic technology and large-scale profiling efforts like those of The Cancer Genome Atlas (TCGA) has dramatically helped the characterization of molecular alterations in gliomas and increased our understanding of the biology behind gliomagenesis (Fig. 1). Recent studies have also highlighted the notion that "molecular subclasses" exist in gliomas. These subclasses may be biologically relevant and may better characterize the clinical behavior of these tumors [1, 2]. These findings suggest that each glioma subclass may represent a distinct oncogenic mechanism arising in distinct pools of precursor cells. Most importantly, these studies also suggest that the different glioma subclasses may need to be treated with different targeted therapies in order to improve clinical outcome.

Amongst all the genetic alterations found in gliomas, IDH mutations stand out. Clinically, glioma patients bearing an IDH mutation tend to be younger and have a much better prognosis [3]. Indeed, unlike IDH wild type gliomas, which are typically treatment-resistant and fast-growing cancers, gliomas harboring IDH mutations have a better prognosis and grow relatively slowly [4]. As an extreme example, oligodendrogliomas, also characterized by 1p/19q co-deletion and frequent mutations of FUBP1 and CIC [5, 6], could be considered as the prototypical IDH mutant cancer since virtually 100 % are IDH mutant. Interestingly oligodendrogliomas display longer survival and enhanced radio- and chemo-sensitivity [7].

Mutations of IDH1 and IDH2 are mutually exclusive and strictly heterozygous. IDH mutations are present in over 70 % of grade II–III gliomas and secondary glioblastomas [8, 9]. IDH1 mutation is much more frequent than IDH2 mutation in gliomas and unlike TP53 mutations and 1p/19q loss, IDH mutations are found in tumors of both astrocytic and oligodendroglial lineages. Moreover, there is no report of an IDH mutation occurring after the acquisition of TP53 mutation or loss of 1p/19q suggesting that IDH1 and IDH2 mutations are early events that probably arise in a common precursor cell of astrocytic and oligodendroglia tumors (i.e., "glioma cell of origin") [10]. Interestingly, it is currently argued that virtually every lower grade glioma and secondary GBM may in fact be IDH mutant while none of the primary GBM would be [11]. The discrepancies could be due to misdiagnosis of primary GBM, secondary GBM and grade III astrocytomas. This school of thought is coherent with the evolution of the disease based on genomic alterations (Fig. 1). Noteworthy, IDH mutations have also been identified in ~17 % of patients with Acute Myeloid Leukemia (AML) where it does not predict better survival but is instead associated

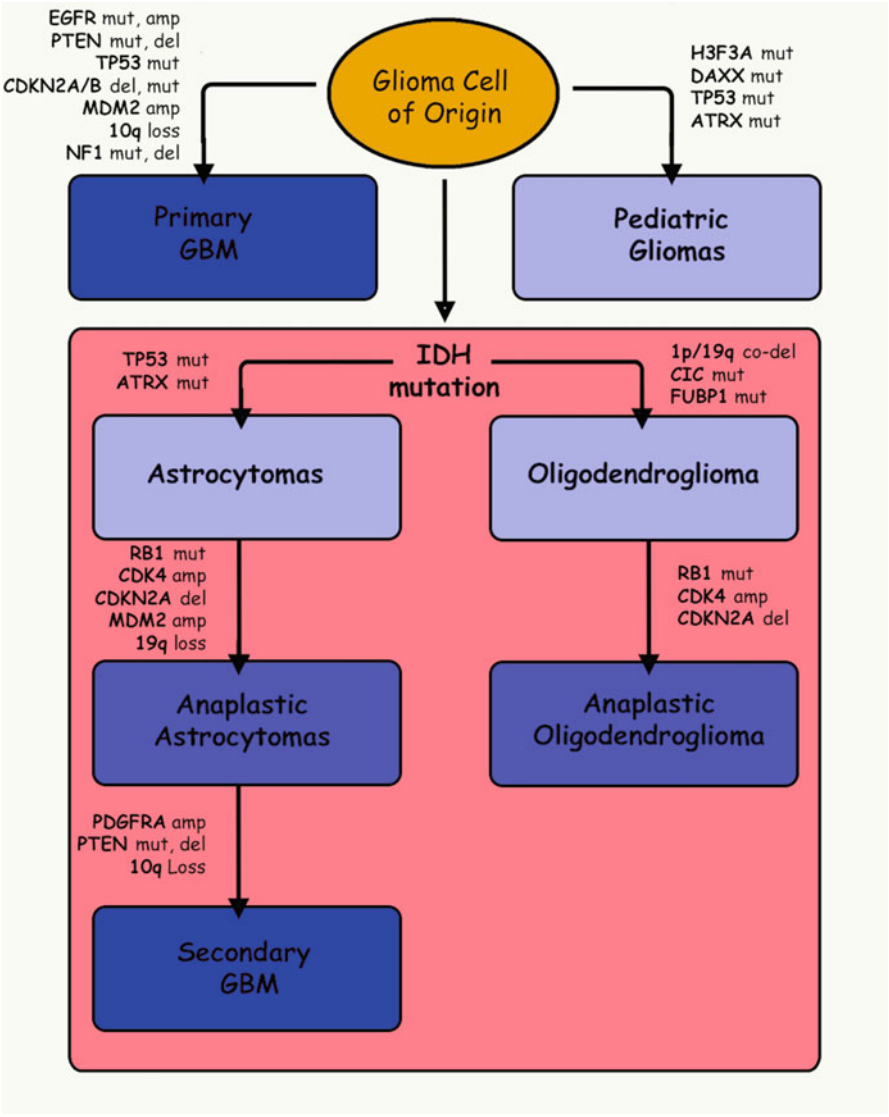


Fig. 1 Glioma subtypes and molecular alterations. Representation of glioma subtypes and the major associated molecular alterations. IDH mutation is the earliest genetic event found in lower grade gliomas and secondary GBMs. IDH mutation is thought to arise in the glioma cell of origin and be the initial event promoting tumorigenesis

with worse prognosis in a subset of patients [12–16]. Of note, IDH mutations are also detected at high frequency in several cartilaginous tumors [17, 18] and in many other cancers, albeit at much lower frequencies [19–25]. It is now clear that IDH mutations are key events in gliomagenesis and in the development of AML.

However, it remains unknown why they are associated with prolonged survival in glioma but not in AML.

In this chapter, we will discuss IDH mutations, their roles in tumorigenesis, their potential pro-survival effects in gliomas and the therapeutic strategies being developed for IDH mutant gliomas.

2 IDH Mutation

The most common IDH mutations affect residue R132 (IDH1), its analogous residue R172 (IDH2) and the non-analogous R140 (IDH2). Several IDH1^{R132} variants have been identified (R132H, R132C, R132G, R132S, and R132L), R132H being the most common (~90 % in gliomas, 50 % in AML). Likewise, three IDH2^{R140} variants have been identified (R140Q, R140L, and R140W), R140Q being the predominant one in AML (~95 %). Finally, different IDH2^{R172} variants have also been detected, R172K representing the large majority of cases. These mutated arginine residues fall within the catalytic domain involved in isocitrate binding, leading to the initial assumption that these mutations affect IDH enzymatic activity and result in loss of function. Further biochemical studies established that mutant IDH1 and IDH2 enzymes are unable to efficiently catalyze the oxidative decarboxylation of isocitrate [26]. It was then quickly hypothesized that IDH mutations may induce a disruption of the TCA Cycle, promoting the metabolic shift toward aerobic glycolysis as proposed by Otto Warburg over 80 years ago (Warburg effect). This created a lot of enthusiasm in the glioma research community and more globally in the field of cancer metabolism and energetics, a field that had barely progressed since Warburg [27–29] and the more recent discovery of mutations affecting the Fumarate Hydratase (FH) and Succinate Dehydrogenase (SDH) in cancer [30–32]. However, unlike FH and SDH, IDH1 and 2 are not directly involved in the TCA cycle. The IDH family consists of IDH1, 2, and 3. All catalyze the same oxidative decarboxylation of isocitrate to produce CO_2 and α-ketoglutarate (αKG), using NADP+ (IDH1 and IDH2) or NAD+ (IDH3) as electron acceptors and generating NADPH or NADH, respectively. Furthermore, while IDH1 and IDH2 are similar homodimeric enzymes, IDH3 is a heterotetrameric enzyme. Another fundamental difference between the three enzymes resides in their subcellular localization. IDH1 is localized in both peroxisomes and the cytosol, whereas IDH2 and IDH3 localize to mitochondria. Although IDH3 is well characterized and is the isocitrate dehydrogenase directly involved in the TCA cycle, the exact roles of IDH1 and 2 in cellular metabolism remains unclear. Importantly, while IDH1 and IDH2 mutations converge functionally, there is no report to date of mutations in any genes encoding IDH3 subunits, suggesting that IDH mutations do not directly affect the TCA cycle.

Despite the fact that these mutations were first recognized as loss of function mutations, a pivotal study reported soon afterwards that substitutions of IDH1^{R132} or IDH2^{R172} results in a loss of affinity for isocitrate along with an increased affinity for αKG and NADPH leading to a neomorphic enzymatic activity of the mutated

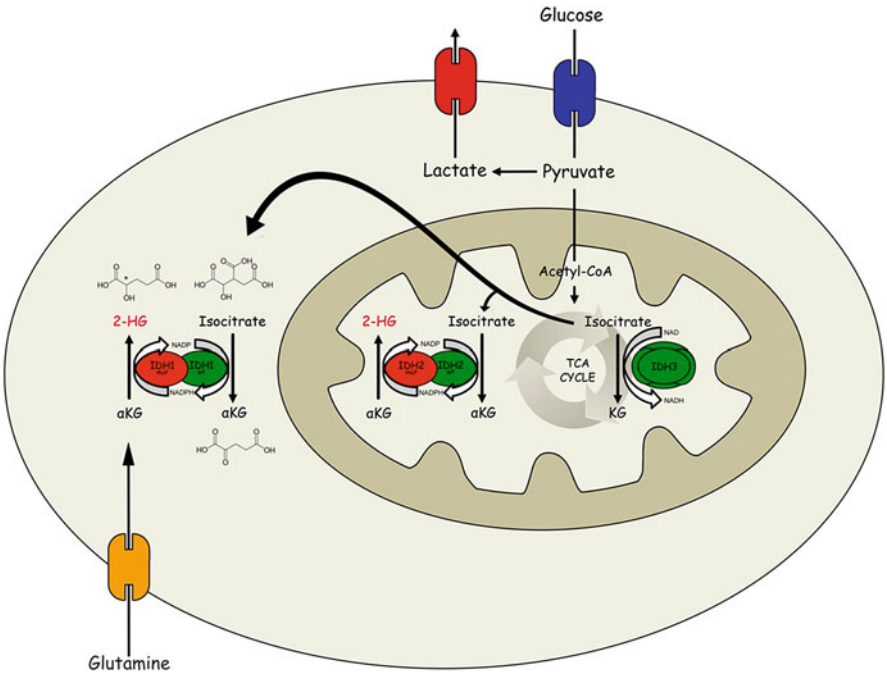


Fig. 2 IDH mutation. Representation of IDH1 and IDH2 mutant enzymes, their subcellular localization and enzymatic activity. IDH mutations confer a neomorphic enzyme activity: the reduction of αKG to 2-hydroxyglutarate (2HG) coupled with the oxidation of NADPH to NADP

enzyme; the reduction of αKG to 2-hydroxyglutarate (2HG) coupled with the oxidation of NADPH to NADP (Fig. 2) [33]. This neo-activity of the mutant enzyme has since been very well documented. With the exception of extremely rare IDH mutations in thyroid cancer that do not result in 2HG production, but can result in loss-of-function [34], IDH mutations affecting $IDH1^{R132}$, $IDH2^{R172}$, and $IDH2^{R140}$, as well as other less common mutations, were all shown to produce 2HG. All 2HG-producing IDH mutant enzymes were interestingly shown to exclusively produce the D-enantiomeric isoform of 2HG. However, these mutations are not necessarily equivalent. First, different diseases have varying frequencies of different IDH mutations. For example IDH1 mutation is predominant in gliomas while mutation of IDH2 is prevalent in AML. Moreover, $IDH2^{R140}$ is the residue most commonly mutated in AML, while mutations affecting $IDH2^{R140}$ remains to be reported in gliomas, which are more frequently mutated at $IDH2^{R172}$. Interestingly, it was recently shown that the different subcellular localization of IDH1 and IDH2 is extremely important for the production of 2HG and may influence the function of IDH mutation in various types of cancer. Indeed, αKG is present abundantly in the mitochondria allowing mutated IDH2 to produce 2HG efficiently and independently of the presence of the wild type enzyme. This is coherent with the rare but reported

homozygosity of mutant IDH2 [35]. On the contrary, IDH1 mutant enzyme is substrate-limited. Indeed, αKG is not abundant in the cytoplasm forcing mutated IDH1 to completely rely on its wild type counterpart to produce 2HG explaining why IDH1 mutation is strictly heterozygous [36, 37]. Moreover, it was shown that different mutations vary in their ability to produce 2HG. As an example, $IDH2^{R172}$ mutations consistently lead to greater 2HG production than $IDH2^{R140}$ mutations [36]. Thus, IDH subcellular localization and specific mutation can affect 2HG production and may explain differences in mechanisms, roles, and impact on prognosis in various IDH mutant cancers.

3 Role of IDH Mutation in Tumorigenesis

3.1 NADPH/NADP+ Balance

IDH enzymes are the most important producers of NADPH. As such, IDH mutations may alter the cellular NADPH/NADP+ ratio, which can lead to mitochondrial dysfunction and disruption of metabolic pathways dependent on a specific NADPH/NADP+ ratio. The alteration of the NADPH/NADP+ balance also causes a disruption of the potential of reduction of Reactive Oxygen Species (ROS) leading to increased oxidative stress (Fig. 3), causing DNA damage further promoting malignant transformation [38, 39]. Interestingly, it was also suggested that 2HG itself may promote oxidative stress and contribute to this effect [40, 41].

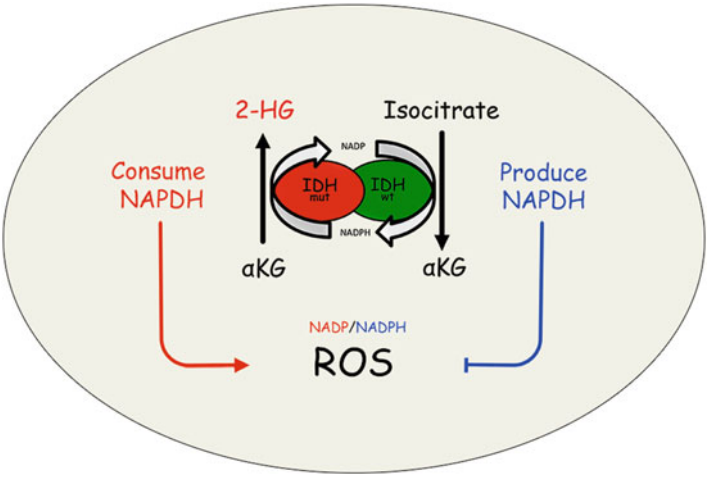


Fig. 3 IDH mutant enzymes promote oxidative stress. IDH mutant enzymes alter the cellular NADPH/NADP+ balance, causing a disruption of the potential of reduction of Reactive Oxygen Species (ROS) leading to increased oxidative stress. This can lead to the promotion of DNA damage but also sensitization to therapy

3.2 α-Ketoglutarate (αKG)

While wild type IDH enzyme produces αKG, the mutant enzyme consumes αKG. Interestingly, decreased αKG levels may impact the activity of several αKG-dependent dioxygenases; dioxygenases use αKG as substrate to catalyze many reactions involved in essential processes such as DNA repair, epigenetics, oxygen sensing, hypoxia adaptation, extracellular matrix remodeling, and fatty-acid metabolism [42]. Importantly, αKG is also an essential intermediary metabolite of the TCA cycle. Decreased αKG levels could thus result in disruption of the TCA cycle and promotion of the glycolytic shift. However, no decrease in αKG levels has been confirmed in IDH mutant cells. Instead, IDH mutant cells were shown to maintain αKG levels by increasing glutamine degradation. This has been proposed to trigger an addiction of the IDH mutant cells to glutamine, offering interesting therapeutic strategies for IDH mutant gliomas, which are illustrated by the increased sensitivity of IDH mutant cells toward Glutaminase inhibitors [43].

