Alessandra Rogato · Valeria Zazzu
Mario Guarracino *Editors*

Dynamics of Mathematical Models in Biology

Bringing Mathematics to Life

Springer

Dynamics of Mathematical Models in Biology

Alessandra Rogato • Valeria Zazzu
Mario Guarracino
Editors

Dynamics of Mathematical Models in Biology

Bringing Mathematics to Life

Springer

Editors

Alessandra Rogato
National Research Council
Institute of Biosciences and Bioresources
Naples, Italy

Integrative Marine Ecology
Stazione Zoologica Anton Dohrn
Napoli, Italy

Valeria Zazzu
National Research Council
Institute of Genetics and Biophysics, ABT
Naples, Italy

Mario Guarracino
Laboratory for Genomics, Transcriptomics
and Proteomics (Lab-GTP)
High Performance Computing
and Networking Institute (ICAR)
National Research Council (CNR)
Naples, Italy

ISBN 978-3-319-45722-2 ISBN 978-3-319-45723-9 (eBook)
DOI 10.1007/978-3-319-45723-9

Library of Congress Control Number: 2016955240

Mathematics Subject Classification (2010): 00B15 92B05

Printed on acid-free paper

This Springer imprint is published by Springer Nature
The registered company is Springer International Publishing AG
The registered company address is: Gewerbestrasse 11, 6330 Cham, Switzerland

Preface

Despite Galileo's claim that mathematics is the language of nature, the two disciplines of mathematics and life sciences had been considered two planets belonging to two very far galaxies which would never meet. The two communities were vastly different and it seemed impossible for them to collaborate. Only recently when life scientists began producing experimental data at an unprecedentedly high pace, did it become clear that mathematical models were necessary to interpret such data and to structure them, with the ultimate goal of unveiling biological mechanisms, to making new discoveries, and to making predictions.

There are very few examples of events that bring the two communities together to discuss research questions. For this reason we decided to create a series of annual workshops to gather a multidisciplinary and international community. "Bringing Maths to Life" enabled the two communities of life scientists and mathematicians to exchange a bidirectional flow of ideas. The broad community of mathematician enabled life scientists to introduce new algorithms, methods, and software that may be useful to model life. Biologists enabled scientists to pose new challenges for mathematicians, thereby bringing to life novel opportunities for mathematicians to explore interesting problems. From this workshop many ideas and collaboration began. In the second year of the workshop, the *leitmotiv* had surrounded the concepts of time and dynamicity of nature. In the necessary simplifications applied during the modeling process, time is sometimes not accounted for in an attempt to avoid exponential complexity in computations. Nevertheless time imposed a different thought paradigm, which in turn created more elegant mathematical models.

The second workshop, held during October 19–21, 2015, in Naples (Italy) featured three main sessions. "Dynamics of genomes and genetic variation" was the topic of the first session. In this session, we discussed the molecular mechanisms and evolutionary processes that shape the structure and function of genomes and that govern genome dynamics. The session on "Dynamics of motifs" provided an overview of current methods for motif searching in DNA, RNA, and proteins, a key process to discover emergent properties of cells, tissues, and organisms. The third session was dedicated to the "Dynamics of biological networks." Networks representing complex biological functions and activities are useful to interpret processes in the cell, and several mathematical models and algorithms are now available for their integration, analysis, and characterization. As mentioned above, in the necessary simplifications applied during the modeling process, time was often not accounted for in an effort to avoid exponential complexity in computations.

