Big Data and Cancer Research

Binay Panda

1 Introduction

The advent of high-throughput technology has revolutionized biological sciences in the last two decades enabling experiments on the whole genome scale. Data from such large-scale experiments are interpreted at system's level to understand the interplay among genome, transcriptome, epigenome, proteome, metabolome, and regulome. This has enhanced our ability to study disease systems, and the interplay between molecular data with clinical and epidemiological data, with habits, diet, and environment. A disproportionate amount of data has been generated in the last 5 years on disease genomes, especially using tumor tissues from different subsites, using high-throughput sequencing (HTS) instruments. Before elaborating the use of HTS technology in generating cancer-related data, it is important to describe briefly the history of DNA sequencing and the revolution of second and third generation of DNA sequencers that resulted in much of today's data deluge.

2 Sequencing Revolution

The history of DNA sequencing goes back to the late 1970s when Maxam and Gilbert [1] and Sanger, Nicklen and Coulson [2] independently showed that a stretch of DNA can be sequenced either by using chemical modification method or by chain termination method using di-deoxy nucleotides, respectively. Maxam and Gilbert's method of DNA sequencing did not gain popularity due to the usage of toxic chemicals and the di-deoxy chain termination method proposed by Professor

B. Panda (✉)
Ganit Labs, Bio-IT Centre, Institute of Bioinformatics and Applied Biotechnology,
Biotech Park, Electronic City Phase I, Bangalore 560100, India
e-mail: binay@ganitlabs.in

Fred Sanger became the de facto standard and method of choice for researchers working in the field of DNA sequencing. Many of the present-day high-throughput next-generation sequencing methods (described later) use the same principle of sequencing-by-synthesis originally proposed by Sanger. The pace, ease, and automation of the process have since grown further with the advent of PCR and other incremental, yet significant, discoveries including introduction of error-free, high fidelity enzymes, use of modified nucleotides, and better optical detection devices. It is essentially the same technology, first proposed and used by Fred Sanger [2], with modifications that led to the completion of the first draft of the Human Genome Project [3, 4] that ushered in a new era of DNA sequencing.

The idea behind some of the first generation high-throughput sequencing (HTS) assays was to take a known chemistry (predominantly the Sanger's sequencing-by-synthesis chemistry) and parallelize the assay to read hundreds of millions of growing chains of DNA rather than tens or hundreds as done with capillary Sanger sequencing. The processes for HTS comprise mainly of four distinct steps, template preparation, sequencing, image capture, and data analysis (Fig. 1). Different HTS platforms use different template preparation methods, chemistry to sequence DNA, and imaging technology that result in differences in throughput, accuracy, and running costs among platforms. As most imaging
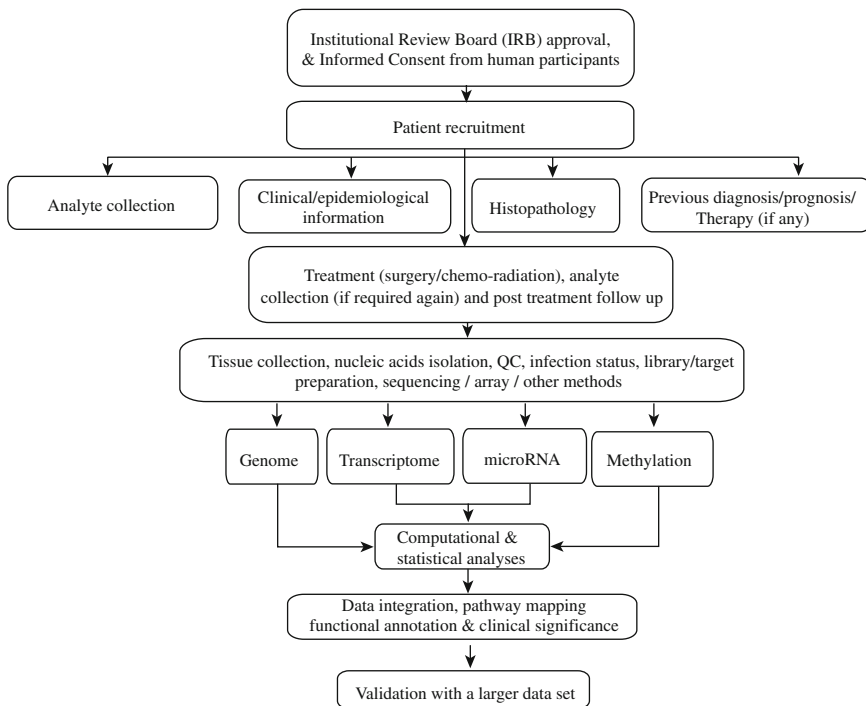

Fig. 1 Steps involved in HTS assays involving cancer patient samples and variant discovery, validation and interpretation

systems are not designed to detect single fluorescent events, clonal amplification of templates prior to imaging is incorporated as a part of template preparation before optical reading of the signal. In some cases, as in the case of single molecule sequencing, templates are not amplified but read directly to give base-level information. Some platforms are better suited than others for certain types of biological applications [5]. For discovery of actionable variants in tumors, accuracy is more important over all other parameters. Therefore, some HTS platforms are better suited to study tumor genomes over others. However, as the cost per base goes down, accuracy is increasingly achieved by higher coverage, thereby compensating errors with higher number of overlapping reads. The first publications on the human genome resequencing using the HTS system appeared in 2008 using pyrosequencing [6] and sequencing-by-synthesis using reversible terminator chemistry [7]. Since that time, the field that has gained the most amount of information using HTS platforms is cancer science. The discovery of novel DNA sequence variants in multiple cancer types using HTS platforms along with the advances in analytical methods has enabled us with the tools that have the potential to change the way cancer is currently diagnosed, treated, and managed.

3   Primary Data Generation in Cancer Studies

Various steps involved in a typical high-throughput experiment involving cancer tissue are depicted in Fig. 1. Briefly, when the patient is admitted in the hospital, clinical, epidemiological and information on habits and previous diagnosis, and treatment (if any) is recorded. Any study involving human subjects must be preapproved by an institutional review/ethics board with informed consent from all participants. Following this, analytes, full history of patients, including information on habits, and previous diagnosis/treatment (if any) are collected. Then the patients undergo treatment (surgery/chemoradiation) and the tumor tissue is collected and stored properly till further use. Once the tumor/adjacent normal/blood is collected, nucleic acids are isolated, checked for quality, and used in library/target preparation for HTS or microarray experiments. Once the raw data is collected, the data is analyzed by computational and statistical means before being integrated with clinical and epidemiological features to come up with a set of biomarkers, which is then validated in a larger cohort of patients.

4   High-Throughput Data

HTS platforms generate terabytes of data per instrument per run per week. For example, the Illumina HiSeq 4000 can generate nearly 3 terabytes of data per run in 7 days (or >400 Gb of data per day). This pose challenges for data storage, analysis, sharing, interpreting, and archiving.