3.3 2-Hydroxyglutarate (2HG), an "Onco-Metabolite"

2HG is a direct product of mutant IDH enzymes, which can be detected in patients' blood, urine and brain with noninvasive methods [44–46]. As such, 2HG is an important biomarker for IDH mutant cancer and is under investigation for clinical purposes. More importantly, 2HG is directly linked to disease development. 2HG is a metabolite normally produced from αKG in a reaction coupled with the oxidation of γ-hydroxybutyrate into succinic semi-aldehyde by the hydroxyacid oxoacid transhydrogenase (HOT). Under physiological conditions, cellular levels of 2HG are maintained by hydroxyglutarate dehydrogenase (HGDH), producing αKG from 2HG. There are two enantiomers of 2HG: D-2HG and L-2HG. Alterations in the metabolism of both enantiomers are associated with pathological acidurias. Elevated D-2HG and L-2HG in urine, plasma, and CSF characterize rare neurometabolic disorders (2-hydroxyglutarate aciduria: D or L-2HGA) most often caused by mutations in d-*2HGDH* (or l-*2HGDH*) [47]. Interestingly, L-2HGA is associated with a higher risk of developing malignant brain tumors [48], suggesting that 2HG has oncogenic properties.

Recently, 2HG has been proposed to be an "onco-metabolite." The first evidence of a potential role of 2HG in cancer came from studies in primary neurons, in which 2HG activates the *N*-methyl-D-aspartic acid receptors, thereby disrupting Ca^{2+} homeostasis and increasing reactive oxygen species. In the same model, 2HG has been shown to inhibit the ATP synthase complex, resulting in mitochondrial dysfunction [49, 50]. More recently, several studies have shown that 2HG competitively inhibits several αKG-dependent enzymes [51]. 2HG and αKG being structurally very similar, it has been proposed that 2HG could competitively inhibit αKG-dependent enzymes. Interestingly, this converges toward our current understanding of the impact of alterations in SDH and FH [52, 53], which result

in elevated levels of succinate and fumarate, interfering with the activity of αKG-dependent dioxygenases. Via inhibition of αKG-dependent dioxygenases, 2HG may alter matrix remodeling, epigenetic regulation and metabolism.

3.3.1 Matrix Remodeling

The extracellular matrix (ECM) is an essential part of the tumor microenvironment. Disruption and/or modifications of the ECM structure and composition are essential for tumorigenesis and affect proliferation, differentiation and invasion. Interestingly, 2HG has been shown to inhibit collagen hydroxylases that are essential for collagen maturation potentially resulting in disturbed ECM. Interestingly, it was also proposed that 2HG, like lactate, may induce acidification of the extracellular environment causing ECM disturbance and promoting invasion. However, the importance and relevance of an ECM remodeling dependent on IDH mutation remains to be established.

3.3.2 Epigenetic Alterations

The most interesting and most described conseuence of IDH mutation to date is its impact on epigenetic regulation. Alteration of DNA methylation is a hallmark of human cancers. Typically, cancers present a global DNA hypomethylation and more local hypermethylated foci. Promoter CpG island hypermethylation is often observed in cancer and leads to transcriptional silencing of associated genes, typically tumor suppressors. A large-scale analysis of the epigenome from TCGA samples has identified glioma samples with DNA hypermethylated loci characterizing a Glioma CpG Island Methylator Phenotype (G-CIMP) [54]. G-CIMP-positive samples were tightly associated with IDH1 mutation and further studies have shown that IDH mutation is in fact directly responsible the establishment of G-CIMP. Indeed, it was shown that 2HG competitively inhibits α-KG-dependent 5-methylcytosine hydroxylases involved in DNA demethylation [55–57], resulting in promoter CpG island hypermethylation.

2HG also competitively inhibits αKG-dependent Jumonji-C domain Histone Demethylases (JHDMs) [58, 59], essential regulators of posttranslational histone tail methylation, necessary for proper regulation of chromatin organization and gene expression. Thus, through alterations of histones and DNA methylation, IDH mutation affects global and/or local DNA and histone methylation patterns, potentially altering the expression of oncogenes, tumor suppressors and metabolic genes. Consequently, via 2HG-dependent epigenetic alterations, IDH mutation may interfere with differentiation, proliferation, survival, and metabolism (Fig. 4). As an example, 2HG-dependent inhibition of histone demethylases was reported to be sufficient to block cell differentiation [58]. In addition, via promoter methylation, IDH is involved in the repression of essential metabolic genes, tumor suppressor and DNA repair genes affecting proliferation and survival [3, 60].

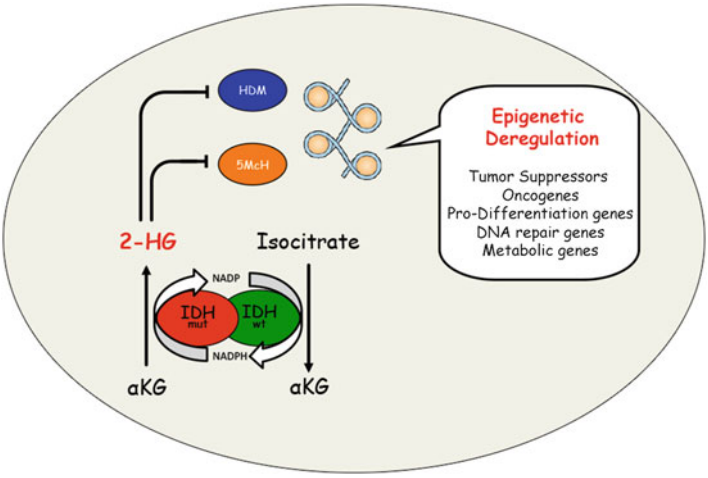


Fig. 4 IDH mutation leads to epigenetic alterations. 2HG competitively inhibits 5-methylcytosine hydroxylases (5McH) and Jumonji-C domain histone demethylases (HDM) leading to disorganization of chromatin structure and overall deregulation of gene expression. 2HG-dependent epigenetic alterations may promote tumorigenesis via deregulation of differentiation, proliferation, survival, and metabolism

3.3.3 Metabolism

IDH mutations were first hypothesized to directly impact energy metabolism and promote aerobic glycolysis, which is the metabolic pathway favored by cancer cells to support rapid proliferation [61]. However, IDH mutation has recently been shown to downregulate the hypoxia inducible factor-1α (HIF1α) pathway, leading to the repression of important glycolytic genes typically overexpressed in cancer. Indeed, EGLN prolyl 4-hydroxylase, which is responsible for the oxygen-dependent degradation of HIF1α, is an αKG-dependent enzyme whose activity was shown to be affected by 2HG. Although stabilization of HIF1α by 2HG was first reported (Fig. 5a) [26, 51], more recent studies have clarified that 2HG activates EGLN promoting HIF1α degradation [61–64] and downregulating HIF1α target genes [60], including many essential for aerobic glycolysis (Fig. 5b).

Interestingly, a HIF1α target gene essential for aerobic glycolysis, Lactate Dehydrogenase A (LDHA), was reported as silenced in IDH mutant glioma. Silencing of LDHA was shown to be associated with increased methylation of its promoter. The global downregulation of the HIF1α pathway through 2HG-dependent promotion of HIF1α degradation [62], in conjunction with the silencing of LDHA, was suggested to limit the glycolytic capacity of IDH mutant cells. Moreover, an α-ketoglutarate-dependent histone demethylase (JMJD2C), also potentially inhibited by 2HG, is required for the expression of several glycolysis essential genes downstream of HIF1α [65]. It thus appears that several mechanisms triggered by IDH mutant enzyme may act in concert to suppress glycolytic genes, suggesting

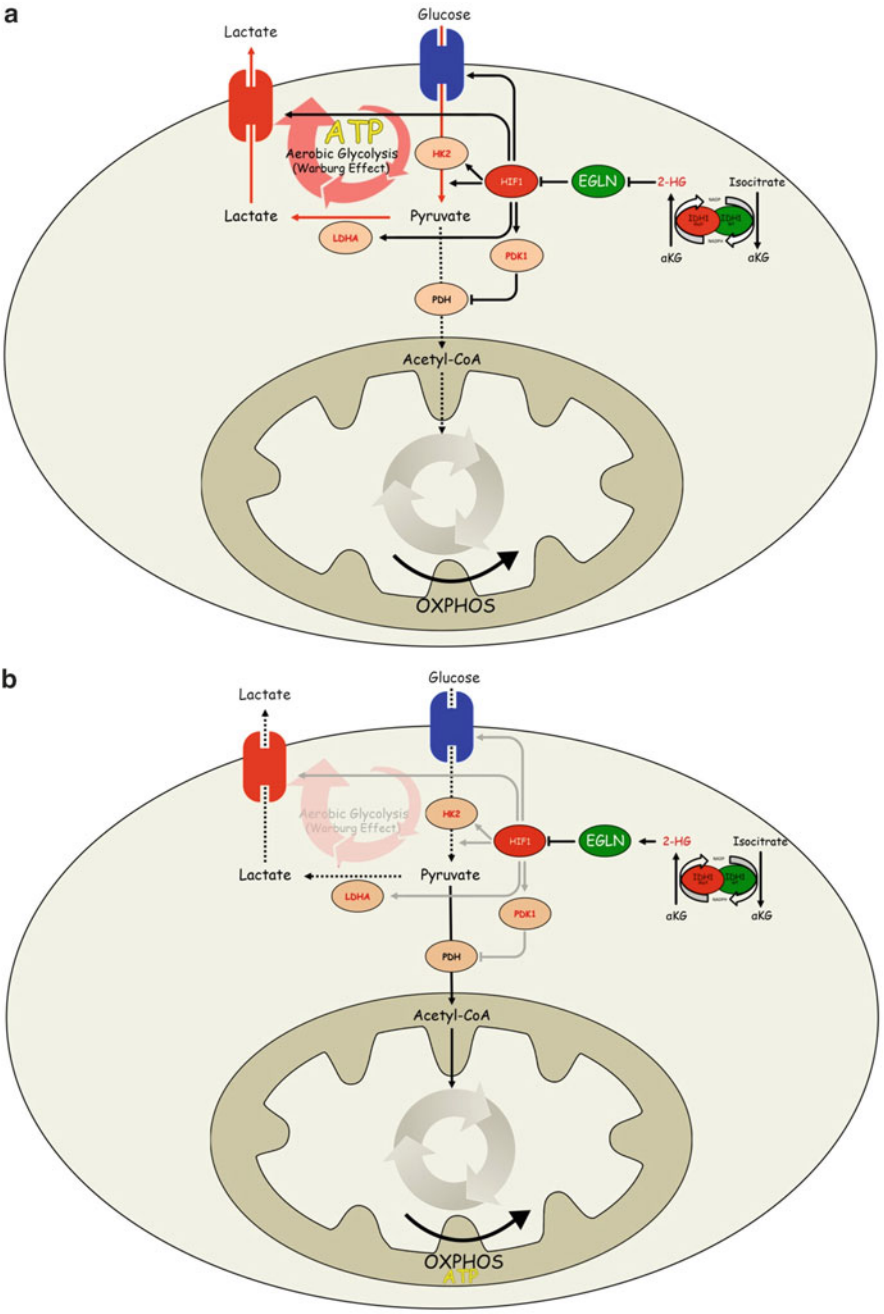


Fig. 5 IDH mutation and metabolism. (**a**) 2HG was first reported to inhibit EGLN prolyl 4-hydroxylase, resulting in stabilization of HIF1α and promotion of the Warburg effect. (**b**) However, more recent studies have clarified that 2HG activates EGLN, promoting HIF1α degradation and potentially preventing the glycolytic shift towards aerobic glycolysis

that IDH mutant gliomas may have selected defects in glycolysis. Consequently, IDH mutant gliomas may rely on oxidative phosphorylation for energy production although, to date, few elements support this [66, 67].

Downregulation of glycolytic genes in IDH mutant gliomas is counterintuitive. Indeed, IDH mutation playing a role in the prevention of the glycolytic switch would be contrary to classical steps toward tumorigenesis. This unique feature may help explain the slow disease progression and better survival of this subgroup of gliomas. However, one may ask how the downregulation of the HIF1α pathway makes sense in IDH mutant gliomas. Interestingly, although the L-enantiomer of 2HG (L-2HG) is a more potent inhibitor of αKG-dependent enzymes, 2HG-dependent transformation is specific to the D-enantiomer (D-2HG). The fact that only D-2HG induces transformation may be explained by differential effects of these two enantiomers on EGLN. Indeed, while D-2HG acts as an agonist of EGLN, L-2HG is an antagonist [64]. Moreover, a key experiment recently demonstrated that knockdown of EGLN blocks IDH mutant-dependent leukemogenesis [64]. Downregulation of the HIF1α pathway may thus be an essential characteristic of IDH mutant cancers. Furthermore, several studies have shown that HIF1α can inhibit hematopoietic stem cells (HSC) and leukemic cell proliferation [68–73], suggesting that the inhibition of HIF1α may be essential for the IDH mutant-dependent tumorigenic process. However, in gliomas, this may have the unique consequence of limiting the fundamental ability of cancer cells to undergo a metabolic shift towards aerobic glycolysis.

Noteworthy, IDH mutations may also promote other metabolic pathways and trigger dependence on specific fuel sources offering therapeutic opportunities (glutaminolysis, fatty acid metabolism, oxidative phosphorylations). Identifying and disrupting certain components of these metabolic pathways will provide intervention points for the discovery and development of novel therapeutics.

4 Pro-survival Effects of IDH Mutation

Glioma patients bearing an IDH mutation demonstrate longer survival independently of age and clinical performance status. Whether IDH mutations are directly responsible for this better prognosis remains an open question. However, several intriguing elements are raising the possibility that IDH mutation may carry pro-survival and sensitizing effects in glioma.

IDH mutant gliomas are less necrotic and show less frequent vascular abnormality. In addition, IDH mutant gliomas have less edema and present less enhancing lesions. IDH mutations could be directly involved in those favorable prognostic factors via its action on the HIF1α pathway. Indeed, downregulation of the HIF1α pathway could directly impact vascularization, edema, blood–brain barrier permeability and necrosis.