In this volume we collect many of the important ideas that derived from the workshop which are representative of the research questions that can be posed within such multidisciplinary applications. In the first chapter, Verena Thormann and colleagues describe the transcriptional regulation (when $1 + 1 \neq 2$). In the chapter "Differential Network Analysis and Graph Classification: A Glocal Approach", a glocal approach to differential network analysis and graph classification is introduced by Giuseppe Jurmal and colleagues. Maria Pia Saccomanni and Karl Thomaseth discuss the identifiability of differential equation models that are used in systems biology. In the chapter "Boolean Dynamics of Regulatory Compound Circuits", Elisabeth Remy et al. discuss the regulatory circuits and their dynamics. Target genes of homologous transcription factors are differentially analyzed by Elijah K. Lowe and colleagues. In the chapter "Reconstructing a Genetic Network from Gene Perturbations in Secretory Pathway of Cancer Cell Lines", a pipeline for gene regulatory networks reconstruction is proposed by Marina Piccirillo et al., and in the chapter "Dissecting the Functions of the Secretory Pathway by Transcriptional Profiling" the functions of secretory pathways are analyzed starting from transcriptional profiling by Sonali Gopichand Chavan and colleagues. Saraunas Germanas et al. propose a Beta-Binomial model to detect rare mutations in NGS experiments. In the chapter "An Overview of Genotyping by Sequencing in Crop Species and Its Application in Pepper", pepper genotyping by sequencing is discussed by Francesca Taranto et al. Irma Terracciano and colleagues describe in the chapter "Hybridization-Based Enrichment and Next Generation Sequencing to Explore Genetic Diversity in Plants" how to explore genetic diversity in plants. Lastly, in the chapter "DecontaMiner: A Pipeline for the Detection and Analysis of Contaminating Sequences in Human NGS Sequencing Data", Ilaria Granata and colleagues describe a pipeline for the detection of sequences belonging to contaminating organisms in human NGS sequencing data.

We would like to acknowledge the work and support we have received for realizing this volume.

The workshop has been organized by Alessandra Rogato (Institute of Biosciences and Bioresources), Valeria Zazzu and Enza Colonna (Institute of Genetics and Biophysics "Adriano Buzzati-Traverso"), and Mario Guarracino (High Performance Computing and Networking Institute and Institute for Higher Mathematics "F. Saveri") from the Italian National Research Council (CNR), Italy. Gerardo Toraldo from the Department of Mathematics and Applications "Renato Caccioppoli," University of Naples Federico II, contributed to the organization. The initiative has been supported by the Italian National Research Council (CNR), the Institute for High Mathematics "F. Saveri" (INDAM), the High Performance

Computing and Networking Institute (ICAR), the Institute of Biosciences and Bioresources (IBBR), the Institute of Genetics and Biophysics “Adriano Buzzati-Traverso” (IGB-ABT), the LABGTP, and the University of Naples Federico II.

Naples, Italy

Alessandra Rogato
Valeria Zazzu
Mario Guarracino

Contents

Contributors

Maria I. Arnone Stazione Zoologica Anton Dohrn, Naples, Italy

Marina Borschiwer Max Planck Institute for Molecular Genetics, Berlin, Germany

Concita Cantarella Consiglio per la ricerca in agricoltura e l'analisi dell'economia agraria (CREA) - Centro di ricerca per l'orticoltura, Pontecagnano Faiano (SA), Italy

Sonali Gopichand Chavan Institute of Protein Biochemistry, National Research Council (CNR-IBP), Naples, Italy

Claudia Cuomo Stazione Zoologica Anton Dohrn, Naples, Italy

Nunzio D'Agostino Consiglio per la ricerca in agricoltura e l'analisi dell'economia agraria (CREA) – Centro di ricerca per l'orticoltura, Pontecagnano Faiano (SA), Italy

Michele Filosi Fondazione Bruno Kessler, Trento, Italy

Cesare Furlanello Fondazione Bruno Kessler, Trento, Italy

Sarunas Germanas Institute of Mathematics and Informatics, Vilnius University, Vilnius, Lithuania

Ilaria Granata ICAR-CNR, Naples, Italy

Mario Guarracino Laboratory for Genomics, Transcriptomics and Proteomics (LAB-GTP), High Performance Computing and Networking Institute (ICAR), National Research Council (CNR), Naples, Italy