Although there are many different HTS instruments in the market, the bulk of the cancer data so far have been generated using the Illumina's sequencing-by-synthesis chemistry. Therefore, a detailed description is provided on the data size, types, and complexity involved in cancer data generated by the Illumina instruments. Below is a description of different data types, usually produced during the course of a cancer high-throughput discovery study.

Despite the fact that the process of high-throughput data generation using Illumina sequencing instruments has become streamlined, never-the-less, there are inherent limitations on the quality of data generated. Some of the limitations are high degree of errors in sequencing reads (making some clinical test providers sequence up to $1000\times$ coverage or more per nucleotide to attain the requisite accuracy), shorter sequencing reads (HiSeq series of instruments do not produce data with longer than 150 nt read length), the assay not interrogating the low-complexity regions of the genome, and higher per sample cost (to gain the requisite accuracy, one needs to spend thousands of dollars per sample even for a small gene panel test). Details on different data types generated by Illumina HiSeq instrument, their approximate sizes and file type descriptions are provided in Table 1.

5 Primary Data Analysis

The cancer data analysis schema is represented in Fig. 2. First, the raw image data from the sequencing instruments are converted into *fastq* format, which is considered as the primary data files for all subsequent analysis. Before analyzing the data, the quality of the *fastq* files is checked by using tools like FastQC (http://www.bioinformatics.babraham.ac.uk/projects/fastqc/), or with *in-house* scripts to reduce sequencing quality-related bias in subsequent analysis steps. Next, sequencing reads are aligned against a reference sequence. Broadly, the alignment tools fall into two major classes, depending on which of the indexing algorithm it uses: (hash table-based or Burrows Wheeler transformation (BWT)-based). Some of the commonly used alignment tools that use hash table-based approach are Bfast [8], Ssaha [9], Smalt [10], Stampy [11] and Novoalign [12] and the ones that are BWT-based are Bowtie [13], Bowtie2 [14], and BWA [15]. BWA is the most widely used aligner by the research community. Lately, many of the alignment programs are made parallel to gain speed [16–19]. Most aligners report the results of the alignment in the form of Sequence Alignment/Map (SAM, and its binary form the BAM) format [20] that stores different flags associated with each read aligned. Before processing the aligned files for calling single (SNVs)—and/or multi (indels)—nucleotide variants, copy number variants (CNVs), and other structural variants (SVs), a few filtering, and quality checks on the SAM/BAM files are performed. These include removal of duplicates and reads mapped at multiple locations in the genome, realigning reads with known Indels, and recalibrating base quality scores with respect to the known SNVs. Once the SAM/BAM files are checked for quality, the files are used for variant calling. Although there are

Table 1 Different types of cancer data generated from a typical Illumina sequencing instrument and their descriptions.

Data types	Description
Raw data	• Multiple instruments, from multiple vendors and produced by multiple scientists • Produced as image data and never leaves the primary instrument that drives the sequencing instruments • Image data (in multi-terabytes) are automatically discarded after the initial QC • Image files are converted to base call files (BCL) after the initial QC and transferred to a server for further analyses
Primary data	• Usable data that is derived from raw data and serves as the first-level primary data for all subsequent analyses • Usually in the form of fastq files • Usually kept for 3–5 years depending on the project duration and complexity before being archived • A single fastq file size can vary anywhere between 5–15 Gb (for a 50 MB human exome at 50–100X) to 200–350 Gb (for a 30–50X whole human genome)
Secondary data	• The fastq files are aligned to a reference genome and kept as aligned file (SAM/BAM format) • A single SAM/BAM file can vary in size (2–5 GB in size for a single 50 MB human exome at 50–100X).
Tertiary data	• Files produced after the variant calls are made (in most cases VCF files) • Usually vary in size (40–250 Mb for a 50 MB human exome data with 50–100X coverage) • Kept for selected project-specific files for up to a year and then archived
Final results	• Amalgamated with other data types from multiple instruments • Used in publication • Between 10Mb - Gb in size depending on the data type for each sample • Kept for a very long time

multiple tools for calling variants, the widely used popular one is the genome analysis toolkit (GATK) [21, 22] developed at the Broad Institute, USA. GATK implements variant quality score recalibration and posterior probability calculations to minimize the false positive rates in the pool of variants called [22]. Variants are stored in a file format called variant call format (VCF), which is used by various secondary and tertiary analysis tools. Another commonly used file format for cancer data analysis is called mutation annotation format (MAF), initially made to analyze data coming out from the cancer genome atlas (TCGA) consortium. The MAF format lists all the mutations and stores much more information about the variants and alignment than the VCF files.
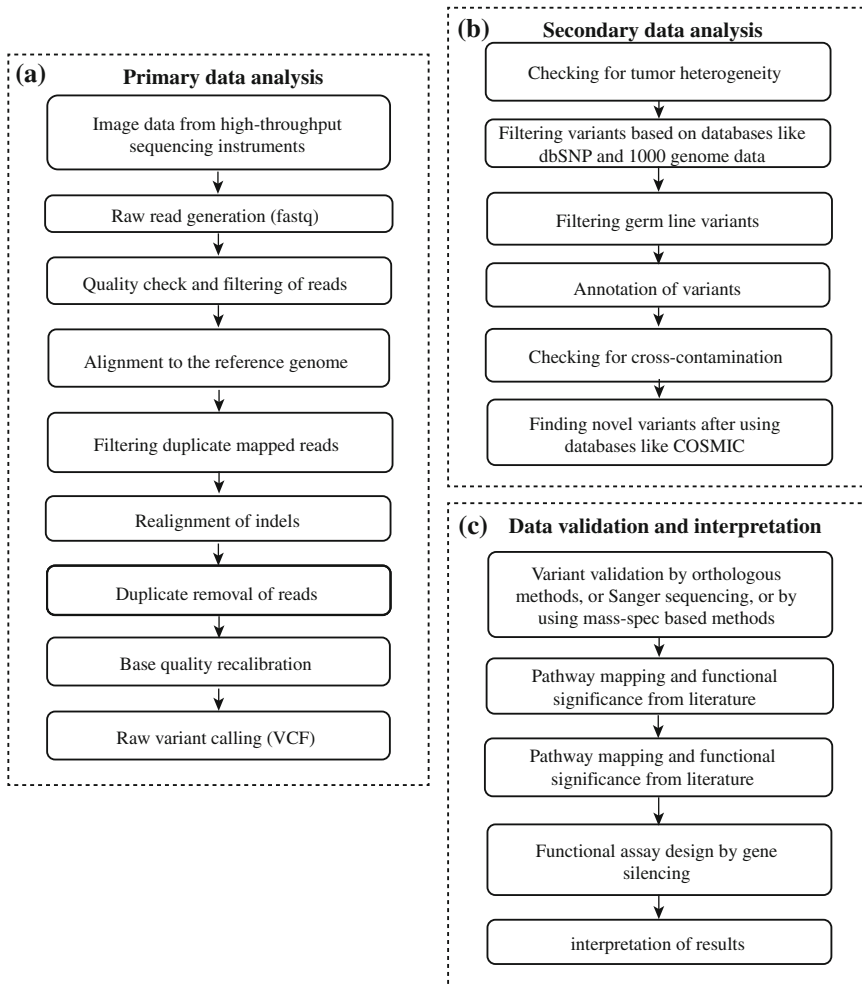
(a) **Primary data analysis**