The impact of IDH mutation on growth and proliferation is uncertain. Indeed, engineering of several IDH mutant cell lines via overexpression of mutant IDH1 and 2 have shown contradictory results reporting both reduced and enhanced

proliferation and growth in vitro and in vivo. However, it is worth mentioning that patient-derived cells with endogenous expression of IDH mutant enzyme are typically refractory to in vitro culture conditions [74]. Moreover, several reported IDH1 mutant-derived cell lines present a loss of one of the IDH1 alleles. Two lines derived from IDH mutant gliomas were reported to present a loss of heterozygosity at the IDH1 locus affecting the mutant allele [60]. Another study reported an IDH mutant line [75] presenting a loss of the wild type IDH1 allele [76]. Interestingly, loss of IDH1 wild type allele has also been reported in vivo and may result in decreased 2HG [36, 37]. These reports, although not conclusive, argue for a selection pressure against the IDH mutant phenotype at least in vitro.

Finally, IDH mutations in gliomas predict response to Temozolomide (TMZ) and Ionizing Radiation (IR) [4, 77, 78]. Whether IDH mutation directly sensitizes gliomas to alkylating drugs (TMZ) and IR is an extremely important question that is being answered. As mentioned previously, IDH mutation triggers global hypermethylation leading to promoter methylation and silencing of O6-methylguanine-DNA methyltransferase (MGMT), a DNA repair protein involved in repairing alkylated DNA such as those produced by TMZ. IDH mutation thus sensitize glioma cells to TMZ, improving patient survival [54, 55, 58, 79]. Experimentally, in vitro overexpression of $IDH1^{R132H}$ and $IDH2^{R172K}$ has been shown to promote apoptosis of cells treated with IR or TMZ while IDH1 wild type overexpression appears protective [80, 81]. Moreover, IDH1 is the main producer of NADPH, while IDH1 mutation results in a strong reduction of this production. Thus, IDH1 mutation limit the reduction of the ROS generated by radio-chemotherapies therefore sensitizing glioma cells to current standard of care and prolonging survival (Fig. 3) [82]. Noteworthy, in AML, Glucose-6-phosphate Dehydrogenase (G6PDH) is the predominant NADPH generator, not IDH. Furthermore, AML treatment does not typically include IR, which may help explain why IDH mutations do not predict enhanced treatment sensitivity and longer survival in AML.

Of note, these pro-survival effects of IDH mutation may highlight therapeutic opportunities for IDH wild type gliomas. Although a deeper understanding of IDH mutation is still needed, it may be possible to recreate certain pro-survival effect conferred by IDH mutations. For example, inhibitors of NADPH production may induce some of these pro-survival effects in IDH wild type gliomas via sensitization to current standard of care.

5 Therapeutic Opportunities

Considerable effort has been recently directed toward the synthesis of IDH mutant selective inhibitors to disrupt the production of 2HG [83, 84]. To date, two selective IDH1 and IDH2 mutant inhibitors have been reported and well documented. AGI-6780 reduces 2HG production in a dose-dependent manner, inducing the differentiation of $IDH2^{R140Q}$ primary AML cells [85]. A second compound, AGI-5198, also decreases 2HG levels and induces differentiation of an $IDH1^{R132H}$ oligodendroglioma

cell line [86]. Interestingly, induction of differentiation by both inhibitors is associated in a dose-dependent manner with loss of the repressive histone marks, H3K9me3 and H3K27me3. However, AGI-5198 did not reverse DNA hypermethylation associated with IDH mutation. These studies demonstrate that IDH mutant enzymes are targetable by small molecules with positive effects on proliferation and differentiation. These inhibitors are able to reverse certain epigenetic marks triggered by IDH mutation and promote differentiation. Interestingly, the efficacy of these inhibitors may not be limited to their effects on histone epigenetic marks as illustrated by some interesting results showing that AGI-5198 doses that were insufficient to reverse histone methylation still suppressed tumor growth. The strategy aiming to inhibit IDH mutant enzymes thus holds promising therapeutic opportunities for the treatment of IDH mutant gliomas and other IDH mutant cancers.

Based on the idea that reversing IDH mutant-dependent epigenetic changes may be an interesting therapeutic option, several groups started to evaluate the efficacy of DNMT inhibitors in IDH mutant gliomas. Decitabine and 5-azacytidine are two very interesting DNMT inhibitors because they are well known FDA approved drugs that can pass the blood–brain barrier. *Turcan* et al. observed that decitabine treatment promotes changes in methylation markers leading to the re-expression of differentiation genes, and long lasting differentiation effects [87]. Another group examined the impact of 5-azacytidine and a reported reversion of the G-CIMP state concomitant with the upregulation of differentiation genes and decreased tumor growth in vivo [88].

Finally, since IDH mutations may directly confer vulnerabilities to oxidative stress, increasing this stress may represent an interesting therapeutic strategy for IDH mutant gliomas.

6 Conclusion

Although IDH mutations were first hypothesized to directly impact energy metabolism and promote aerobic glycolysis, which is the metabolic pathway typically favored by cancer cells to support rapid proliferation, the hypothesis that IDH mutations directly alter cancer cell metabolism remains unproven. IDH mutation is the earliest genetic event found in IDH mutant gliomas and is presumed to arise in the glioma cell of origin. This genetic alteration and the subsequent production of 2HG appear to drive tumorigenesis by blocking cancer cells in an immature proliferative state. According to the most recent studies, the main oncogenic properties of IDH mutations may be dependent on alterations of the epigenetic state via 2HG-dependent inhibition of histone demethylases and 5-methylcytosine hydroxylases. The occurrence of IDH mutations in a progenitor cell may hence maintain an undifferentiated, proliferative state through remodeling of epigenetic marks. These epigenetic alterations may be the initial event leading to tumorigenesis, which may be further promoted by secondary mutations and genetic alterations (Fig. 6). According to most recent studies, inhibition of IDH mutant enzymes appears

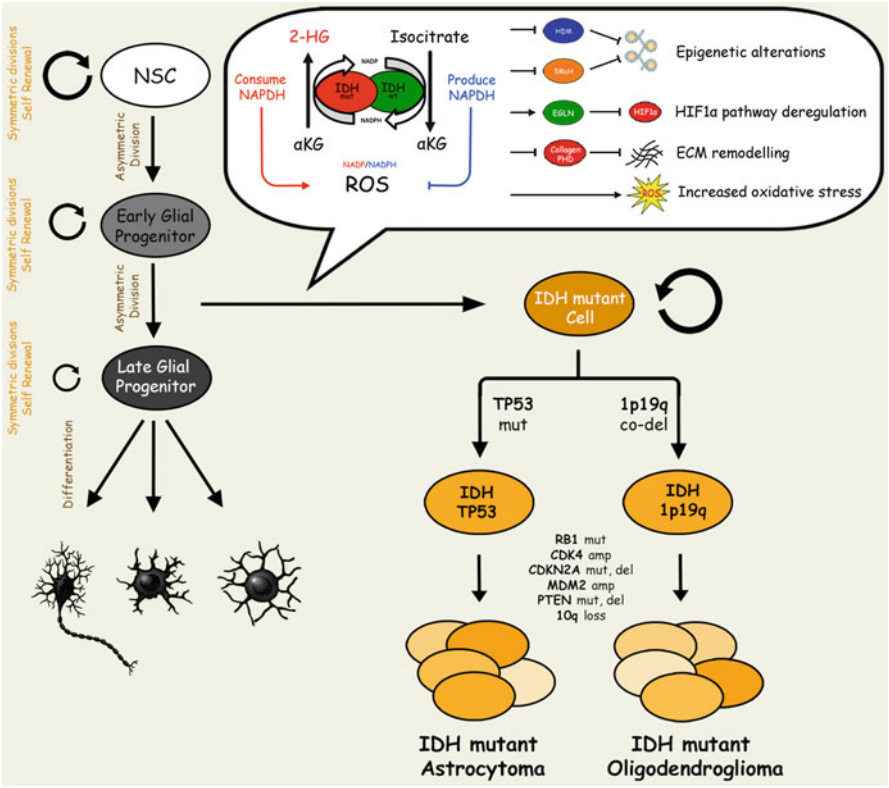


Fig. 6 IDH mutation: earliest genetic event in IDH mutant gliomas. IDH mutation is an early genetic event that may arise in an early progenitor cell. This genetic alteration and the subsequent production of 2HG appear to drive tumorigenesis by blocking cancer cells in an immature proliferative state. IDH mutation lead to epigenetic alterations, deregulation of the HIF1α pathway, ECM alterations, and increased oxidative stress, which may help maintain an undifferentiated, proliferative state. IDH mutation may be the initial event leading to tumorigenesis, which may be further promoted by secondary mutations and genetic alterations

effective in reversing some epigenetic marks and inducing differentiation as well as cytostatic effects. However, one has to keep in mind the intriguing possibility that even though IDH mutation is essential for tumor initiation it may be dispensable once a durable transformed state is acquired through the acquisition of additional mutations. Certain pro-survival effects of IDH mutations may even limit tumor progression beyond a certain stage and sensitize those tumors to current therapies, for example, IDH mutation via regulation of HIF1α and though epigenetic alterations deregulates expression of key genes essential for aerobic glycolysis preventing the glycolytic shift in IDH mutant gliomas, thus limiting the rapid growth typical of IDH wild type gliomas. Furthermore, IDH mutant enzymes alter the production of NADPH, limiting the reduction of Reactive Oxygen Species (ROS) and potentially leading to sensitization effects to therapy. Although such phenomena

may help explain the better prognosis and slow disease progression of IDH mutant gliomas, they also highlight unexpected consequences of IDH mutation that will influence our therapeutic strategies to control IDH mutant gliomas. One could indeed wonder if the axis of research aiming to inhibit IDH mutant enzyme is wise when it could, paradoxically, represent a step toward increased aggressiveness and resistance to current therapies. A better understanding of the roles and impact of IDH mutation in gliomas is thus paramount in order to develop therapeutic strategies that disrupt the oncogenic properties of IDH mutations while promoting properties that may contribute to the slower growth, enhanced sensitivity, and overall longer survival characteristic of IDH mutant gliomas.

References

1. Verhaak RG, Hoadley KA, Purdom E, et al. Integrated genomic analysis identifies clinically relevant subtypes of glioblastoma characterized by abnormalities in PDGFRA, IDH1, EGFR, and NF1. Cancer Cell. 2010;17:98–110.
2. Phillips HS, Kharbanda S, Chen R, et al. Molecular subclasses of high-grade glioma predict prognosis, delineate a pattern of disease progression, and resemble stages in neurogenesis. Cancer Cell. 2006;9:157–73.
3. Sanson M, Marie Y, Paris S, et al. Isocitrate dehydrogenase 1 codon 132 mutation is an important prognostic biomarker in gliomas. J Clin Oncol. 2009;27:4150–4.
4. Houillier C, Wang X, Kaloshi G, et al. IDH1 or IDH2 mutations predict longer survival and response to temozolomide in low-grade gliomas. Neurology. 2010;75:1560–6.
5. Bettegowda C, Agrawal N, Jiao Y, et al. Mutations in CIC and FUBP1 contribute to human oligodendroglioma. Science. 2011;333:1453–5.
6. Yip S, Butterfield YS, Morozova O, et al. Concurrent CIC mutations, IDH mutations, and 1p/19q loss distinguish oligodendrogliomas from other cancers. J Pathol. 2012;226:7–16.
7. Cairncross G, Jenkins R. Gliomas with 1p/19q codeletion: a.k.a. oligodendroglioma. Cancer J. 2008;14:352–7.
8. Parsons DW, Jones S, Zhang X, et al. An integrated genomic analysis of human glioblastoma multiforme. Science. 2008;321:1807–12.
9. Yan H, Parsons DW, Jin G, et al. IDH1 and IDH2 mutations in gliomas. N Engl J Med. 2009;360:765–73.
10. Watanabe T, Nobusawa S, Kleihues P, Ohgaki H. IDH1 mutations are early events in the development of astrocytomas and oligodendrogliomas. Am J Pathol. 2009;174:1149–53.
11. Ohgaki H, Kleihues P. The definition of primary and secondary glioblastoma. Clin Cancer Res. 2013;19:764–72.
12. Marcucci G, Maharry K, Wu YZ, et al. IDH1 and IDH2 gene mutations identify novel molecular subsets within de novo cytogenetically normal acute myeloid leukemia: a Cancer and Leukemia Group B study. J Clin Oncol. 2010;28:2348–55.
13. Mardis ER, Ding L, Dooling DJ, et al. Recurring mutations found by sequencing an acute myeloid leukemia genome. N Engl J Med. 2009;361:1058–66.
14. Paschka P, Schlenk RF, Gaidzik VI, et al. IDH1 and IDH2 mutations are frequent genetic alterations in acute myeloid leukemia and confer adverse prognosis in cytogenetically normal acute myeloid leukemia with NPM1 mutation without FLT3 internal tandem duplication. J Clin Oncol. 2010;28:3636–43.
15. Ward PS, Patel J, Wise DR, et al. The common feature of leukemia-associated IDH1 and IDH2 mutations is a neomorphic enzyme activity converting alpha-ketoglutarate to 2-hydroxyglutarate. Cancer Cell. 2010;17:225–34.