Audrone Jakaitiene Department of Human and Medical Genetics, Faculty of Medicine, Vilnius University, Vilnius, Lithuania

Giuseppe Jurman Fondazione Bruno Kessler, Trento, Italy

Elijah K. Lowe Stazione Zoologica Anton Dohrn, Naples, Italy

Beacon Center for Evolution in Action, Michigan State University, East Lansing, MI, USA

Alberto Luini Institute of Protein Biochemistry at National Research Council (CNR), Naples, Italy

Istituto di Ricovero e Cura a Carattere Scientifico SDN, Naples, Italy

Sebastiaan H. Meijsing Max Planck Institute for Molecular Genetics, Berlin, Germany

Brigitte Mossé Aix Marseille Université, CNRS, Centrale Marseille, I2M, UMR 7373, Marseille, France

Seetharaman Parashuraman Institute of Protein Biochemistry, National Research Council (CNR-IBP), Naples, Italy

Marina Piccirilo Laboratory for Genomics, Transcriptomics and Proteomics (LAB-GTP), High Performance Computing and Networking Institute (ICAR) at National Research Council (CNR), Naples, Italy

Elisabeth Remy Aix Marseille Université, CNRS, Centrale Marseille, I2M, UMR 7373, Marseille, France

Samantha Riccadonna Centro Ricerca e Innovazione, Fondazione Edmund Mach, San Michele all'Adige, Italy

Prathyush Deepth Roy Institute of Protein Biochemistry at National Research Council (CNR), Naples, Italy

Maria Pia Saccomani Department of Information Engineering, University of Padova, Padova, Italy

Mara Sangiovanni ICAR-CNR, Naples, Italy

Francesca Taranto Consiglio per la ricerca in agricoltura e l'analisi dell'economia agraria (CREA) – Centro di ricerca per l'orticoltura, Pontecagnano Faiano (SA), Italy

Irma Terracciano Consiglio per la ricerca in agricoltura e l'analisi dell'economia agraria (CREA) - Centro di ricerca per l'orticoltura, Pontecagnano Faiano (SA), Italy

Denis Thieffry Computational Systems Biology team, Institut de Biologie de l'Ecole Normale Supérieure (IBENS), CNRS UMR8197, INSERM U1024, Ecole Normale Supérieure, PSL Research University, Paris, France

Karl Thomaseth Institute of Electronics, Computer and Telecommunication Engineering (IEIIT-CNR) c/o DEI, Padova, Italy

Verena Thormann Max Planck Institute for Molecular Genetics, Berlin, Germany

Kumar Parijat Tripathi Laboratory for Genomics, Transcriptomics and Proteomics (LAB-GTP), High Performance Computing and Networking Institute (ICAR) at National Research Council (CNR), Naples, Italy

Pasquale Tripodi Consiglio per la ricerca in agricoltura e l'analisi dell'economia agraria (CREA) – Centro di ricerca per l'orticoltura, Pontecagnano Faiano (SA), Italy

Alessandra Varavallo Institute of Protein Biochemistry, National Research Council (CNR-IBP), Naples, Italy