Image data from high-throughput sequencing instruments

↓

Raw read generation (fastq)

↓

Quality check and filtering of reads

↓

Alignment to the reference genome

↓

Filtering duplicate mapped reads

↓

Realignment of indels

↓

Duplicate removal of reads

↓

Base quality recalibration

↓

Raw variant calling (VCF)

(b) **Secondary data analysis**

Checking for tumor heterogeneity

↓

Filtering variants based on databases like dbSNP and 1000 genome data

↓

Filtering germ line variants

↓

Annotation of variants

↓

Checking for cross-contamination

↓

Finding novel variants after using databases like COSMIC

(c) **Data validation and interpretation**

Variant validation by orthologous methods, or Sanger sequencing, or by using mass-spec based methods

↓

Pathway mapping and functional significance from literature

↓

Pathway mapping and functional significance from literature

↓

Functional assay design by gene silencing

↓

interpretation of results

Fig. 2 Cancer data analysis schema

6   Secondary Data Analysis

Before secondary analysis, usually the PASS variants produced by GATK (standard call confidence $>= 50$) within specific genomic bait (used in exome or gene panels) are filtered and taken for further use. Tumor-specific variants are detected by filtering out the variants found in its corresponding/paired normal sample. During this process of calling tumor-specific variants, sequencing reads representing a particular variant in a tumor sample that have no corresponding reads in matched normal

in the same location are ignored (using the callable filter of the variant caller). Then common SNPs (found in the normal population) are filtered out using a list of variants found in the databases like dbSNP and 1000 genome project. Therefore, only variants that are represented by sequencing reads both in tumor and its matched normal samples are considered. Optimization methods/workflows have been designed to analytically access the best combination of tools (both alignment and variant calling) to increase the sensitivity of variant detection [23]. The sensitivity of the alignment and variant calling tools are usually assessed by a set of metrics like aligner and variant caller-specific base quality plots of the variants called, transition/transversion (Ti/Tv) ratios, and SNP rediscovery rate using microarrays [23]. Further cross-contamination in tumor samples are assessed using tools like ContEst [24]. Searching variants against the known cancer-specific variants in databases like COSMIC [25–27] is the first step to find out whether the variant/gene is unique/novel or have been found in the same or other cancer types previously. There are cancer-specific tools to perform annotation and functional analysis. The common annotation tools are ANNOVAR [28] and VEP [29]. CRAVAT [30] provides predictive scores for different types of variants (both somatic and germline) and annotations from published literature and databases. It uses a specific cancer database with the CHASM analysis option. Genes with a CHASM score of a certain value are considered significant for comparison with other functional analyses. IntoGen [31], MutSigCV [32], and MuSiC2 [33] are other tools that are used for annotation and functional analyses of somatic variants.

7 Data Validation, Visualization, and Interpretation

Once annotated, the cancer-specific genes are validated in the same discovery set and also using a larger set of validation samples. Validation is largely done using either an orthologous sequencing method/chemistry, mass-spec-based mutation detection methods, and/or using Sanger sequencing technique. Finally the validated variants are mapped to pathways using tools like Graphite Web [34] that employs both the topological and multivariate pathway analyses with an interactive network for data visualizations. Once the network of genes is obtained, the interactions are drawn using tools like CytoScape [35–37]. Variants can also be visualized by using Circos [38], a cancer-specific portal like cbio portal [39] or with a viewer like the integrative genomics viewer (IGV) [40]. Finally, the genes that are altered in a specific cancer tissue are validated using functional screening methods using specific gene knockouts to understand their function and relationship with other genes.

8   High-Throughput Data on Human Cancers

The current projects on cancer genomics are aimed to produce a large amount of
sequence information as primary output and information on variant data (somatic
mutations, insertions and deletions, copy number variations and other structural
variations in the genome). In order to analyze the large amount of data,
high-performance compute clusters (HPC) with large memory and storage capacity
are required. Additionally, higher frequency, high-throughput multi-core chips
along with the ability to do high-volume data analysis in memory are often
required. Due to the sheer number of files, and not just the size of the files, that need
to be processed, the read/write capability is an important parameter for sequence
analysis. For effective storage and analysis of sequencing and related metadata,
network access storage systems, providing file-level access, are recommended.
Additionally, there is a need for an effective database for data organization for easy
access, management, and data update. Several data portals, primarily made by the
large consortia are developed. Prominent among them are: The Cancer Genome
Atlas (TCGA, https://tcga-data.nci.nih.gov/tcga/) data portal; cbio data portal [39]
(developed at the Memorial Sloan-Kettering Cancer Center, http://www.cbioportal.
org); the International Cancer Genome Consortium (ICGC) data portal (https://dcc.
icgc.org); and the Sanger Institute's Catalogue of Somatic Mutations in Cancer
(COSMIC) database [25] portal (http://cancer.sanger.ac.uk/cosmic).

Although biological databases are created using many different platforms, the
most common among them are MySQL and Oracle. MySQL is more popular
database because of its open source. Although the consortia-led efforts (like TCGA
and ICGC) have resulted in large and comprehensive databases covering most
cancer types, the sites are not user-friendly and do not accept external data for
integration and visualization. Therefore, efforts like cbio portal (http://www.
cbioportal.org) are required to integrate data and user-friendly data search and
retrieval. However, such efforts have to balance keeping in mind the cost and time
required versus usability and additional value addition from the new database. The
common databases use software systems known as Relational Database Manage-
ment Systems (RDBMS) that use SQL (Structured Query Language) for querying
and maintaining the databases. MySQL is a widely used open source RDBMS.
Although most biological database uses MySQL or other RDBMS, it has its lim-
itations as far as large data is concerned. First, big data is assumed to come in
structured, semi-structured, and unstructured manner. Second, traditional SQL
databases and other RDBMS lack ability to scale out a requirement for databases
containing large amount of data. Third, RDBMS cannot scale out with inexpensive
hardware. All these make RDBMS unsuitable for large data uses. This is primarily
filled by other databases like NoSQL that are document-oriented graph databases
that are non-relational, friendly to HPC environment, schema-less, and built to scale
[41]. One of the important parameters in a database is the ability to take care of
future increase in data size and complexity (Fig. 3), therefore having an ability to
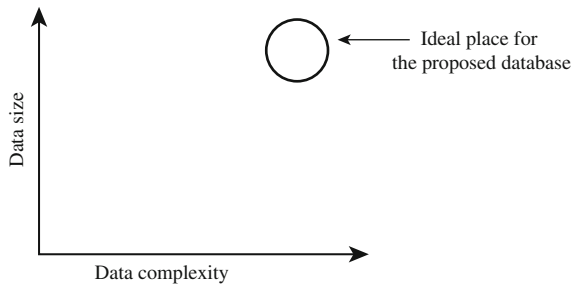scale in both these parameters. Although it is a good idea to think of databases that

Fig. 3 Two important parameters of big data and the place for an ideal database

have the ability to scale out, and accommodate variety and volume of future data increase, due to simplicity and ease of use, most small labs stick with MySQL database that uses variety of data, commonly used middleware and web server, and browser for data retrieval and visualization.