16. Abbas S, Lugthart S, Kavelaars F, et al. Acquired mutations in the genes encoding Idh1 and Idh2 both are recurrent aberrations in acute myeloid leukemia (Aml): prevalence and prognostic value. Haematol Hematol J. 2010;95:455.
17. Amary MF, Bacsi K, Maggiani F, et al. IDH1 and IDH2 mutations are frequent events in central chondrosarcoma and central and periosteal chondromas but not in other mesenchymal tumours. J Pathol. 2011;224:334–43.
18. Amary MF, Damato S, Halai D, et al. Ollier disease and Maffucci syndrome are caused by somatic mosaic mutations of IDH1 and IDH2. Nat Genet. 2011;43:1262–5.
19. Kosmider O, Gelsi-Boyer V, Slama L, et al. Mutations of IDH1 and IDH2 genes in early and accelerated phases of myelodysplastic syndromes and MDS/myeloproliferative neoplasms. Leukemia. 2010;24:1094–6.
20. Zhang Y, Wei H, Tang K, et al. Mutation analysis of isocitrate dehydrogenase in acute lymphoblastic leukemia. Genet Test Mol Biomarkers. 2012;16:991–5.
21. Cairns RA, Iqbal J, Lemonnier F, et al. IDH2 mutations are frequent in angioimmunoblastic T-cell lymphoma. Blood. 2012;119:1901–3.
22. Kipp BR, Voss JS, Kerr SE, et al. Isocitrate dehydrogenase 1 and 2 mutations in cholangiocarcinoma. Hum Pathol. 2012;43:1552–8.
23. Gaal J, Burnichon N, Korpershoek E, et al. Isocitrate dehydrogenase mutations are rare in pheochromocytomas and paragangliomas. J Clin Endocrinol Metab. 2010;95:1274–8.
24. Schaap FG, French PJ, Bovee JV. Mutations in the isocitrate dehydrogenase genes IDH1 and IDH2 in tumors. Adv Anat Pathol. 2013;20:32–8.
25. Kato Kaneko M, Liu X, Oki H, et al. Isocitrate dehydrogenase mutation is frequently observed in giant cell tumor of bone. Cancer Sci. 2014;105:744–8.
26. Zhao S, Lin Y, Xu W, et al. Glioma-derived mutations in IDH1 dominantly inhibit IDH1 catalytic activity and induce HIF-1alpha. Science. 2009;324:261–5.
27. Warburg O. On the origin of cancer cells. Science. 1956;123:309–14.
28. Warburg O. The metabolism of tumours: investigations from the Kaiser Wilhelm Institute for Biology, Berlin-Dahlem. London: Constable & Co. Ltd.; 1930.
29. Warburg O. On respiratory impairment in cancer cells. Science. 1956;124:269–70.
30. Baysal BE, Ferrell RE, Willett-Brozick JE, et al. Mutations in SDHD, a mitochondrial complex II gene, in hereditary paraganglioma. Science. 2000;287:848–51.
31. King A, Selak MA, Gottlieb E. Succinate dehydrogenase and fumarate hydratase: linking mitochondrial dysfunction and cancer. Oncogene. 2006;25:4675–82.
32. Tomlinson IP, Alam NA, Rowan AJ, et al. Germline mutations in FH predispose to dominantly inherited uterine fibroids, skin leiomyomata and papillary renal cell cancer. Nat Genet. 2002;30:406–10.
33. Dang L, White DW, Gross S, et al. Cancer-associated IDH1 mutations produce 2-hydroxyglutarate. Nature. 2009;462:739–44.
34. Ward PS, Cross JR, Lu C, et al. Identification of additional IDH mutations associated with oncometabolite R(-)-2-hydroxyglutarate production. Oncogene. 2012;31:2491–8.
35. Pichler MM, Bodner C, Fischer C, et al. Evaluation of mutations in the isocitrate dehydrogenase genes in therapy-related and secondary acute myeloid leukaemia identifies a patient with clonal evolution to IDH2 R172K homozygosity due to uniparental disomy. Br J Haematol. 2011;152:669–72.
36. Ward PS, Lu C, Cross JR, et al. The potential for isocitrate dehydrogenase mutations to produce 2-hydroxyglutarate depends on allele specificity and subcellular compartmentalization. J Biol Chem. 2013;288:3804–15.
37. Jin G, Reitman ZJ, Duncan CG, et al. Disruption of wild type IDH1 suppresses D-2-hydroxyglutarate production in IDH1-mutated gliomas. Cancer Res. 2013;73:496.
38. Ying W. NAD+/NADH and NADP+/NADPH in cellular functions and cell death: regulation and biological consequences. Antioxid Redox Signal. 2008;10:179–206.
39. Bleeker FE, Atai NA, Lamba S, et al. The prognostic IDH1(R132) mutation is associated with reduced NADP+-dependent IDH activity in glioblastoma. Acta Neuropathol. 2010;119:487–94.
40. Gilbert MR, Liu Y, Neltner J, et al. Autophagy and oxidative stress in gliomas with IDH1 mutations. Acta Neuropathol. 2014;127:221–33.

41. Latini A, Scussiato K, Rosa RB, et al. D-2-hydroxyglutaric acid induces oxidative stress in cerebral cortex of young rats. Eur J Neurosci. 2003;17:2017–22.
42. Loenarz C, Schofield CJ. Expanding chemical biology of 2-oxoglutarate oxygenases. Nat Chem Biol. 2008;4:152–6.
43. Seltzer MJ, Bennett BD, Joshi AD, et al. Inhibition of glutaminase preferentially slows growth of glioma cells with mutant IDH1. Cancer Res. 2010;70:8981–7.
44. Chaumeil MM, Larson PE, Yoshihara HA, et al. Non-invasive in vivo assessment of IDH1 mutational status in glioma. Nat Commun. 2013;4:2429.
45. Andronesi OC, Kim GS, Gerstner E, et al. Detection of 2-hydroxyglutarate in IDH-mutated glioma patients by in vivo spectral-editing and 2D correlation magnetic resonance spectroscopy. Sci Transl Med. 2012;4:116ra114.
46. Pope WB, Prins RM, Albert Thomas M, et al. Non-invasive detection of 2-hydroxyglutarate and other metabolites in IDH1 mutant glioma patients using magnetic resonance spectroscopy. J Neurooncol. 2012;107:197–205.
47. Struys EA. D-2-Hydroxyglutaric aciduria: unravelling the biochemical pathway and the genetic defect. J Inherit Metab Dis. 2006;29:21–9.
48. Moroni I, Bugiani M, D'Incerti L, et al. L-2-hydroxyglutaric aciduria and brain malignant tumors: a predisposing condition? Neurology. 2004;62:1882–4.
49. Latini A, da Silva CG, Ferreira GC, et al. Mitochondrial energy metabolism is markedly impaired by D-2-hydroxyglutaric acid in rat tissues. Mol Genet Metab. 2005;86:188–99.
50. Kolker S, Pawlak V, Ahlemeyer B, et al. NMDA receptor activation and respiratory chain complex V inhibition contribute to neurodegeneration in d-2-hydroxyglutaric aciduria. Eur J Neurosci. 2002;16:21–8.
51. Xu W, Yang H, Liu Y, et al. Oncometabolite 2-hydroxyglutarate is a competitive inhibitor of alpha-ketoglutarate-dependent dioxygenases. Cancer Cell. 2011;19:17–30.
52. Frezza C, Pollard PJ, Gottlieb E. Inborn and acquired metabolic defects in cancer. J Mol Med (Berl). 2011;89:213–20.
53. Kaelin WG. SDH5 mutations and familial paraganglioma: somewhere warburg is smiling. Cancer Cell. 2009;16:180–2.
54. Noushmehr H, Weisenberger DJ, Diefes K, et al. Identification of a CpG island methylator phenotype that defines a distinct subgroup of glioma. Cancer Cell. 2010;17:510–22.
55. Turcan S, Rohle D, Goenka A, et al. IDH1 mutation is sufficient to establish the glioma hypermethylator phenotype. Nature. 2012;483:479–83.
56. Figueroa ME, Abdel-Wahab O, Lu C, et al. Leukemic IDH1 and IDH2 mutations result in a hypermethylation phenotype, disrupt TET2 function, and impair hematopoietic differentiation. Cancer Cell. 2010;18:553–67.
57. He YF, Li BZ, Li Z, et al. Tet-mediated formation of 5-carboxylcytosine and its excision by TDG in mammalian DNA. Science. 2011;333:1303–7.
58. Lu C, Ward PS, Kapoor GS, et al. IDH mutation impairs histone demethylation and results in a block to cell differentiation. Nature. 2012;483:474–8.
59. Chowdhury R, Yeoh KK, Tian YM, et al. The oncometabolite 2-hydroxyglutarate inhibits histone lysine demethylases. EMBO Rep. 2011;12:463–9.
60. Chesnelong C, Chaumeil MM, Blough MD, et al. Lactate dehydrogenase A silencing in IDH mutant gliomas. Neuro Oncol. 2014;16:686–95.
61. Kaelin Jr WG. Cancer and altered metabolism: potential importance of hypoxia-inducible factor and 2-oxoglutarate-dependent dioxygenases. Cold Spring Harb Symp Quant Biol. 2011;76:335–45.
62. Koivunen P, Lee S, Duncan CG, et al. Transformation by the (R)-enantiomer of 2-hydroxyglutarate linked to EGLN activation. Nature. 2012;483:484–8.
63. Losman JA, Kaelin Jr WG. What a difference a hydroxyl makes: mutant IDH, (R)-2-hydroxyglutarate, and cancer. Genes Dev. 2013;27:836–52.
64. Losman JA, Looper RE, Koivunen P, et al. (R)-2-hydroxyglutarate is sufficient to promote leukemogenesis and its effects are reversible. Science. 2013;339:1621–5.
65. Luo W, Chang R, Zhong J, Pandey A, Semenza GL. Histone demethylase JMJD2C is a coactivator for hypoxia-inducible factor 1 that is required for breast cancer progression. Proc Natl Acad Sci U S A. 2012;109:E3367.

66. Navis AC, Niclou SP, Fack F, et al. Increased mitochondrial activity in a novel IDH1-R132H mutant human oligodendroglioma xenograft model: in situ detection of 2-HG and alpha-KG. Acta Neuropathologica Commun. 2013;1:18.
67. Grassian AR, Parker SJ, Davidson SM, et al. IDH1 mutations alter citric acid cycle metabolism and increase dependence on oxidative mitochondrial metabolism. Cancer Res. 2014;74:3317.
68. Takubo K, Goda N, Yamada W, et al. Regulation of the HIF-1alpha level is essential for hematopoietic stem cells. Cell Stem Cell. 2010;7:391–402.
69. Forristal CE, Winkler IG, Nowlan B, Barbier V, Walkinshaw G, Levesque JP. Pharmacologic stabilization of HIF-1alpha increases hematopoietic stem cell quiescence in vivo and accelerates blood recovery after severe irradiation. Blood. 2013;121:759–69.
70. Chow DC, Wenning LA, Miller WM, Papoutsakis ET. Modeling pO(2) distributions in the bone marrow hematopoietic compartment. I Krogh's model. Biophysical journal. 2001;81:675–84.
71. Jensen PO, Mortensen BT, Hodgkiss RJ, et al. Increased cellular hypoxia and reduced proliferation of both normal and leukaemic cells during progression of acute myeloid leukaemia in rats. Cell Prolif. 2000;33:381–95.
72. Liu W, Guo M, Xu YB, et al. Induction of tumor arrest and differentiation with prolonged survival by intermittent hypoxia in a mouse model of acute myeloid leukemia. Blood. 2006;107:698–707.
73. Song LP, Zhang J, Wu SF, et al. Hypoxia-inducible factor-1alpha-induced differentiation of myeloid leukemic cells is its transcriptional activity independent. Oncogene. 2008;27:519–27.
74. Piaskowski S, Bienkowski M, Stoczynska-Fidelus E, et al. Glioma cells showing IDH1 mutation cannot be propagated in standard cell culture conditions. Br J Cancer. 2011;104:968–70.
75. Luchman HA, Stechishin OD, Dang NH, et al. An in vivo patient-derived model of endogenous IDH1-mutant glioma. Neuro Oncol. 2012;14:184–91.
76. Luchman HA, Chesnelong C, Cairncross JG, Weiss S. Spontaneous loss of heterozygosity leading to homozygous R132H in a patient-derived IDH1 mutant cell line. Neuro Oncol. 2013;15:979.
77. Okita Y, Narita Y, Miyakita Y, et al. IDH1/2 mutation is a prognostic marker for survival and predicts response to chemotherapy for grade II gliomas concomitantly treated with radiation therapy. Int J Oncol. 2012;41:1325–36.
78. SongTao Q, Lei Y, Si G, et al. IDH mutations predict longer survival and response to temozolomide in secondary glioblastoma. Cancer Sci. 2012;103:269–73.
79. Hegi ME, Diserens AC, Gorlia T, et al. MGMT gene silencing and benefit from temozolomide in glioblastoma. N Engl J Med. 2005;352:997–1003.
80. Li S, Chou AP, Chen W, et al. Overexpression of isocitrate dehydrogenase mutant proteins renders glioma cells more sensitive to radiation. Neuro Oncol. 2013;15:57–68.
81. Wang JB, Dong DF, Wang MD, Gao K. IDH1 overexpression induced chemotherapy resistance and IDH1 mutation enhanced chemotherapy sensitivity in Glioma cells in vitro and in vivo. Asian Pac J Cancer Prev. 2014;15:427–32.
82. Baldewpersad Tewarie NM, Burgers IA, Dawood Y, et al. NADP+ -dependent IDH1 R132 mutation and its relevance for glioma patient survival. Med Hypotheses. 2013;80:728–31.
83. Zheng B, Yao Y, Liu Z, et al. Crystallographic investigation and selective inhibition of mutant isocitrate dehydrogenase. ACS Med Chem Lett. 2013;4:542–6.
84. Davis M, Pragani R, Popovici-Muller J, et al. ML309: a potent inhibitor of R132H mutant IDH1 capable of reducing 2-hydroxyglutarate production in U87 MG glioblastoma cells. Probe Reports from the NIH Molecular Libraries Program. Bethesda, MD: NIH; 2010.
85. Wang F, Travins J, DeLaBarre B, et al. Targeted inhibition of mutant IDH2 in leukemia cells induces cellular differentiation. Science. 2013;340:622–6.
86. Rohle D, Popovici-Muller J, Palaskas N, et al. An inhibitor of mutant IDH1 delays growth and promotes differentiation of glioma cells. Science. 2013;340:626–30.
87. Turcan S, Fabius AW, Borodovsky A, et al. Efficient induction of differentiation and growth inhibition in IDH1 mutant glioma cells by the DNMT Inhibitor Decitabine. Oncotarget. 2013;4:1729–36.
88. Borodovsky A, Salmasi V, Turcan S, et al. 5-azacytidine reduces methylation, promotes differentiation and induces tumor regression in a patient-derived IDH1 mutant glioma xenograft. Oncotarget. 2013;4:1737–47.

Utilization of Multigene Panels in Hereditary Cancer Predisposition Testing

Holly LaDuca, Tina Pesaran, Aaron M. Elliott, Virginia Speare, Jill S. Dolinsky, Chia-Ling Gau, and Elizabeth Chao

Abstract Hereditary cancer diagnostics is rapidly evolving with the increased availability and uptake of next-generation sequencing (NGS)-based multigene panels. Multigene panels offer several advantages such as time- and cost-effectiveness, and have been shown to be a useful diagnostic tool, particularly for cases suggestive of multiple different hereditary cancer conditions and for atypical phenotypes. However, there are many important considerations in the clinical use of multigene panels in hereditary cancer predisposition testing, from both clinic and laboratory perspectives. There are currently limited resources to guide clinicians in ordering multigene panels and managing patients with significant findings in lesser known genes. In addition, the development of clinical grade NGS-based panels is complex, and laboratories differ in various aspects of testing methodology. In this chapter, we review the various aspects of multigene panel workflow including target enrichment, NGS, bioinformatics, and interpretation of results. Results from our laboratory's experience with over 20,000 hereditary cancer panel cases are also summarized, with a focus on frequently mutated moderate penetrance genes, atypical phenotypes, and mosaic results.

H. LaDuca (✉) • T. Pesaran • V. Speare • J.S. Dolinsky • C.-L. Gau
Department of Clinical Diagnostics, Ambry Genetics,
15 Argonaut, Aliso Viejo, CA 92656, USA
e-mail: hladuca@ambrygen.com; tpesaran@ambrygen.com; vspeare@ambrygen.com; jdolinsky@ambrygen.com; cgau@ambrygen.com

A.M. Elliott
Department of Research and Development, Ambry Genetics,
15 Argonaut, Aliso Viejo, CA 92656, USA
e-mail: aelliott@ambrygen.com

E. Chao
Department of Clinical Diagnostics, Ambry Genetics,
15 Argonaut, Aliso Viejo, CA 92656, USA

Department of Pediatrics, Division of Genetics and Metabolism,
University of California, Irvine, Irvine, CA 92697, USA
e-mail: echao@ambrygen.com

W. Wu, H. Choudhry (eds.), *Next Generation Sequencing in Cancer Research, Volume 2*, DOI 10.1007/978-3-319-15811-2_26

1 Introduction

Next-generation sequencing (NGS)-based multigene panel testing is becoming increasingly utilized in hereditary cancer diagnostics, with a number of labs now offering this type of testing [1]. It is well known that NGS is more time-efficient and cost-effective than conventional sequencing methods [2, 3]. The ability to sequence multiple genes simultaneously gives the clinician information that would previously have required extensive time and resources to collect. As such, the introduction of NGS into the realm of hereditary cancer predisposition testing is transforming the way that we assess patients who are at risk for hereditary cancer. Several groups have validated NGS multigene cancer panels and have shown complete concordance with results from conventional testing methods [2–7]. Validations have included a variety of alteration types ranging from point mutations to genomic rearrangements. There are multiple important considerations in designing NGS panels for use in clinical practice to ensure the necessary clinical sensitivity and specificity, for example confirming results with Sanger sequencing to rule out false positives and follow-up analysis of no/low coverage regions with conventional testing methods. In this chapter, we review various aspects of multigene panel workflow in hereditary cancer diagnosis, with a focus on medical interpretation of variants, and present our clinical diagnostic laboratory's experience with multigene panel testing.