9 Large-Scale Cancer Genome Projects

Advances in technology have fuelled interest in the cancer research community that has resulted in several large publicly funded consortia-based efforts to catalogue changes in primary tumors of various types. Some of the notable and prominent efforts in this direction are, The Cancer Genome Atlas (TCGA) project (http://www.cancergenome.nih.gov/), the International Cancer Genome Consortium (ICGC) project (https://icgc.org) [42], the Cancer Genome Project (http://www.sanger.ac.uk/genetics/CGP/), and the Therapeutically applicable Research to Generate Effective Treatments (http://target.cancer.gov/) project. The National Cancer Institute (NCI) and the National Human Genome Research Institute (NHGRI) of USA initially launched the TCGA as a pilot project in 2006 even before the first human resequencing work using HTS platforms was published. The TCGA effort plans to produce a comprehensive understanding of the molecular basis of cancer and currently has grown to include samples from more than 11,000 patients across 33 different cancer types. The ICGC is an international consortium that plans to obtain a comprehensive description of various molecular changes (genomic, transcriptomic and epigenomic) in 50 different tumor types and/or subtypes. ICGC currently has participants from 18 countries studying cancer samples from more than 12000 donors on 21 tumor types. All the consortia projects are producing substantial resource for the wider cancer research community.

Till date, HTS data on several cancer types have been generated and data analysis confirmed the presence of somatic mutations in important genes, significant changes in gene/miRNA expression, hyper- and hypo-methylation in gene promoters, and structural variations in the cancer genomes [32, 43–66]. Additionally, comparative

analysis of different analytical tools have been published for cancer data analysis [23, 67–83]. Pan-cancer analyses projects have also have come up with specific and shared regions among different cancer types [32, 65, 84–88].

10   Cancer Research-Specific Challenges

There are challenges related to HTS assays using tumor tissues. For a HTS assay to have clinical utility, several challenges need to be overcome. The challenges can be clinical, technical, biological, statistical, regulatory, and market-related and are outlined in Table 2.

Clinical challenges: First among the clinical challenges is related to sample quantity. For retrospective studies to be meaningful, assays must be robust to use nucleic acids derived from formalin-fixed paraffin-embedded (FFPE) tissues. Tissue sections extracted are not often big to yield sufficient quantity of nucleic acids that can be used for sequencing, and validation studies even with the newer assays that

Table 2 Challenges of making high-throughput assays, especially sequencing-based assays, meaningful in clinics

Type of challenge	Issues
Clinical	• Sample quantity from tissue biopsies • Sample quality (FFPE tissue, chemical-crosslinking, etc.) • Tumor heterogeneity
Biological	• Background somatic mutation (needle in a haystack problem) • Varying mutation rate • Finding the right matched normal • Lack of proper controls for data assessment
Technical	• Sequencing error rate • Repeats in the genome • Read length • Stretch of homopolymers in the genome • Regions of low-complexity in the genome • Nucleotide composition
Statistical	• Large sample number requirement to catch somatic mutations at low frequency
Regulatory	• Lack of standards for clinical evaluation • Lack of controls to judge sequencing and variant calling error and accuracy • Lack of proper regulatory guidelines
Market-related	• Price • Competition • Acceptability

use only tens of nanograms of nucleic acids as starting material. Even if one manages to get enough nucleic acids from the FFPE tissues, the quality of nucleic acids extracted is not the best of quality and is often fragmented. Additionally, chemical modifications like the presence of cross-links and depurination and the presence of certain impurities in the FFPE-extracted DNA make them less amenable to alterations required for high-throughput assays. Therefore, FFPE-extracted DNA can have a stronger influence on the HTS assays. Cancer tissues are heterogeneous [32, 89] and in certain cases extremely heterogeneous (for example in pancreatic adenocarcinoma) that cautions overinterpreting HTS data from a lump of tumor tissue as shown in metastatic renal-cell carcinoma [90]. Therefore, in heterogenous tumors, the mutational burden may be underestimated. Studying such intra-tumor heterogeneity may aid the case for combination therapeutic approaches in cancer [91]. Analytical methods have been devised in the past to detect tumor heterogeneity [73, 92, 93].

Biological challenges: The next challenge is biological where finding somatic mutations, especially, those present at very low frequency, among the sea of normal background is really difficult. The use of a matched normal sample for cancer sequencing is essential to find somatic variants but, at times, the matched normal tissue might be hard to get and therefore, variants found in lymphocytes from the same patients are often used as normal samples. Another problem in sequencing tumor tissue DNA is cross-contamination. Analytical tools have been developed to detect the level of cross-contamination in tumor tissues from both sequencing and array data [24, 94]. To overcome both the heterogeneity and the cross-contamination issue, the best way is to perform DNA/RNA sequencing derived from a single tumor cell. Single-cell genomics is likely to help and improve detection, progression, and prediction of therapeutic efficacy of cancer [95]. Several reports have been published on single-cell sequencing of different cancers and analytical tool development to analyze data from a single tumor cell [96–108]. Although the problem of heterogeneity is overcome with single-cell sequencing, the fundamental questions may still linger, i.e., how many single cells have to be sequenced and if the signature is different in different single tumor cells. Additionally, there are limitations to the current protocols for isolation of single tumor cells and the inaccuracies involved in whole genome amplification of genomic DNA derived from a single cell. Therefore, capturing minute amounts of genetic material and amplifying them remain as one the greatest challenges in single cell genomics [109, 110].

Technical challenges: The third type of challenge is related to technical issues with current generation of sequencing instruments. Depending on the instrument in use, there could be an issue related to high error rate, length of the read, homopolymer stretches, and GC-rich regions in the genome. Additionally, accurate sequencing and assembling correct haplotype structures for certain regions of the genome, like the human leukocyte antigen (HLA) region, are challenging due to shorter read lengths generated in second generation DNA sequencers, presence of polymorphic exons and pseudogenes, and repeat rich region.