2 Workflow Considerations for Clinical Hereditary Cancer Panels

With the popularity of NGS increasing and the multitude of labs currently offering NGS-based multigene panel testing, it is important to have a good understanding of the various technologies and their limitations. NGS tests are more complex than Sanger sequencing-based testing methods, with very few labs employing similar workflows. Labs may differ in the target enrichment strategies, NGS platforms, and bioinformatics pipelines used for data analysis, each significantly impacting the accuracy and reliability of a test. Labs may choose a specific method and workflow based on sample volume, turn-around-time, cost or experience.

2.1 Target Enrichment

DNA target enrichment for NGS is performed either with polymerase chain reaction (PCR) primers or probes, each having unique advantages and disadvantages for variant detection. Primers, due to their small size (18–25 base pairs), have very high specificity enabling the selective enrichment of highly homologous and GC-rich regions. For diagnostic testing, this is extremely important as most targeted regions

can be covered 100 %, avoiding labor intensive Sanger sequencing to "fill-in" low coverage regions. In addition, the high on target specificity (typically 80–90 %) allows more samples to be multiplexed per sequencing run, bringing down the costs and turn-around-time of high volume testing [5]. The extremely high coverage achieved using primer based enrichment also enables the detection of low level mosaicism, which could be missed when using other approaches. Unfortunately, to perform in high-throughput, most primer based target enrichment technologies require expensive instrumentation which can limit those users without the sample volume or resources to support this type of investment. In addition, primer based target enrichment can produce false negatives when a variant is located underneath a primer binding site, affecting hybridization and resulting in allele drop-out [8, 9]. False negatives can be limited by designing primers away from common variants or known mutations and tiling amplicons to provide redundancy and sequence under other amplicon primer binding sites. These approaches were used in the design of our laboratory's cancer NGS tests, which have been shown to detect several mutations missed with previous testing methods [5].

By comparison, probe based target enrichment uses long oligonucleotides (100–120 base pairs) to hybridize and pull down regions of interest. Allele drop-out is reduced with this method as the hybridization properties of long probes allow tolerance to mismatched nucleotides. However, allelic ratios could be skewed as less efficient probe binding on one allele can cause bias. This method also results in less target enrichment derived false positives than primer-based methods due to the use of fragmented DNA and the removal of PCR duplicates during data analysis. Therefore, false positives are randomly distributed throughout the dataset with numerous unique reads used to make a call. Unlike primer based enrichment, where panels over ~100 genes becomes very cumbersome and expensive to process, there is no size limitation for probe based enrichment. For example, exome enrichment is performed using a probe library. Although the length of probes provides some advantages over primer based enrichment, it also results in several disadvantages. The main disadvantage is target specificity. Typically only ~50–60 % of reads are on target as the probes also pull down homologous regions in the genome [10]. This also makes capturing genes with highly homologous pseudogenes such as *PMS2* and *CHEK2* extremely challenging and unreliable, requiring labs to perform Sanger sequencing for these regions for full coverage. In addition, high GC-rich regions can be difficult to target with probe based enrichment. For these reasons, most probe based enrichment methods are only able to capture ~90–95 % of regions of interest, resulting in labs having to perform Sanger sequencing to fill-in the gaps.

2.2 Next-Generation Sequence Analysis

Currently the majority of diagnostic labs use Illumina sequencing instruments for testing due to the high accuracy and numerous throughput options which can accommodate any lab's testing volume (Illumina, San Diego, CA). The consistent use of the same

sequencing platform between labs removes some variability in the workflow as the base-calling error rates and sequencing related artifacts should be similar between assays interrogating the same genes. A small number of labs have implemented Ion Torrent or Ion Proton sequencing platforms as the instrument of choice for panel testing (Thermo Fisher Scientific, Waltham, MA). Unlike Illumina's sequencing by synthesis technology which utilizes reversible terminators to ensure only one nucleotide is incorporated at a time, Ion Torrent semiconductor technology uses flow based chemistry with the signal intensity proportional to the number of hydrogen ions released during nucleotide incorporation. As a result, similar to Roche 454 sequencing, there is a high error rate in homopolymer stretches making it difficult to accurately detect insertions and deletions in these regions (454 Life Sciences, Branford, CT) [11]. Importantly, homopolymers comprise a significant percentage of a gene's sequence and are hotspots for true mutations. For example, homopolymers (≥4 base pairs) account for 7.5 % of reportable *BRCA1* sequence and 11 % of *BRCA2* sequence. Therefore, to reliably call mutations in these regions Sanger sequencing would have to be performed.

2.3 Deletion/Duplication Analysis

Variability also exists in the methods used for gross deletion/duplication analysis. One approach involves normalized depth of coverage and split read analysis of NGS data to assess for deletion/duplications [12]. Software such as CNVseq can also be used to facilitate deletion/duplication analysis from NGS data [13]. This method is cost-effective since it does not require a separate testing method to be utilized; however, there are limitations depending on methods used for capture. Nonlinear PCR amplification can bias read depth coverage, and novel breakpoints can challenge alignment algorithms. Another option for concurrent deletion/duplication analysis of multiple genes is a targeted chromosomal microarray. Typically these rely on array-based comparative genomic hybridization as a methodology to compare copy number in patient DNA with control DNA following fluorescent labelling and hybridization. Exon-level coverage may not be achievable across all promoter and coding regions due to technical limitations of the DNA sequence such as high GC-content and/or sequence specificity and tandem repeats. Conventional deletion/duplication methods such as Multiplex Ligation-Dependent Probe Amplification (MLPA), or quantitative-PCR are not cost-effective for testing a large number of genes in parallel.

2.4 Bioinformatics

Finally, a significant amount of variability between different labs' panel NGS tests occurs in the bioinformatics and software employed for data analysis. There are a multitude of commercial software programs available to aid in the filtering and

analysis of sequencing data. However, many of these off the shelf programs are not flexible and cannot be tailored to analyze select genes or complex sequence regions on a panel independently. It is important for an analysis pipeline to be tailored for each gene on the panel based on its sequencing profile and account for the methods used in target enrichment and sequencing. For example, our approach has been to develop a custom pipeline where the primer sequences of tiled amplicons are trimmed off to avoid diluting out variants underneath the primer binding sites. Data from the analysis of our first 3,000 BRCAplus™ patients illustrated that we would have missed two pathogenic mutations in *BRCA1* and *BRCA2* if primer trimming was not incorporated into our pipeline [5]. Labs without extensive bioinformatics expertise and a tailored bioinformatics pipeline will compromise assay sensitivity and specificity resulting in missed mutations. Current algorithms for variant calling in clinical laboratory settings typically aim to maximize variant detection sensitivity. Because this may be achieved at the expense of false positive results, accuracy of results reporting may depend on Sanger sequencing confirmation of reported variants for some time as chemistries and bioinformatics algorithms continue to improve.

3 Medical Interpretation

A key component of any molecular diagnostic testing is accurate interpretation of detected variants. Various organizations such as the American College of Medical Genetics and Genomics (ACMG) and the International Agency for Research and Cancer (IARC) have developed guidelines for the interpretation and reporting of sequence variants to help standardize the interpretation and presentation of genetic testing results [14, 15]. While these guidelines provide a basic framework for assessment of variants that can be used for a wide range of autosomal dominant genes, variant assessment should occur in a gene-specific context, with disease phenotype, inheritance pattern, mechanism, and prevalence being considered, along with protein structure and function. Examples of other gene-specific nuances that should be considered include additional tests such as tumor studies, variation in nonsense-mediated decay, and alternate splicing. Several groups such as the International Society for Gastrointestinal Hereditary Tumours (InSiGHT) and IARC have developed more specific guidelines for mismatch repair genes (*MLH1*, *MSH2*, *MSH6*, and *PMS2*) and *BRCA1/2*, respectively [16, 17].

A common theme in published variant assessment guidelines is the reliance on multiple lines of evidence when interpreting the significance of sequence variation including published functional and splicing studies, case reports and case–control studies, information from locus-specific and population frequency databases, co-segregation, co-occurrence, and results from in silico prediction models. When reviewing any literature pertaining to variant assessment, the methods and study design should be carefully vetted for the strength and significance of the results.

3.1 Functional Studies

Functional studies aim to evaluate the effect of an alteration in vitro. They can be used as a strong line of evidence when results from multiple independent researchers are concordant. However, the study design should be carefully evaluated on a gene-by-gene basis. A review of the available functional literature and correlation with in vivo pathogenicity is highly advised. For example, the model organism can have a major impact on how well the results correlate with disease in vivo. Thompson et al. compared the results of in vitro studies in yeast and mammalian cells for mismatch repair genes [17]. For variants considered benign (class 1) based on multifactorial analysis or having a general population frequency >1 %, they found discordant results in 8/19 (42 %) yeast based assays and in 1/18 (5.5 %) mammalian cell assays. It should also be noted that normal functional and/or expression studies do not always correlate with pathogenicity as some mutations might not affect protein function but could affect other factors such as stability, expression, cellular localization, or binding.

3.2 Case–Control Studies

Case–control studies can be used to estimate the associated risk of a variant with disease. In these studies the allele or genotype frequency of a variant is compared between affected and unaffected populations or control cohorts. These controls generally consist of healthy individuals and are ideally matched for age, sex and ethnicity. Case–control data can be particularly helpful for classification of variants in high penetrance genes that are not rare (i.e., >0.1 %), as statistically significant odds ratios can then be defined in reasonably sized cohorts. However, rare variants and variants in moderate penetrance genes require much larger data sets to reach statistical significance. For rare variants the number of controls necessary to achieve the power of 90 % with the significance level of 0.05 would be generally greater than 15,000 if the allele frequency is ~0.1 % and greater than 30,000 if the allele frequency is ~0.05 % [18].

3.3 Phenotype Data

When evaluated in the context of penetrance and variability, phenotype data can be a powerful piece of evidence for classifying variants in genes associated with rare diseases (affecting < 1/2,000) and well-defined clinical diagnostic criteria [19]. Phenotype data is available from multiple sources such as published literature, online databases, and internal laboratory data. Clinical history information provided by clinicians on test requisition forms may not be complete; therefore, this information should be confirmed with clinicians if phenotype data is being used in variant classification.

Careful consideration should also be given to selection biases potentially influencing phenotype data. For example, phenotype data for patients tested for single genes, or a small panel of genes is typically biased, as genetic testing in this scenario was prompted based on the patient meeting clinical criteria for the disease. As such, use of phenotype data for variant classification in this context has limitations. History-weighing algorithms have been established for the use of such data; however, these analyses require very large datasets and are limited to high penetrance genes and alleles [20]. With NGS multigene panels for hereditary cancer syndromes, phenotype data is less biased. For example, if an alteration in a gene with a well-defined clinical presentation is observed in a large number of individuals without that phenotype (e.g., absence of diffuse gastric cancer or lobular breast cancer in carriers of *CDH1* variants), the likelihood of pathogenicity is decreased, although association of the variant with a "non-classic" phenotype cannot be ruled out. Statistical analysis of patterns of co-segregation with disease or family history can be performed and provide an extra line of evidence; the power of such analysis, however, depends on the accuracy of the penetrance estimation and may require sampling of additional individuals in the pedigree.

3.4 Population Frequency and Prevalence in Healthy Controls

In general, detection of a variant in control populations argues against its role in genetic disease. However, variants in less penetrant disease related genes may be found in both control and experimental populations. Online databases such as the Single Nucleotide Polymorphism Database (dbSNP), 1000 Genomes, and the Exome Sequencing Project (ESP) provide data on the frequency of a variant in general/control populations [21–24]. When drawing conclusions based on allele frequency the size of the cohort should be considered and statistical significance established. Although this was previously somewhat difficult to achieve and required pooling of cohorts, databases such as the ESP provide large general population frequencies for European Americans and African Americans. While the 1000 Genomes database genotyped a smaller cohort, it provides allele frequencies for a wider range of ethnicities. Statistical methods can be used to define thresholds for significance within smaller cohorts.

In our laboratory variants with a frequency >1 % in a large general population cohort such as ESP or a large negative control group are considered benign polymorphisms as they are too frequent to cause disease compared with the prevalence of hereditary cancer syndromes; however, this threshold varies between laboratories. Likewise, variants with allele frequency > 1 % at statistical significance in subpopulations can be considered benign polymorphisms, providing that the lower 95 % confidence interval is also >1 %. Rarely there are pathogenic founder mutations found in >1 % within an ethnic group, such as the *BRCA1* c.68_69delAG (also known as 187delAG) and *BRCA2* c.5946delT (also known as 6174delT); however, these are usually well-characterized [25, 26].

3.5 Co-segregation Analysis

Segregation analysis measures the degree to which a variant segregates with disease in one or more families. Typically segregation analysis relies on determining the likelihood a given variant is causing disease within large families with multiple affected and unaffected family members as measured by a logarithm of odds (LOD) score (base 10) [27]. Additional approaches to segregation analysis have been published that allow for analysis of smaller families, which can be used to determine odds of causality and can be combined from small families sharing the same variant [28]. Segregation analysis can be confounded by phenocopies within the family, particularly when investigating common disease. Analysis of a large number of individuals in a single family may be required to accurately associate the variant with hereditary disease. For rare, well-characterized hereditary cancer syndromes with classic phenotypes (e.g., Li–Fraumeni syndrome), fewer observations are needed to reach statistical significance as phenocopies are rare. Caution should be used when interpreting co-segregation data, as the possibility may exist that the alteration in question is in linkage disequilibrium with an unidentified causal mutation.

3.6 Co-occurrence

Mutation co-occurrence is the observation of a variant in conjunction with a known pathogenic mutation either in the same gene (*in cis* or *in trans*) or in another gene that at least in part explains the clinical phenotypes in the family. For some genes such as *BRCA1*, with few exceptions, homozygous and compound heterozygous loss-of-function mutations are embryonic lethal [29, 30]. For other genes, such as the mismatch repair genes and *BRCA2*, homozygous and compound heterozygous loss-of-function mutations lead to a severe phenotype such as constitutional mismatch repair deficiency syndrome and Fanconi Anemia, respectively [31–34]. For these genes, if a variant of unknown significance is confirmed to be *in trans* with a known deleterious mutation in the same gene, the likelihood that the variant is pathogenic is significantly reduced in the absence of a severe phenotype. In addition, variants found in the presence of known pathogenic mutations in other genes of the same pathway, or other genes that clearly explain the clinical phenotypes found in that family can also be used as evidence against pathogenicity. Cases have been reported, however, of families and individuals with mutations in more than one gene with or without overlapping phenotypes [35]. Therefore, co-occurrence with a mutation in another gene should be used cautiously and requires larger empirical data sets. Co-occurrence in some moderate penetrance genes such as *CHEK2* is not uncommon. While biallelic mutation carriers tend to have more severe phenotypes, co-occurrence in this setting is not as informative as evidence against pathogenicity [36, 37].

3.7 Evolutionary Conservation

Evolutionary conservation describes how well a nucleotide or amino acid position has been preserved throughout evolution in various species. In general, the degree of conservation can reflect functional significance; where highly conserved amino acid positions in highly conserved domains are more likely to have functional significance and less conserved amino acid positions are less likely to be functionally significant [16, 38]. In addition, observation of a specific amino acid change as the reference allele in many species suggests that the change is tolerated and therefore less likely to be significant. The depth of conservation should also be considered as some genes may not be as relevant in lower species and therefore not under strict evolutionary constraints. In this example, looking too deeply might underestimate the significance of conservation at that position. Interpretation of conservation data should be gene and context specific.

Multiple computational algorithms such as Sorting Intolerant from Tolerant (SIFT) [39], Polymorphism Phenotyping (PolyPhen) [40] and Combined Annotation Dependent Depletion (CADD) [41] as well as gene specific algorithms such as Align-GVGD (A-GVGD) [42, 43] and MAPP-MMR [44] have been developed to evaluate the evolutionary and functional significance of amino acid changes. These algorithms yield varying degrees of false positive rates and should not be used alone to classify variants. However, once their limitations are taken into consideration, they can be used to reflect evolutionary conservation as part of a classification scheme that relies on multiple lines of evidence.

4 Our Clinical Diagnostic Laboratory's Multigene Panel Experience

4.1 Methods

4.1.1 Study Subjects

Study subjects included 21,151 patients who had hereditary cancer multigene panel testing (BreastNext™, OvaNext™, ColoNext™, CancerNext™, and BRCAPlus™) through our clinical diagnostic laboratory. All patients were clinician-referred, with demographic information supplied by ordering clinicians on test requisition forms (TRFs) submitted at the time of testing. For the first 2,079 patients who underwent panel testing, clinical history information was also obtained from TRFs as well as any other clinical documentation (e.g., pedigrees, clinic notes) submitted by clinicians at the time of testing [35].

4.1.2 Gene Selection

Four panels were initially designed (BreastNext, OvaNext, ColoNext, and CancerNext) to target 14–22 genes associated with hereditary cancer susceptibility. Genes were selected if medical literature and database review supported a minimum of a twofold increased risk for the cancer type(s) targeted by the panel (breast cancer for BreastNext; breast, ovarian, and uterine cancers for OvaNext; colorectal cancer for ColoNext; breast, ovarian, colorectal, and uterine cancers for CancerNext). The well-known hereditary breast and ovarian cancer genes, *BRCA1* and *BRCA2*, were initially excluded from BreastNext, OvaNext, and CancerNext due to patents held by Myriad Genetics Laboratories, Inc.; however, these genes were added to the appropriate panels following the Supreme Court's ruling in June 2013 that naturally occurring DNA is not patent eligible merely because it has been isolated [45]. The ability to analyze *BRCA1* and *BRCA2* also enabled our design of a high risk breast cancer gene panel (BRCAplus) [5]. Continuous medical literature and database review has resulted in the addition of multiple genes to various panels. For example, *RAD51D* was added to the appropriate panels after multiple large studies confirmed its association with hereditary ovarian cancer [46–50]. Genes included in the aforementioned panels are summarized in Table 1, along with the number of patients who underwent sequencing of each gene as part of a multigene panel.

4.1.3 Target Enrichment and Next-Generation Sequencing Analysis

For all submitted blood and saliva samples, DNA was isolated using a QIAsymphony instrument (Qiagen, Valencia, CA) and then quantified using a UV spectrophotometer (NanoDrop, Thermo Scientific, Pittsburgh) or Infinite F200 (TECAN, San Jose, CA). Custom primers were designed to target regions of interest and also include sequences corresponding to the Illumina NGS adapters. Primers were heavily tiled in regions of interest to limit allele drop-out. Sequence enrichment was carried out by incorporating gDNA into emulsion microdroplets along with primer pairs followed by polymerase chain reaction (PCR) (RainDance Technologies, Billerica, MA). The enriched libraries were then applied to the solid surface flow cell for clonal amplification and sequencing using 150 bp paired-end conditions on the HiSeq2500 (Illumina, San Diego, CA). For all OvaNext, ColoNext, and CancerNext panels, *PMS2* sequence analysis was performed via Sanger sequencing due to pseudogene interference.

Initial data processing and base calling, including extraction of cluster intensities, was done using RTA 1.17.21.3 (Real Time Analysis, HiSeq Control Software version 2.0.10). Sequence quality filtering was executed with the Illumina CASAVA software (ver 1.8.2, Illumina, Hayward, CA). A custom bioinformatics pipeline utilized Novoalign V3.00.05 to align sequence data to the reference human genome (GRCh37) and GATK V2013.1-2.4.9 to generate variant and no/low coverage reports. During variant calling, primer sequences were trimmed off to avoid these sequences being included in the analysis and diluting out true sample sequence under primer sites. For filtering, a minimum quality threshold of Q20 was applied and no/low coverage

Table 1 Genes included on each multigene cancer panel

Gene (# of patients sequenced)	BRCAplus™	BreastNext™	OvaNext™	ColoNext™	CancerNext™
APC (4,748)				•	•
ATM (10,417)		•	•		•
BARD1 (10,417)		•	•		•
BRCA1 (16,430)	•	•	•		•
BRCA2 (16,430)	•	•	•		•
BRIP1 (10,417)		•	•		•
BMPR1A (4,748)				•	•
CDH1 (21,151)	•	•	•	•	•
CDK4 (1,108)					•
CDKN2A (1,108)					•
CHEK2 (12,433)		•	•	•	•
EPCAM (7,110)			•	•	•
MLH1 (7,110)			•	•	•
MRE11A (10,417)		•	•		•
MSH2 (7,110)			•	•	•
MSH6 (7,110)			•	•	•
MUTYH (12,433)		•	•	•	•
NBN (10,417)		•	•		•
NF1 (4,886)		•	•		•
PALB2 (10,417)		•	•		•
PMS2 (7,110)			•	•	•
PTEN (21,151)	•	•	•	•	•
RAD50 (10,417)		•	•		•
RAD51C (10,417)		•	•		•
RAD51D (4,886)		•	•		•
SMAD4 (4,748)				•	•
STK11 (21,151)	•	•	•	•	•
TP53 (21,151)	•	•	•	•	•

regions that are <50× coverage. Additional Sanger sequencing was performed for any no/low coverage regions (<50×). Variant calls other than known non-pathogenic alterations were verified by Sanger sequencing in sense and antisense directions prior to reporting. The complete multigene panel workflow is depicted in Fig. 1.

4.1.4 Deletion/Duplication Analysis

A custom, highly tiled chromosomal microarray (CMA) consisting of approximately 60,000 oligonucleotide probes was developed for gross deletion/duplication analysis using eArray software (Agilent Technologies, https://earray.chem.agilent.com/earray/) and was run concurrently with NGS analysis for all multigene panels.

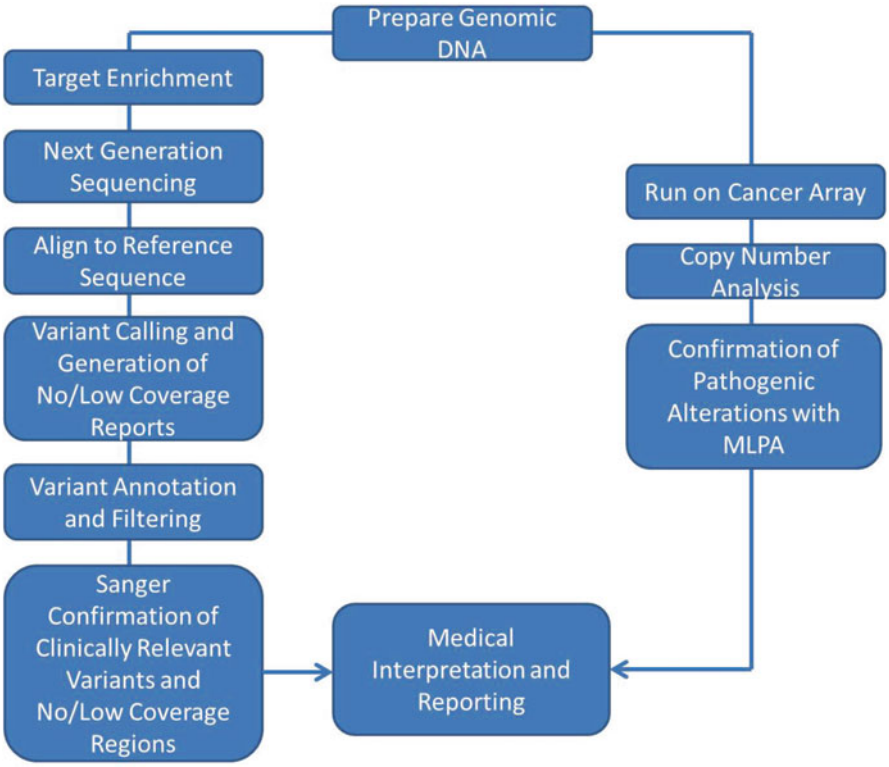


Fig. 1 Multigene panel workflow. After preparing genomic DNA, samples undergo concurrent NGS and gross deletion/duplication analysis. All no/low coverage regions on NGS are subsequently analyzed with Sanger sequencing and all clinically relevant variants are confirmed with conventional testing methods (Sanger sequencing or MLPA) prior to medical interpretation and reporting [5]

Probes were placed every 2.5 Kb of intronic sequence, with increased probe density in exons (average 13 probes per exon), flanking intronic sequences and promoter regions. Pathogenic gross deletions/duplications were confirmed using MLPA analysis (MRC-Holland). *PMS2* deletion/duplication analysis was carried out using MLPA due to pseudogene interference. Follow-up double stranded sequencing of the appropriate pseudogene exons was performed in the event of a deletion in exons 12–15 of *PMS2* [51].

4.1.5 Variant Assessment

All variants, with the exception of previously characterized benign alterations, underwent thorough assessment and review of available evidence by a team of highly trained scientists. Our proprietary five-tier variant classification protocol, based on the ACMG and IARC guidelines, was used to arrive at final variant classifications [14, 15, 35].

4.1.6 Results Reporting

All confirmed variants in coding exons ±5 base pairs into flanking introns and untranslated regions were reported for genes on the panel ordered, with the exception of benign alterations. Known mutations beyond ±5 base pairs were also reported. Four results categories were utilized on test reports: positive (one or more pathogenic mutations or likely pathogenic variants were detected), inconclusive (only variants of uncertain significance (VUS) were detected), negative (no variants or only benign or likely benign variants were detected), and carrier (monoallelic *MUTYH* pathogenic mutation or likely pathogenic variant carrier). For all reports containing pathogenic mutations, likely pathogenic variants, and variants of unknown significance, detailed alteration and gene information was included to support the reported classification and interpretation.

4.2 Results

This cohort consisted of 21,151 individuals who underwent BRCAplus ($n=8,718$), BreastNext ($n=5,323$), OvaNext ($n=2,362$), ColoNext ($n=2,016$), or CancerNext ($n=2,732$) testing through our clinical diagnostic laboratory. The majority of patients were reported to be Caucasian (70.0 %) and female (93.8 %), with an average age at testing of 51.9 years (range birth to 97 years) (Table 2).

A total of 1,691 total mutations and likely pathogenic variants were detected among 1,616 probands with positive results, including 28 biallelic *MUTYH* mutation carriers and 47 probands with two mutations. The majority of mutations were sequence mutations ($n=1,574$, 93.1 %), and the remainder were gross deletions/duplications/triplication ($n=117$, 6.9 %).

4.2.1 Factors Influencing Positive and Inconclusive Rates

Positive, inconclusive, and negative results rates for each panel are shown in Fig. 2. CancerNext yielded the highest percentage of positive results (11.4 %) and BRCAplus yielded the lowest percentage of positive results (4.8 %), which is reflective of the number of genes included on the panel (6 genes on BRCAplus compared to 28 genes on CancerNext). Generally speaking, the mutation detection rate increases with the number of genes included that are relevant to the referring diagnosis, and the inconclusive rate is directly related to the number of genes included on the panel and consequently the number of base pairs sequenced (Fig. 3). The amount of published literature (e.g., functional studies and large case–control studies) on any given gene may also impact the classification of variants as pathogenic, uncertain significance, or benign.

The clinical history of the patient being tested is another factor that influences mutation detection rate. Previous detailed analysis of clinical data from our first 2,079 cancer panel cases demonstrated varying detection by clinical history (Table 3).

Table 2 Patient demographics

	Overall *n* (%)
Total patients	21,151 (100)
Age	
≤30 years	1,074 (5.1)
31–40 years	3,339 (15.8)
41–50 years	5,591 (26.4)
51–64 years	7,004 (33.1)
≥65 years	4,143 (19.6)
Test performed	
BRCAplus	8,718 (41.2)
BreastNext	5,323 (25.2)
CancerNext	2,732 (12.9)
ColoNext	2,016 (9.5)
OvaNext	2,362 (11.2)
Ethnicity	
African American/Black	1,167 (5.5)
Ashkenazi Jewish	1,165 (5.5)
Asian	669 (3.2)
Caucasian	14,805 (70.0)
Hispanic	940 (4.4)
Middle Eastern	115 (0.5)
Multiple Ethnicities	679 (3.2)
Native American	24 (0.1)
Unknown/blank	1,535 (7.3)
Gender	
Female	19,836 (93.8)
Male	1,272 (6.0)

For example, ColoNext cases with ≥10 cumulative adenomatous polyps were almost twice as likely to test positive for a mutation than cases with <10 cumulative adenomatous polyps. The impact of clinical history on VUS rate is not as well explored.