Statistical challenges: One of the biggest challenges to find driver mutations in cancer is related to sample number. Discovering rare driver mutations in cancer is extremely challenging, especially when sample numbers are not adequate. This, so-called, "the long tail phenomenon" is quite common in many of the cancer genome sequencing studies. Discovering rare driver mutations (found at 2 % frequency or lower) requires sequencing a large number of samples. For example, in head and neck cancer, imputations have shown that it will take 2000 tumor:normal samples to be sequenced at 90 % power in 90 % of the genes to find somatic variants present at 2 % frequency or higher [43].

Regulatory and other challenges: In order for cancer personalized medicine to become a reality, proper regulatory and policy framework need to be in place. Issues around how to deal with germline changes along with strict and proper assay and technical controls/standards are needed to be in place to assess biological, clinical, and technical accuracy and authenticity. A great beginning in this direction has already been made by the genome in a bottle consortium (https://sites.stanford.edu/abms/giab) hosted by the National Institute of Standards and Technology of the USA that has come up with reference materials (reference standards, reference methods, and reference data) to be used in sequencing. Finally, in order for cutting edge genomic tests to become a reality, collaboration and cooperation between academic centers and industry are absolutely necessary [111]. Additionally, acceptability criteria and proper pricing control mechanism(s) need to be in place by the government. This is necessary for countries like India where genomic tests are largely unregulated.

11 Conclusion

Cancer research has changed since the introduction of technologies like DNA microarray and high-throughput sequencing. It is now possible to get a genome-wide view on a particular tumor rather than looking at a handful of genes. The biggest challenge for finding actionable variants in cancer remains at the level of data analysis and understanding of their functional importance. Recent demonstrations [112–115] of gene editing systems like CRISPR-Cas9 in understanding the function of cancer-related genes and their role(s) in carcinogenesis and metastasis will play a big role in the future. Further, high-throughput sequencing technology can be used to providing information on individual cancer regulome by integrating information on genetic variants, transcript variants, regulatory proteins binding to DNA and RNA, DNA and protein methylation, and metabolites. Finally, for big data to bear fruits in cancer diagnosis, prognosis, and treatment, processes like; simplified data analytics platforms; accurate sequencing chemistry; standards for measuring clinical accuracy, precision and sensitivity; proper country-specific regulatory guidelines and stringent yet ethical framework against data misuse; need to be in place [111].

References

1. Maxam AM, Gilbert W (1977) A new method for sequencing DNA. Proc Natl Acad Sci USA 74:560–564
2. Sanger F, Nicklen S, Coulson AR (1977) DNA sequencing with chain-terminating inhibitors. Proc Natl Acad Sci USA 74:5463–5467
3. Venter JC, Adams MD, Myers EW, Li PW, Mural RJ, Sutton GG, Smith HO, Yandell M, Evans CA, Holt RA et al (2001) The sequence of the human genome. Science 291:1304–1351
4. Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, Baldwin J, Devon K, Dewar K, Doyle M, FitzHugh W et al (2001) Initial sequencing and analysis of the human genome. Nature 409:860–921
5. Metzker ML (2010) Sequencing technologies - the next generation. Nat Rev Genet 11:31–46
6. Wheeler DA, Srinivasan M, Egholm M, Shen Y, Chen L, McGuire A, He W, Chen YJ, Makhijani V, Roth GT et al (2008) The complete genome of an individual by massively parallel DNA sequencing. Nature 452:872–876
7. Bentley DR, Balasubramanian S, Swerdlow HP, Smith GP, Milton J, Brown CG, Hall KP, Evers DJ, Barnes CL, Bignell HR et al (2008) Accurate whole human genome sequencing using reversible terminator chemistry. Nature 456:53–59
8. Homer N, Merriman B, Nelson SF (2009) BFAST: an alignment tool for large scale genome resequencing. PLoS ONE 4:e7767
9. Ning Z, Cox AJ, Mullikin JC (2001) SSAHA: a fast search method for large DNA databases. Genome Res 11:1725–1729
10. SMALT [http://www.sanger.ac.uk/resources/software/smalt/]
11. Lunter G, Goodson M (2011) Stampy: a statistical algorithm for sensitive and fast mapping of Illumina sequence reads. Genome Res 21:936–939
12. Novoalign (www.novocraft.com)
13. Langmead B (2010) Aligning short sequencing reads with Bowtie. Curr Protoc Bioinform., Chap 11:Unit 11–17
14. Langmead B, Salzberg SL (2012) Fast gapped-read alignment with Bowtie 2. Nat Methods 9:357–359
15. Li H, Durbin R (2009) Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics 25:1754–1760
16. Liu Y, Schmidt B, Maskell DL (2012) CUSHAW: a CUDA compatible short read aligner to large genomes based on the Burrows-Wheeler transform. Bioinformatics 28:1830–1837
17. Klus P, Lam S, Lyberg D, Cheung MS, Pullan G, McFarlane I, Yeo G, Lam BY (2012) BarraCUDA—a fast short read sequence aligner using graphics processing units. BMC Res Notes 5:27
18. Gupta S, Choudhury S, Panda B (2014) MUSIC: A hybrid-computing environment for Burrows-Wheeler alignment for massive amount of short read sequence data. MECBME 2014 (indexed in IEEE Xplore)

19. Schatz MC, Trapnell C, Delcher AL, Varshney A (2007) High-throughput sequence alignment using graphics processing units. BMC Bioinform 8:474

20. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R (2009) The sequence alignment/map format and SAMtools. Bioinformatics 25:2078–2079

21. McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, Garimella K, Altshuler D, Gabriel S, Daly M, DePristo MA (2010) The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res 20:1297–1303

22. DePristo MA, Banks E, Poplin R, Garimella KV, Maguire JR, Hartl C, Philippakis AA, del Angel G, Rivas MA, Hanna M et al (2011) A framework for variation discovery and genotyping using next-generation DNA sequencing data. Nat Genet 43:491–498

23. Pattnaik S, Vaidyanathan S, Pooja DG, Deepak S, Panda B (2012) Customisation of the exome data analysis pipeline using a combinatorial approach. PLoS ONE 7:e30080

24. Cibulskis K, McKenna A, Fennell T, Banks E, DePristo M, Getz G (2011) ContEst: estimating cross-contamination of human samples in next-generation sequencing data. Bioinformatics 27:2601–2602

25. Forbes SA, Beare D, Gunasekaran P, Leung K, Bindal N, Boutselakis H, Ding M, Bamford S, Cole C, Ward S et al (2015) COSMIC: exploring the world's knowledge of somatic mutations in human cancer. Nucleic Acids Res 43:D805–D811

26. Forbes SA, Bindal N, Bamford S, Cole C, Kok CY, Beare D, Jia M, Shepherd R, Leung K, Menzies A et al (2011) COSMIC: mining complete cancer genomes in the Catalogue of Somatic Mutations in Cancer. Nucleic Acids Res 39:D945–D950

27. Forbes SA, Tang G, Bindal N, Bamford S, Dawson E, Cole C, Kok CY, Jia M, Ewing R, Menzies A et al (2010) COSMIC (the Catalogue of Somatic Mutations in Cancer): a resource to investigate acquired mutations in human cancer. Nucleic Acids Res 38:D652–D657