4.2.2 Frequently Mutated Moderate-Penetrance Genes

The multigene panels included in our analysis contain genes that are known to confer both high and intermediate risk for cancer [35]. The intermediate risk genes (relative risk two to fourfold) were identified as candidates for hereditary predisposition based on an understanding of cellular pathways important to the integrity of the genome such as DNA repair (genes of Fanconi Anemia-BRCA pathway), signal transduction (e.g., *PTEN*), and cell cycle checkpoints (e.g., *CHEK2*) [52]. Variants in some of the genes may be found with high frequency in the population, making it difficult to study the association of these variants with disease, as large cohorts are

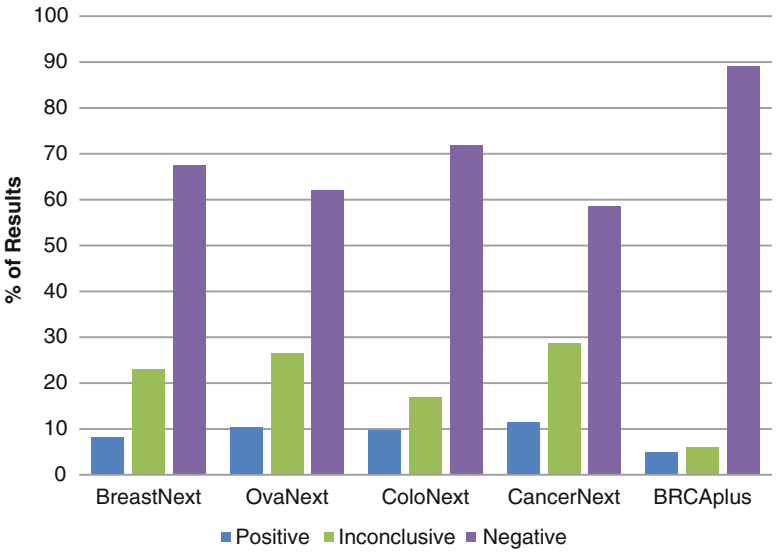


Fig. 2 Result rates by multigene panel. Positive, inconclusive, and negative result rates are shown for each multigene panel. Positive and inconclusive rates were highest for CancerNext, which includes the most genes and were lowest for BRCAplus, which includes the least number of genes. Individuals whose molecular findings only included a monoallelic *MUTYH* mutation were excluded from calculations

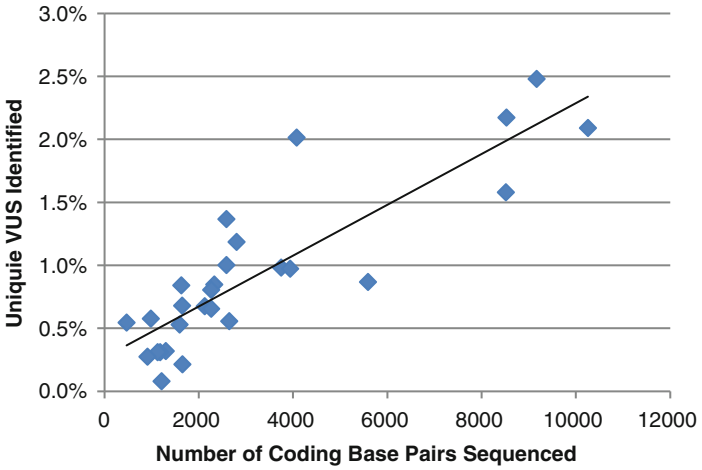


Fig. 3 Number of unique sequence VUS identified compared to number of coding base pairs sequenced. The percentage of unique VUS identified per patients sequenced is directly related to the number of coding base pairs sequenced, with VUS rates highest for larger genes and lowest for smaller genes. Each dot represents a gene on the multigene panels. *EPCAM* is the only gene not represented in this figure, as only deletion/duplication analysis is performed for this gene

Table 3 Positive and inconclusive rates by clinician-reported clinical history [35]

Characteristic (*n*)	% Positive results	% Inconclusive results
BreastNext (*n* = 874)		
High-risk breast/ovarian criteria (239)	10.9	20.9
Triple negative breast cancer (76)	5.3	30.3
Multiple breast cancer primaries (148)	8.8	23.6
Breast cancer diagnosed <35 years (136)	7.4	24.3
Breast cancer diagnosed <50 years (528)	9.3	19.9
OvaNext (*n* = 223)		
High-risk breast/ovarian criteria (37)	5.4	16.2
Breast cancer diagnosed <50 years (43)	11.6	23.3
Ovarian cancer at any age (111)	6.3	30.6
ColoNext (*n* = 557)		
Colorectal cancer diagnosed <50 years (168)	13.1	13.7
2–9 Cumulative adenomas (120)	7.5	20.8
10+ Cumulative adenomas (90)	14.4	12.2
CancerNext (*n* = 425)		
High-risk breast/ovarian criteria (47)	12.8	21.3
Breast cancer diagnosed <50 years (90)	10.0	25.6
Ovarian cancer at any age (32)	9.4	25.0
Colorectal cancer diagnosed <50 years (23)	8.7	21.7
Multiple primary tumors (160)	7.5	21.9

required to show statistical significance. We observed three intermediate risk genes that were mutated in ≥ 1 % of individuals tested: *CHEK2*, *ATM*, and *PALB2* (Fig. 4). The BreastNext panel yielded the highest mutation frequency for all three of these genes with frequencies of 2.7 %, 1.2 %, and 1.3 %, respectively (data not shown).

Our observed mutation frequencies for *ATM*, *CHEK2*, and *PALB2* are higher than in other published multigene panel cohorts (Table 4). The differences in detection rates among study populations may be explained by differing study cohorts, variable methods of variant classification or the possibility of false negatives due to inconsistencies in analysis for deletions and duplications. Walsh et al. did not detect any *ATM* mutations in their cohort of 360 ovarian cancer patients unselected for age or family history, which may reflect our current understanding that *ATM* mutations are not associated with ovarian cancer risk [53]. Castera et al. included only mutations inducing premature termination codons and considered missense mutations with an A-GVGD score > C45 and a MAF in ESP samples >0.01 as potentially deleterious [4]. Pathogenic mutations exist that have an A-GVGD score of <C45 and a MAF of >0.1 %, for example, the p.S428F mutation in *CHEK2* which has an A-GVGD score of C15 and is seen in 0.02 % of the population. Furthermore, gross deletion/duplication analysis was not performed for any genes other than *BRCA1/2* in their study.

In our study, the likelihood of identifying mutations in moderate penetrance genes exceeds that of identifying mutations in high penetrance genes. Currently, there are limited resources to guide clinicians in managing patients with mutations in moderate penetrance genes, leaving clinicians to extrapolate from published

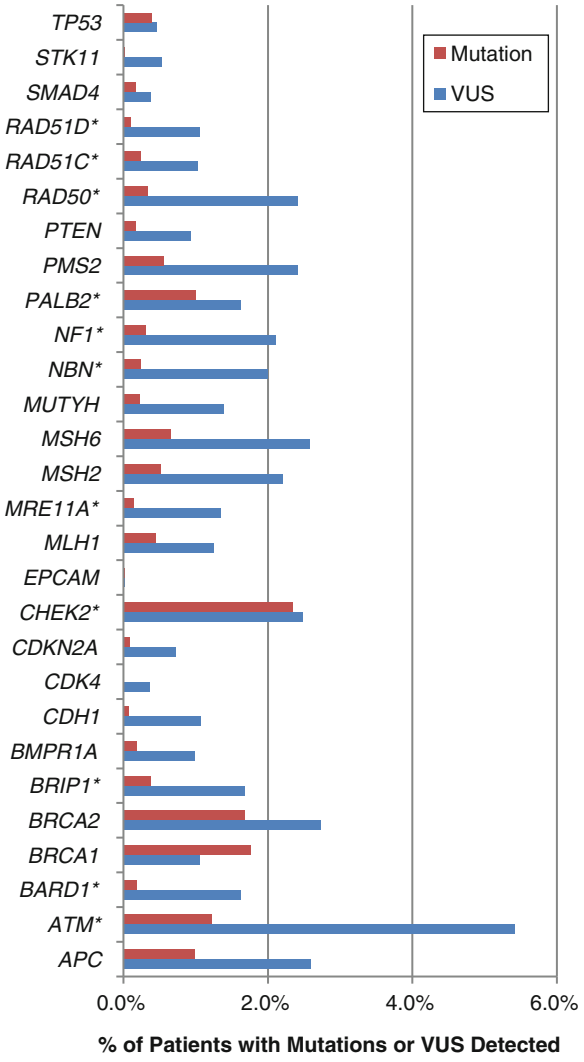


Fig. 4 Mutation and VUS rates by gene. The mutation and VUS rates were calculated based on the number of times each gene was sequenced in our cohort. Moderate penetrance cancer genes are indicated with a *. Aside from *BRCA1/2*, the most frequently mutated genes were *ATM, CHEK2, PALB2*, and *APC*. Of note, 21 of 47 (44.7 %) *APC* mutations reported were the moderate risk allele, p. I1307K, which confers a two-fold increased risk of colorectal cancer but is not causative of the classic or attenuated familial adenomatous polyposis phenotype [70]. Therefore, the rate for *APC* mutations causative of FAP/AFAP was 0.55 %

gene-specific risk estimates, guidelines for other genes conferring similar risks, and patient clinical history. As additional information emerges in published literature, one can anticipate the development of additional guidelines for management of families with mutations in moderate penetrance genes, including but not limited to specific guidelines for surveillance and recommendations for or against risk-reducing surgeries. Genetic testing is already being used to guide targeted cancer therapy in the case of the *BRCA1/2* genes [54]. It has been postulated that carriers of mutations in other genes in the Fanconi Anemia/Homologous recombination DNA repair pathway may also be responsive to similar targeted therapies, i.e., Poly (ADP-ribose) polymerase (PARP) inhibitors [55]. Additional support for this hypothesis might be expected to emerge over time as clinical trials are ongoing [56].

Table 4 Frequency of *ATM*, *CHEK2*, and *PALB2* mutations in multigene panel cohorts

Author	Study population	Carrier frequency[a]		
		ATM	*CHEK2*	*PALB2*
Kurian et al. [7]	141 women referred for *BRCA1/2* testing, with negative results	1.42 %	ND	ND
Castera et al. [4]	708 consecutive patients with clinical histories suspicious for hereditary breast and/or ovarian cancer	0.71 %	0.71 %	0.99 %
Walsh et al. [53]	360 ovarian cancer, fallopian tube, and primary peritoneal cancer patients unselected for age or family history	ND	1.39 %	0.56 %
Tung et al. [71]	1,781 patients referred for *BRCA1/2* testing and 377 *BRCA1/2* negative breast cancer patients	0.60 %	1.58 %	0.60 %
LaDuca et al. (this study)	21,151 patients clinician-referred for hereditary cancer panel testing[b]	1.77 %	1.64 %	1.31 %

[a]ND: not detected
[b]*CHEK2* was sequenced for 12,433 patients and *ATM* and *PALB2* were sequenced for 10,417 patients included in the study

4.2.3 Atypical Phenotypes

We previously analyzed clinical histories of 46 ColoNext probands with mutations in genes with well-established diagnostic criteria and treatment guidelines (*CDH1*, *PTEN*, *SMAD4*, *STK11*, *TP53*, *APC*, *MLH1*, *MSH2*, *MSH6*, *PMS2*, *EPCAM*, and biallelic *MUTYH*) to determine whether they met diagnostic/testing criteria for the corresponding hereditary cancer syndrome [35]. Thirty percent of these probands did not meet criteria for the corresponding syndrome, and in several of these cases, the clinical history was suggestive of multiple different hereditary cancer syndromes. Subsequent analysis of clinical histories for other gene mutation carriers such as *TP53*, *PTEN*, and *CDH1* has also revealed a number of patients not meeting diagnostic/testing criteria for the respective syndrome (unpublished data).

Our results demonstrate that current syndrome-specific testing guidelines are missing a number of patients with hereditary cancer susceptibility and do not address the significant overlap between phenotypes. Historically, research on hereditary cancer syndromes was limited to clinical observation of patients with strong similar phenotypes. Cancer risks and testing guidelines were subsequently developed based on highly penetrant families. Our results also indicate that further research in prospective cohorts is needed to better define phenotypes and penetrance for cancer susceptibility genes. These findings also have implications for the use of phenotype data in variant assessment for well-characterized high penetrance genes. While the identification of a rare variant in a patient with classic disease is a factor in support of pathogenicity, the lack of a classic disease phenotype no longer carries as much weight in support of nonpathogenicity.

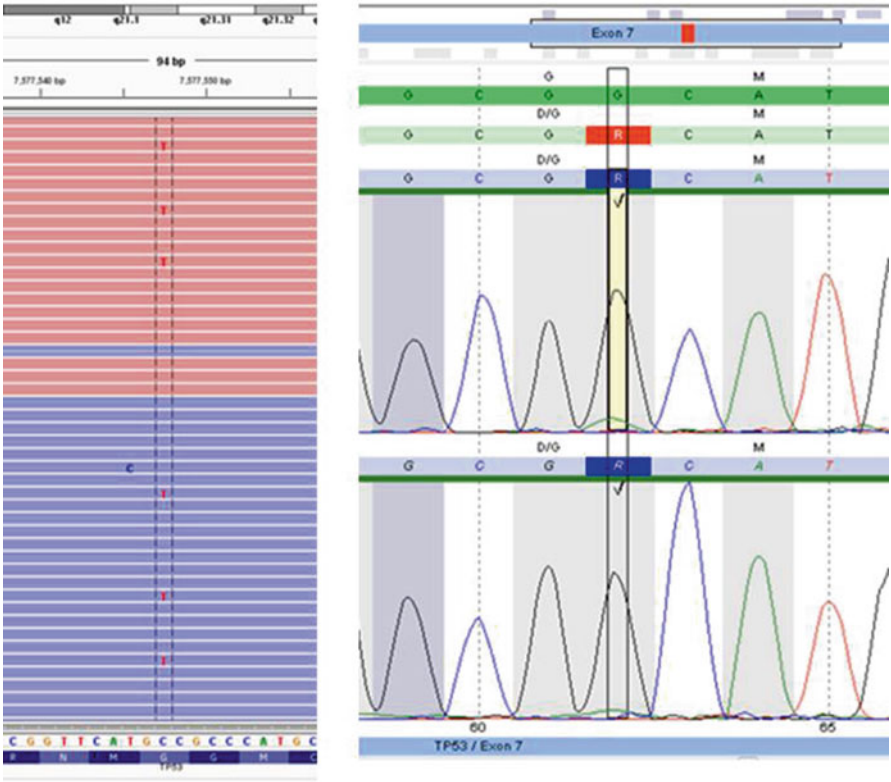


Fig. 5 Mosaic *TP53* results. The *TP53* p.G245D pathogenic mutation was detected on NGS with an allele ratio of 10.7 % (*left*) and subsequently confirmed via Sanger sequencing (*right*)

4.2.4 Mosaic Results

NGS is more sensitive in detection of mosaicism than conventional testing methods. There have been multiple reports of NGS identifying low level mosaicism in hereditary cancer genes [53, 57–60]. In this study, 20 alterations among 19 patients were detected in a significant portion of DNA from the provided sample but at a lower frequency than expected for heterozygous carriers by both NGS and Sanger sequencing (Fig. 5; all data not shown). For these cases, an alternate pair of primers was used in Sanger confirmation to rule out low heterozygous ratio due to allele bias. The majority of these alterations were in *TP53* ($n=13$) and the remaining calls were in *ATM* ($n=1$), *BRCA2* (1), *CHEK2* ($n=4$), and *NF1* ($n=1$). Explanations for lower-than-expected heterozygous ratio include the presence of mosaicism.