28. Wang K, Li M, Hakonarson H (2010) ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. Nucleic Acids Res 38:e164

29. Yourshaw M, Taylor SP, Rao AR, Martin MG, Nelson SF (2015) Rich annotation of DNA sequencing variants by leveraging the Ensembl Variant Effect Predictor with plugins. Brief Bioinform 16:255–264

30. Douville C, Carter H, Kim R, Niknafs N, Diekhans M, Stenson PD, Cooper DN, Ryan M, Karchin R (2013) CRAVAT: cancer-related analysis of variants toolkit. Bioinformatics 29:647–648

31. Gundem G, Perez-Llamas C, Jene-Sanz A, Kedzierska A, Islam A, Deu-Pons J, Furney SJ, Lopez-Bigas N (2010) IntOGen: integration and data mining of multidimensional oncogenomic data. Nat Methods 7:92–93

32. Lawrence MS, Stojanov P, Polak P, Kryukov GV, Cibulskis K, Sivachenko A, Carter SL, Stewart C, Mermel CH, Roberts SA et al (2013) Mutational heterogeneity in cancer and the search for new cancer-associated genes. Nature 499:214–218

33. Dees ND: MuSiC2. 2015

34. Sales G, Calura E, Martini P, Romualdi C (2013) Graphite Web: Web tool for gene set analysis exploiting pathway topology. Nucleic Acids Res 41:W89–W97

35. Lopes CT, Franz M, Kazi F, Donaldson SL, Morris Q, Bader GD (2010) Cytoscape Web: an interactive web-based network browser. Bioinformatics 26:2347–2348

36. Cline MS, Smoot M, Cerami E, Kuchinsky A, Landys N, Workman C, Christmas R, Avila-Campilo I, Creech M, Gross B et al (2007) Integration of biological networks and gene expression data using Cytoscape. Nat Protoc 2:2366–2382

37. Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT, Ramage D, Amin N, Schwikowski B, Ideker T (2003) Cytoscape: a software environment for integrated models of biomolecular interaction networks. Genome Res 13:2498–2504

38. Krzywinski M, Schein J, Birol I, Connors J, Gascoyne R, Horsman D, Jones SJ, Marra MA (2009) Circos: an information aesthetic for comparative genomics. Genome Res 19:1639–1645

39. Cerami E, Gao J, Dogrusoz U, Gross BE, Sumer SO, Aksoy BA, Jacobsen A, Byrne CJ, Heuer ML, Larsson E et al (2012) The cBio cancer genomics portal: an open platform for exploring multidimensional cancer genomics data. Cancer Discov 2:401–404

40. Robinson JT, Thorvaldsdottir H, Winckler W, Guttman M, Lander ES, Getz G, Mesirov JP (2011) Integrative genomics viewer. Nat Biotechnol 29:24–26

41. Hu H, Wen Y, Chua TS, Li X (2014) Toward scalable systems for big data analytics: a technology tutorial. IEEE Access 2:652–687

42. Hudson TJ, Anderson W, Artez A, Barker AD, Bell C, Bernabe RR, Bhan MK, Calvo F, Eerola I, Gerhard DS et al (2010) International network of cancer genome projects. Nature 464:993–998

43. Lawrence MS, Stojanov P, Mermel CH, Robinson JT, Garraway LA, Golub TR, Meyerson M, Gabriel SB, Lander ES, Getz G (2014) Discovery and saturation analysis of cancer genes across 21 tumour types. Nature 505:495–501

44. Stephens PJ, McBride DJ, Lin ML, Varela I, Pleasance ED, Simpson JT, Stebbings LA, Leroy C, Edkins S, Mudie LJ et al (2009) Complex landscapes of somatic rearrangement in human breast cancer genomes. Nature 462:1005–1010

45. van Haaften G, Dalgliesh GL, Davies H, Chen L, Bignell G, Greenman C, Edkins S, Hardy C, O'Meara S, Teague J et al (2009) Somatic mutations of the histone H3K27 demethylase gene UTX in human cancer. Nat Genet 41:521–523

46. Pleasance ED, Cheetham RK, Stephens PJ, McBride DJ, Humphray SJ, Greenman CD, Varela I, Lin ML, Ordonez GR, Bignell GR et al (2010) A comprehensive catalogue of somatic mutations from a human cancer genome. Nature 463:191–196

47. Pleasance ED, Stephens PJ, O'Meara S, McBride DJ, Meynert A, Jones D, Lin ML, Beare D, Lau KW, Greenman C et al (2010) A small-cell lung cancer genome with complex signatures of tobacco exposure. Nature 463:184–190

48. Papaemmanuil E, Cazzola M, Boultwood J, Malcovati L, Vyas P, Bowen D, Pellagatti A, Wainscoat JS, Hellstrom-Lindberg E, Gambacorti-Passerini C et al (2011) Somatic SF3B1 mutation in myelodysplasia with ring sideroblasts. N Engl J Med 365:1384–1395

49. Puente XS, Pinyol M, Quesada V, Conde L, Ordonez GR, Villamor N, Escaramis G, Jares P, Bea S, Gonzalez-Diaz M et al (2011) Whole-genome sequencing identifies recurrent mutations in chronic lymphocytic leukaemia. Nature 475:101–105

50. Stephens PJ, Greenman CD, Fu B, Yang F, Bignell GR, Mudie LJ, Pleasance ED, Lau KW, Beare D, Stebbings LA et al (2011) Massive genomic rearrangement acquired in a single catastrophic event during cancer development. Cell 144:27–40

51. Varela I, Tarpey P, Raine K, Huang D, Ong CK, Stephens P, Davies H, Jones D, Lin ML, Teague J et al (2011) Exome sequencing identifies frequent mutation of the SWI/SNF complex gene PBRM1 in renal carcinoma. Nature 469:539–542

52. Greenman CD, Pleasance ED, Newman S, Yang F, Fu B, Nik-Zainal S, Jones D, Lau KW, Carter N, Edwards PA et al (2012) Estimation of rearrangement phylogeny for cancer genomes. Genome Res 22:346–361

53. Nik-Zainal S, Alexandrov LB, Wedge DC, Van Loo P, Greenman CD, Raine K, Jones D, Hinton J, Marshall J, Stebbings LA et al (2012) Mutational processes molding the genomes of 21 breast cancers. Cell 149:979–993

54. Stephens PJ, Tarpey PS, Davies H, Van Loo P, Greenman C, Wedge DC, Nik-Zainal S, Martin S, Varela I, Bignell GR et al (2012) The landscape of cancer genes and mutational processes in breast cancer. Nature 486:400–404

55. Wang L, Tsutsumi S, Kawaguchi T, Nagasaki K, Tatsuno K, Yamamoto S, Sang F, Sonoda K, Sugawara M, Saiura A et al (2012) Whole-exome sequencing of human pancreatic cancers and characterization of genomic instability caused by MLH1 haploinsufficiency and complete deficiency. Genome Res 22:208–219

56. Cancer Genome Atlas N (2015) Comprehensive genomic characterization of head and neck squamous cell carcinomas. Nature 517:576–582

57. India Project Team of the International Cancer Genome C (2013) Mutational landscape of gingivo-buccal oral squamous cell carcinoma reveals new recurrently-mutated genes and molecular subgroups. Nat Commun 4:2873

58. Barbieri CE, Baca SC, Lawrence MS, Demichelis F, Blattner M, Theurillat JP, White TA, Stojanov P, Van Allen E, Stransky N et al (2012) Exome sequencing identifies recurrent SPOP, FOXA1 and MED12 mutations in prostate cancer. Nat Genet 44:685–689

59. Van Allen EM, Wagle N, Stojanov P, Perrin DL, Cibulskis K, Marlow S, Jane-Valbuena J, Friedrich DC, Kryukov G, Carter SL et al (2014) Whole-exome sequencing and clinical interpretation of formalin-fixed, paraffin-embedded tumor samples to guide precision cancer medicine. Nat Med 20:682–688

60. Wang L, Lawrence MS, Wan Y, Stojanov P, Sougnez C, Stevenson K, Werner L, Sivachenko A, DeLuca DS, Zhang L et al (2011) SF3B1 and other novel cancer genes in chronic lymphocytic leukemia. N Engl J Med 365:2497–2506

61. Craig DW, O'Shaughnessy JA, Kiefer JA, Aldrich J, Sinari S, Moses TM, Wong S, Dinh J, Christoforides A, Blum JL et al (2013) Genome and transcriptome sequencing in prospective metastatic triple-negative breast cancer uncovers therapeutic vulnerabilities. Mol Cancer Ther 12:104–116

62. Beltran H, Rickman DS, Park K, Chae SS, Sboner A, MacDonald TY, Wang Y, Sheikh KL, Terry S, Tagawa ST et al (2011) Molecular characterization of neuroendocrine prostate cancer and identification of new drug targets. Cancer Discov 1:487–495

63. Drier Y, Lawrence MS, Carter SL, Stewart C, Gabriel SB, Lander ES, Meyerson M, Beroukhim R, Getz G (2013) Somatic rearrangements across cancer reveal classes of samples with distinct patterns of DNA breakage and rearrangement-induced hypermutability. Genome Res 23:228–235

64. Eswaran J, Horvath A, Godbole S, Reddy SD, Mudvari P, Ohshiro K, Cyanam D, Nair S, Fuqua SA, Polyak K et al (2013) RNA sequencing of cancer reveals novel splicing alterations. Sci Rep 3:1689

65. Kandoth C, McLellan MD, Vandin F, Ye K, Niu B, Lu C, Xie M, Zhang Q, McMichael JF, Wyczalkowski MA et al (2013) Mutational landscape and significance across 12 major cancer types. Nature 502:333–339

66. Wu X, Cao W, Wang X, Zhang J, Lv Z, Qin X, Wu Y, Chen W (2013) TGM3, a candidate tumor suppressor gene, contributes to human head and neck cancer. Mol Cancer 12:151

67. Merid SK, Goranskaya D, Alexeyenko A (2014) Distinguishing between driver and passenger mutations in individual cancer genomes by network enrichment analysis. BMC Bioinform 15:308

68. Layer RM, Chiang C, Quinlan AR, Hall IM (2014) LUMPY: a probabilistic framework for structural variant discovery. Genome Biol 15:R84

69. Dietlein F, Eschner W (2014) Inferring primary tumor sites from mutation spectra: a meta-analysis of histology-specific aberrations in cancer-derived cell lines. Hum Mol Genet 23:1527–1537

70. Cole C, Krampis K, Karagiannis K, Almeida JS, Faison WJ, Motwani M, Wan Q, Golikov A, Pan Y, Simonyan V, Mazumder R (2014) Non-synonymous variations in cancer and their effects on the human proteome: workflow for NGS data biocuration and proteome-wide analysis of TCGA data. BMC Bioinform 15:28

71. Wittler R (2013) Unraveling overlapping deletions by agglomerative clustering. BMC Genom 14(Suppl 1):S12

72. Trifonov V, Pasqualucci L, Dalla Favera R, Rabadan R (2013) MutComFocal: an integrative approach to identifying recurrent and focal genomic alterations in tumor samples. BMC Syst Biol 7:25

73. Oesper L, Mahmoody A, Raphael BJ (2013) THetA: inferring intra-tumor heterogeneity from high-throughput DNA sequencing data. Genome Biol 14:R80

74. Hansen NF, Gartner JJ, Mei L, Samuels Y, Mullikin JC (2013) Shimmer: detection of genetic alterations in tumors using next-generation sequence data. Bioinformatics 29:1498–1503

75. Hamilton MP, Rajapakshe K, Hartig SM, Reva B, McLellan MD, Kandoth C, Ding L, Zack TI, Gunaratne PH, Wheeler DA et al (2013) Identification of a pan-cancer oncogenic microRNA superfamily anchored by a central core seed motif. Nat Commun 4:2730

76. Chen Y, Yao H, Thompson EJ, Tannir NM, Weinstein JN, Su X (2013) VirusSeq: software to identify viruses and their integration sites using next-generation sequencing of human cancer tissue. Bioinformatics 29:266–267

77. Mosen-Ansorena D, Telleria N, Veganzones S, De la Orden V, Maestro ML, Aransay AM (2014) seqCNA: an R package for DNA copy number analysis in cancer using high-throughput sequencing. BMC Genom 15:178

78. Li Y, Xie X (2014) Deconvolving tumor purity and ploidy by integrating copy number alterations and loss of heterozygosity. Bioinformatics 30:2121–2129

79. Kendall J, Krasnitz A (2014) Computational methods for DNA copy-number analysis of tumors. Methods Mol Biol 1176:243–259

80. Krishnan NM, Gaur P, Chaudhary R, Rao AA, Panda B (2012) COPS: a sensitive and accurate tool for detecting somatic Copy Number Alterations using short-read sequence data from paired samples. PLoS ONE 7:e47812

81. Van Allen EM, Wagle N, Levy MA (2013) Clinical analysis and interpretation of cancer genome data. J Clin Oncol 31:1825–1833

82. Lahti L, Schafer M, Klein HU, Bicciato S, Dugas M (2013) Cancer gene prioritization by integrative analysis of mRNA expression and DNA copy number data: a comparative review. Brief Bioinform 14:27–35

83. Lee LA, Arvai KJ, Jones D (2015) Annotation of sequence variants in cancer samples: processes and pitfalls for routine assays in the clinical laboratory. J Mol Diagn

84. Weinstein JN, Collisson EA, Mills GB, Shaw KR, Ozenberger BA, Ellrott K, Shmulevich I, Sander C, Stuart JM (2013) The cancer genome Atlas Pan-cancer analysis project. Nat Genet 45:1113–1120