Counseling can be challenging in this situation, as there is no easy way to determine which tissues are affected, including gonadal tissue, and so the risks to offspring cannot be clearly delineated. It is also possible that different tissues within

the body may be affected to different degrees and so cancer risks may vary depending on the tissue distribution of the mutation. Another explanation is that the detected mutation is somatic in origin and detection reflects the presence of malignant tissue within the provided patient sample. When the sample is blood or saliva, this is most likely due to the presence of hematologic disease. For this reason, saliva and peripheral blood are not acceptable sample types for patients with active or recent hematologic disease, with cultured fibroblasts or fresh, frozen tissue preferred. Experience suggests that for some patients, however, positive NGS results may be the first indication of an undiagnosed hematologic disease. In such cases our laboratory advises that clinicians correlate results where heterozygous variant read ratios are lower than expected with the patient's clinical history and perform any follow-up analyses that may help to clarify the origin of the reported result such as offering testing to family members, evaluating for potential hematologic disease, and performing testing on additional tissue types such as cultured fibroblasts.

4.2.5 Testing Guidelines for NGS Panels

The multigene panels offered by our laboratory and others, include genes that have not yet been individually evaluated for clinical utility of testing. It is, however, well-established that a molecular diagnosis in individuals diagnosed with cancer has potential implications for future surveillance, risk reduction and potentially treatment [61–63]. Available guidelines for the routine use of genetic testing in oncology include those of the National Comprehensive Cancer Network (NCCN) and American Society of Clinical Oncology (ASCO) [54, 64, 65]. The components of informed consent for genetic testing are also readily available [65, 66]. The NCCN addressed the use of multigene testing in the 2014 updates to their Genetic/Familial High-Risk Assessment: Breast and Ovarian clinical practice guideline, but specific procedures were not recommended.

In our initial study of 2,079 cases of multigene testing, the majority of subjects had previously undergone some genetic testing which proved to be uninformative [35]. As multigene testing has been incorporated into the clinical care of a larger number of patients, the approach is now used as a first line genetic test challenging the convention of sequential gene testing. Individual groups have published their experience and approaches to the use of multigene testing [67, 68]. However, systematic investigation of the personal and societal implications of choosing a multigene approach to genetic testing is needed and studies are underway.

5 Summary and Future Directions

Historically, the identification of patients and families at risk for hereditary cancer predisposition was based on syndrome identification through careful attention to family history and clinical presentation of disease. Limitations to this approach

included the lack of information about the family due to limited communication of diagnoses among family members, small family sizes and variable penetrance of disease-causing genes. Gene discovery relied on the participation of large families in research and the detailed description of disease over three to four generations. With the introduction of a rapid and cost-effective method such as NGS, the approach to gene discovery and identification of at-risk families now may be guided by molecular data (the genotype) as often as it is guided by the phenotypic presentation. The success of this approach is dependent on the reliability of molecular data assemblage and interpretation.

The clinical application of NGS and a multigene approach to testing presents challenges such as the assessment of sequence variants for pathogenicity, the discovery of atypical phenotypes and germ-line mosaicism and a lack of data defining the penetrance of lesser known genes. Clinical recommendations and guidelines are yet to be published regarding the use of multigene testing. More research investigating the utility of the multigene approach is needed and is addressed by working groups such as the *E*vidence-based *N*etwork for the *I*nterpretation of *G*ermline *M*utant *A*lleles (ENIGMA) [69]. Our presentation here of a NGS workflow and strategies for the interpretation of NGS data adds to the growing body of literature needed to translate laboratory innovation to clinical care.

References

1. Pagon, R. GeneTests. 2014. Accessed on September 2, 2014, from: http://www.genetests.org/
2. Pritchard CC, et al. ColoSeq provides comprehensive lynch and polyposis syndrome mutational analysis using massively parallel sequencing. J Mol Diagn. 2012;14(4):357–66.
3. Walsh T, et al. Detection of inherited mutations for breast and ovarian cancer using genomic capture and massively parallel sequencing. Proc Natl Acad Sci U S A. 2010;107(28):12629–33.
4. Castera L, et al. Next-generation sequencing for the diagnosis of hereditary breast and ovarian cancer using genomic capture targeting multiple candidate genes. Eur J Hum Genet. 2014;22:1305.
5. Chong HK, et al. The validation and clinical implementation of BRCAplus: a comprehensive high-risk breast cancer diagnostic assay. PLoS One. 2014;9(5):e97408.
6. Morgan JE, et al. Genetic diagnosis of familial breast cancer using clonal sequencing. Hum Mutat. 2010;31(4):484–91.
7. Kurian AW, et al. Clinical evaluation of a multiple-gene sequencing panel for hereditary cancer risk assessment. J Clin Oncol. 2014;32(19):2001–9.
8. Lam CW, Mak CM. Allele dropout in PCR-based diagnosis of Wilson disease: mechanisms and solutions. Clin Chem. 2006;52(3):517–20.
9. Landsverk ML, et al. Diagnostic approaches to apparent homozygosity. Genet Med. 2012;14(10):877–82.
10. Sulonen AM, et al. Comparison of solution-based exome capture methods for next generation sequencing. Genome Biol. 2011;12(9):R94.
11. Elliott AM, et al. Rapid detection of the ACMG/ACOG-recommended 23 CFTR disease-causing mutations using ion torrent semiconductor sequencing. J Biomol Tech. 2012;23(1):24–30.
12. Nord AS, et al. Accurate and exact CNV identification from targeted high-throughput sequence data. BMC Genomics. 2011;12:184.

13. Xie C, Tammi MT. CNV-seq, a new method to detect copy number variation using high-throughput sequencing. BMC Bioinform. 2009;10:80.
14. Plon SE, et al. Sequence variant classification and reporting: recommendations for improving the interpretation of cancer susceptibility genetic test results. Hum Mutat. 2008;29(11): 1282–91.
15. Richards CS, et al. ACMG recommendations for standards for interpretation and reporting of sequence variations: revisions 2007. Genet Med. 2008;10(4):294–300.
16. Tavtigian SV, et al. Assessing pathogenicity: overview of results from the IARC Unclassified Genetic Variants Working Group. Hum Mutat. 2008;29(11):1261–4.
17. Thompson BA, et al. Application of a 5-tiered scheme for standardized classification of 2,360 unique mismatch repair gene variants in the InSiGHT locus-specific database. Nat Genet. 2014;46(2):107–15.
18. Freidlin B, et al. Trend tests for case-control studies of genetic markers: power, sample size and robustness. Hum Hered. 2002;53(3):146–52.
19. Hennekam RC. Care for patients with ultra-rare disorders. Eur J Med Genet. 2011;54(3): 220–4.
20. Eggington JM, et al. A comprehensive laboratory-based program for classification of variants of uncertain significance in hereditary cancer genes. Clin Genet. 2014;86(3):229–37.
21. Exome Variant Server, NHLBI GO Exome Sequencing Project (ESP). 2013; Seattle, WA.
22. Abecasis GR, et al. An integrated map of genetic variation from 1,092 human genomes. Nature. 2012;491(7422):56–65.
23. Consortium IH. The International HapMap Project. Nature. 2003;426(6968):789–96.
24. Sherry ST, et al. dbSNP: the NCBI database of genetic variation. Nucleic Acids Res. 2001;29(1):308–11.
25. Oddoux C, et al. The carrier frequency of the BRCA2 6174delT mutation among Ashkenazi Jewish individuals is approximately 1 %. Nat Genet. 1996;14(2):188–90.
26. Struewing JP, et al. The carrier frequency of the BRCA1 185delAG mutation is approximately 1 percent in Ashkenazi Jewish individuals. Nat Genet. 1995;11(2):198–200.
27. Morton NE. Sequential tests for the detection of linkage. Am J Hum Genet. 1955;7(3):277–318.
28. Thompson D, Easton DF, Goldgar DE. A full-likelihood method for the evaluation of causality of sequence variants from family data. Am J Hum Genet. 2003;73(3):652–5.
29. Domchek SM, et al. Biallelic deleterious BRCA1 mutations in a woman with early-onset ovarian cancer. Cancer Discov. 2013;3(4):399–405.
30. Judkins T, et al. Application of embryonic lethal or other obvious phenotypes to characterize the clinical significance of genetic variants found in trans with known deleterious mutations. Cancer Res. 2005;65(21):10096–103.
31. Bakry D, et al. Genetic and clinical determinants of constitutional mismatch repair deficiency syndrome: report from the constitutional mismatch repair deficiency consortium. Eur J Cancer. 2014;50(5):987–96.
32. Meyer S, et al. Fanconi anaemia, BRCA2 mutations and childhood cancer: a developmental perspective from clinical and epidemiological observations with implications for genetic counselling. J Med Genet. 2014;51(2):71–5.
33. Myers K, et al. The clinical phenotype of children with Fanconi anemia caused by biallelic FANCD1/BRCA2 mutations. Pediatr Blood Cancer. 2012;58(3):462–5.
34. Wimmer K, et al. Diagnostic criteria for constitutional mismatch repair deficiency syndrome: suggestions of the European consortium 'care for CMMRD' (C4CMMRD). J Med Genet. 2014;51(6):355–65.
35. Laduca H, et al. Utilization of multigene panels in hereditary cancer predisposition testing: analysis of more than 2,000 patients. Genet Med. 2014;16:830.
36. Adank MA, et al. CHEK2*1100delC homozygosity is associated with a high breast cancer risk in women. J Med Genet. 2011;48(12):860–3.
37. Huijts PE, et al. CHEK2*1100delC homozygosity in the Netherlands–prevalence and risk of breast and lung cancer. Eur J Hum Genet. 2014;22(1):46–51.

38. Thusberg J, Vihinen M. Pathogenic or not? And if so, then how? Studying the effects of missense mutations using bioinformatics methods. Hum Mutat. 2009;30(5):703–14.
39. Ng PC, Henikoff S. Predicting the effects of amino acid substitutions on protein function. Annu Rev Genomics Hum Genet. 2006;7:61–80.
40. Adzhubei IA, et al. A method and server for predicting damaging missense mutations. Nat Methods. 2010;7(4):248–9.
41. Kircher M, et al. A general framework for estimating the relative pathogenicity of human genetic variants. Nat Genet. 2014;46(3):310–5.
42. Mathe E, et al. Computational approaches for predicting the biological effect of p53 missense mutations: a comparison of three sequence analysis based methods. Nucleic Acids Res. 2006;34(5):1317–25.
43. Tavtigian SV, et al. Comprehensive statistical study of 452 BRCA1 missense substitutions with classification of eight recurrent substitutions as neutral. J Med Genet. 2006;43(4): 295–305.
44. Chao EC, et al. Accurate classification of MLH1/MSH2 missense variants with multivariate analysis of protein polymorphisms-mismatch repair (MAPP-MMR). Hum Mutat. 2008;29(6):852–60.
45. Association for Molecular Pathology et al. v. Myriad Genetics Inc., et al. in 569 U. S. ____ (2013). 2013.
46. Loveday C, et al. Germline mutations in RAD51D confer susceptibility to ovarian cancer. Nat Genet. 2011;43(9):879–82.
47. Gutierrez-Enriquez S, et al. About 1 % of the breast and ovarian Spanish families testing negative for BRCA1 and BRCA2 are carriers of RAD51D pathogenic variants. Int J Cancer. 2014;134(9):2088–97.
48. Osher DJ, et al. Mutation analysis of RAD51D in non-BRCA1/2 ovarian and breast cancer families. Br J Cancer. 2012;106(8):1460–3.
49. Thompson ER, et al. Analysis of RAD51D in ovarian cancer patients and families with a history of ovarian or breast cancer. PLoS One. 2013;8(1):e54772.
50. Wickramanayake A, et al. Loss of function germline mutations in RAD51D in women with ovarian carcinoma. Gynecol Oncol. 2012;127(3):552–5.
51. Vaughn CP, et al. The frequency of previously undetectable deletions involving 3′ Exons of the PMS2 gene. Genes Chromosomes Cancer. 2013;52(1):107–12.
52. Pennington KP, Swisher EM. Hereditary ovarian cancer: beyond the usual suspects. Gynecol Oncol. 2012;124(2):347–53.
53. Walsh T, et al. Mutations in 12 genes for inherited ovarian, fallopian tube, and peritoneal carcinoma identified by massively parallel sequencing. Proc Natl Acad Sci U S A. 2011;108(44):18032–7.
54. The NCCN Clinical Practice Guidelines in Oncology™ Genetic/Familial High-Risk Assessment: Breast and Ovarian V3.2013. National Comprehensive Cancer Network, Inc. 2013; Available from: http://www.nccn.org/
55. McCabe N, et al. Deficiency in the repair of DNA damage by homologous recombination and sensitivity to poly(ADP-ribose) polymerase inhibition. Cancer Res. 2006;66(16):8109–15.
56. National Cancer Institute, Clinical Trials. 2014.
57. Behjati S, et al. A pathogenic mosaic TP53 mutation in two germ layers detected by next generation sequencing. PLoS One. 2014;9(5):e96531.
58. Chen Z, et al. Enhanced sensitivity for detection of low-level germline mosaic RB1 mutations in sporadic retinoblastoma cases using deep semiconductor sequencing. Hum Mutat. 2014;35(3):384–91.
59. Coppin L, et al. VHL mosaicism can be detected by clinical next-generation sequencing and is not restricted to patients with a mild phenotype. Eur J Hum Genet. 2014;22(9):1149–52.
60. Pritchard CC, et al. A mosaic PTEN mutation causing Cowden syndrome identified by deep sequencing. Genet Med. 2013;15(12):1004–7.
61. Narod SA, et al. Should all BRCA1 mutation carriers with stage I breast cancer receive chemotherapy? Breast Cancer Res Treat. 2013;138(1):273–9.

62. Rebbeck TR, et al. Bilateral prophylactic mastectomy reduces breast cancer risk in BRCA1 and BRCA2 mutation carriers: the PROSE Study Group. J Clin Oncol. 2004;22(6):1055–62.
63. Vasen HF, et al. Revised guidelines for the clinical management of Lynch syndrome (HNPCC): recommendations by a group of European experts. Gut. 2013;62(6):812–23.
64. The NCCN Clinical Practice Guidelines in Oncology™ Colorectal Cancer Screening V1.2013. 2013; Available from: http://www.nccn.org/
65. Robson ME, et al. American Society of Clinical Oncology policy statement update: genetic and genomic testing for cancer susceptibility. J Clin Oncol. 2010;28(5):893–901.
66. Riley BD, et al. Essential elements of genetic cancer risk assessment, counseling, and testing: updated recommendations of the National Society of Genetic Counselors. J Genet Couns. 2012;21(2):151–61.
67. Fecteau H, et al. The evolution of cancer risk assessment in the era of next generation sequencing. J Genet Couns. 2014;23(4):633–9.
68. Mauer CB, et al. The integration of next-generation sequencing panels in the clinical cancer genetics practice: an institutional experience. Genet Med. 2014;16:407.
69. Spurdle AB, et al. ENIGMA–evidence-based network for the interpretation of germline mutant alleles: an international initiative to evaluate risk and clinical significance associated with sequence variation in BRCA1 and BRCA2 genes. Hum Mutat. 2012;33(1):2–7.
70. Liang J, et al. APC polymorphisms and the risk of colorectal neoplasia: a HuGE review and meta-analysis. Am J Epidemiol. 2013;177(11):1169–79.
71. Tung N, et al. Frequency of mutations in individuals with breast cancer referred for BRCA1 and BRCA2 testing using next-generation sequencing with a 25-gene panel. Cancer. 2015;121:25.

Index

W. Wu, H. Choudhry (eds.), *Next Generation Sequencing in Cancer Research, Volume 2*, DOI 10.1007/978-3-319-15811-2