85. Zack TI, Schumacher SE, Carter SL, Cherniack AD, Saksena G, Tabak B, Lawrence MS, Zhang CZ, Wala J, Mermel CH et al (2013) Pan-cancer patterns of somatic copy number alteration. Nat Genet 45:1134–1140

86. Gross AM, Orosco RK, Shen JP, Egloff AM, Carter H, Hofree M, Choueiri M, Coffey CS, Lippman SM, Hayes DN et al (2014) Multi-tiered genomic analysis of head and neck cancer ties TP53 mutation to 3p loss. Nat Genet 46:939–943

87. Pan-cancer initiative finds patterns of drivers (2013) Cancer Discov 3:1320

88. Taking pan-cancer analysis global (2013) Nat Genet 45:1263

89. Russnes HG, Navin N, Hicks J, Borresen-Dale AL (2011) Insight into the heterogeneity of breast cancer through next-generation sequencing. J Clin Invest 121:3810–3818

90. Gerlinger M, Rowan AJ, Horswell S, Larkin J, Endesfelder D, Gronroos E, Martinez P, Matthews N, Stewart A, Tarpey P et al (2012) Intratumor heterogeneity and branched evolution revealed by multiregion sequencing. N Engl J Med 366:883–892

91. Swanton C (2012) Intratumor heterogeneity: evolution through space and time. Cancer Res 72:4875–4882

92. Oesper L, Satas G, Raphael BJ (2014) Quantifying tumor heterogeneity in whole-genome and whole-exome sequencing data. Bioinformatics 30:3532–3540

93. Hajirasouliha I, Mahmoody A, Raphael BJ (2014) A combinatorial approach for analyzing intra-tumor heterogeneity from high-throughput sequencing data. Bioinformatics 30:i78–i86

94. Jun G, Flickinger M, Hetrick KN, Romm JM, Doheny KF, Abecasis GR, Boehnke M, Kang HM (2012) Detecting and estimating contamination of human DNA samples in sequencing and array-based genotype data. Am J Hum Genet 91:839–848

95. Navin N, Hicks J (2011) Future medical applications of single-cell sequencing in cancer. Genome Med 3:31

96. Ji C, Miao Z, He X (2015) A simple strategy for reducing false negatives in calling variants from single-cell sequencing data. PLoS ONE 10:e0123789

97. Yu C, Yu J, Yao X, Wu WK, Lu Y, Tang S, Li X, Bao L, Li X, Hou Y et al (2014) Discovery of biclonal origin and a novel oncogene SLC12A5 in colon cancer by single-cell sequencing. Cell Res 24:701–712

98. Ting DT, Wittner BS, Ligorio M, Vincent Jordan N, Shah AM, Miyamoto DT, Aceto N, Bersani F, Brannigan BW, Xega K et al (2014) Single-cell RNA sequencing identifies extracellular matrix gene expression by pancreatic circulating tumor cells. Cell Rep 8:1905–1918

99. Kim KI, Simon R (2014) Using single cell sequencing data to model the evolutionary history of a tumor. BMC Bioinform 15:27

100. Xu Y, Hu H, Zheng J, Li B (2013) Feasibility of whole RNA sequencing from single-cell mRNA amplification. Genet Res Int 2013:724124

101. Voet T, Kumar P, Van Loo P, Cooke SL, Marshall J, Lin ML, Zamani Esteki M, Van der Aa N, Mateiu L, McBride DJ et al (2013) Single-cell paired-end genome sequencing reveals structural variation per cell cycle. Nucleic Acids Res 41:6119–6138

102. Korfhage C, Fisch E, Fricke E, Baedker S, Loeffert D (2013) Whole-genome amplification of single-cell genomes for next-generation sequencing. Curr Protoc Mol Biol 104:Unit 7–14

103. Geurts-Giele WR, Dirkx-van der Velden AW, Bartalits NM, Verhoog LC, Hanselaar WE, Dinjens WN (2013) Molecular diagnostics of a single multifocal non-small cell lung cancer case using targeted next generation sequencing. Virchows Arch 462:249–254

104. Xu X, Hou Y, Yin X, Bao L, Tang A, Song L, Li F, Tsang S, Wu K, Wu H et al (2012) Single-cell exome sequencing reveals single-nucleotide mutation characteristics of a kidney tumor. Cell 148:886–895

105. Li Y, Xu X, Song L, Hou Y, Li Z, Tsang S, Li F, Im KM, Wu K, Wu H et al (2012) Single-cell sequencing analysis characterizes common and cell-lineage-specific mutations in a muscle-invasive bladder cancer. Gigascience 1:12

106. Hou Y, Song L, Zhu P, Zhang B, Tao Y, Xu X, Li F, Wu K, Liang J, Shao D et al (2012) Single-cell exome sequencing and monoclonal evolution of a JAK2-negative myeloproliferative neoplasm. Cell 148:873–885

107. Novak R, Zeng Y, Shuga J, Venugopalan G, Fletcher DA, Smith MT, Mathies RA (2011) Single-cell multiplex gene detection and sequencing with microfluidically generated agarose emulsions. Angew Chem Int Ed Engl 50:390–395

108. Navin N, Kendall J, Troge J, Andrews P, Rodgers L, McIndoo J, Cook K, Stepansky A, Levy D, Esposito D et al (2011) Tumour evolution inferred by single-cell sequencing. Nature 472:90–94

109. Lasken RS (2013) Single-cell sequencing in its prime. Nat Biotechnol 31:211–212

110. Nawy T (2014) Single-cell sequencing. Nat Methods 11:18

111. Panda B (2012) Whither genomic diagnostics tests in India? Indian J Med Paediatr Oncol 33:250–252

112. Xue W, Chen S, Yin H, Tammela T, Papagiannakopoulos T, Joshi NS, Cai W, Yang G, Bronson R, Crowley DG et al (2014) CRISPR-mediated direct mutation of cancer genes in the mouse liver. Nature 514:380–384

113. Sanchez-Rivera FJ, Papagiannakopoulos T, Romero R, Tammela T, Bauer MR, Bhutkar A, Joshi NS, Subbaraj L, Bronson RT, Xue W, Jacks T (2014) Rapid modelling of cooperating genetic events in cancer through somatic genome editing. Nature 516:428–431

114. Matano M, Date S, Shimokawa M, Takano A, Fujii M, Ohta Y, Watanabe T, Kanai T, Sato T (2015) Modeling colorectal cancer using CRISPR-Cas9-mediated engineering of human intestinal organoids. Nat Med 21:256–262

115. Chen S, Sanjana NE, Zheng K, Shalem O, Lee K, Shi X, Scott DA, Song J, Pan JQ, Weissleder R et al (2015) Genome-wide CRISPR screen in a mouse model of tumor growth and metastasis. Cell 160:1246–1260