Roberto Visintainer Fondazione Bruno Kessler, Trento, Italy

Transcriptional Regulation: When $1 + 1 \neq 2$

Verena Thormann, Marina Borschiwer, and Sebastiaan H. Meijsing

Abstract One of the fascinating questions in biology is to understand how an identical genome can give rise to distinct tissues with different functions, for example, brain and muscle. A key role in selectively decoding the genome is played by transcription factors (TFs), which bind to specific DNA sequences to help specify if and how much of a gene is expressed in a particular tissue. In a simple scenario, binding of TFs near a gene would result in activation of gene expression whereas in the absence of binding the gene would not be expressed. One of the objectives of computational biology is to use the genomic sequence to predict where TFs bind and to both qualitatively and quantitatively predict which genes it regulates. In this chapter, we will discuss how the information encoded in the genome in the form of DNA can serve as a discreet code where combinations of As, Ts, Cs, and Gs specify which TFs can bind. Further, structural features of DNA can be read by proteins to influence their structure and fine-tune their activity towards target genes. In practice, predicting genome-wide binding patterns of TFs based on sequence is problematic and even when we know where TFs bind, all bets appear to be off regarding the effect of TF binding on the regulation of genes. At the moment it sometimes seems as if $1 + 1 \neq 2$ when studying gene regulation. However, this mostly reflects our lack of understanding of the signaling inputs that specify if a gene is activated and at which level it is expressed. For example, in this chapter we will discuss how taking the three-dimensional organization of the genome and the chromatin context in which these binding sites are embedded into account can improve the link between binding of TFs and the regulation of genes. Eventually, by adding more and more pieces of the puzzle, we hope to identify what is missing in our current equations to model gene expression.

Keywords Transcriptional regulation • Transcription factor • Glucocorticoid receptor • Computational modeling • Hi-C • Chromatin

V. Thormann • M. Borschiwer • S.H. Meijsing (✉)
Max Planck Institute for Molecular Genetics, Ihnestrasse 63-73, 14195 Berlin, Germany
e-mail: meijsing@molgen.mpg.de

A. Rogato et al. (eds.), *Dynamics of Mathematical Models in Biology*,
DOI 10.1007/978-3-319-45723-9_1

1 Introduction

How can muscle cells have a distinct phenotype compared to blood cells although both share the same genetic information? One answer to this fundamental question is that different sets of genes are expressed in different cell types. Therefore, a detailed understanding of the mechanisms that control the expression of genes is needed to better understand how cells adopt and change their identity. Two key players in regulating the expression of genes are *cis*- and *trans*-acting elements. The *cis* elements are DNA sequences encoded in the genome that can be bound by *trans*-acting transcription factors (TFs), which in turn can influence the recruitment or activity of the RNA polymerase to influence the expression of genes. Notably, only about 1 % of the genome codes for proteins, which leaves a large fraction of the genome available for potential regulatory functions.

Activation of the right genes at the right place and at the right time is critical, as the misexpression of genes can have pathological consequences. For example, *Sonic hedgehog* is an essential gene involved in embryonic development of the limbs and a failure to express this gene results in severe limb malformations, e.g., hands with only one digit [1]. Similarly, expressing a gene in the wrong place can have detrimental effects as was shown in the fruit fly *Drosophila melanogaster* where misexpression of the *Antp* gene in the head leads to the growth of legs instead of antennas [2]. In addition to expressing the right genes at the right place, proper development requires genes to be expressed at the right level and a failure to express genes at the right dosage can lead to impaired development and disease. One well-known example is Down syndrome, where an extra copy of chromosome 21 and the resulting increased gene-dosage results in several severe developmental defects. Similarly, expressing too little of the tumor suppressor gene *p53* results in an increased chance to develop cancer [3].

To regulate the expression of genes, TFs are recruited to specific regulatory sequences, encoded in the genome (Fig. 1) [4]. These transcription factor binding sites (TFBS) are specific DNA recognition sequences located in regulatory regions. Typically, TFBSs for different TFs are found in clusters that can be referred to as enhancers. These enhancers act on the promoter of genes to influence the recruitment or activity of RNA polymerase and ultimately influence if and how much of a gene is expressed (Fig. 1). Enhancers can be located proximal to the promoter or at a large distance from the transcriptional start site (TSS) of genes [5], which raises the question how enhancers that are remote from the promoter in linear space can influence events at the promoter of genes. One explanation is the fact that looping of the DNA and its three-dimensional organization in the nucleus can bring together sequences that are remote in linear space [6]. Other levels of genome organization that influence the functioning of enhancers include the fact that the DNA in the nucleus is wrapped around histone proteins to form nucleosomes. The tails of these histones can be post-translationally modified and specific modifications were shown to correlate with the activity of enhancer elements [7].

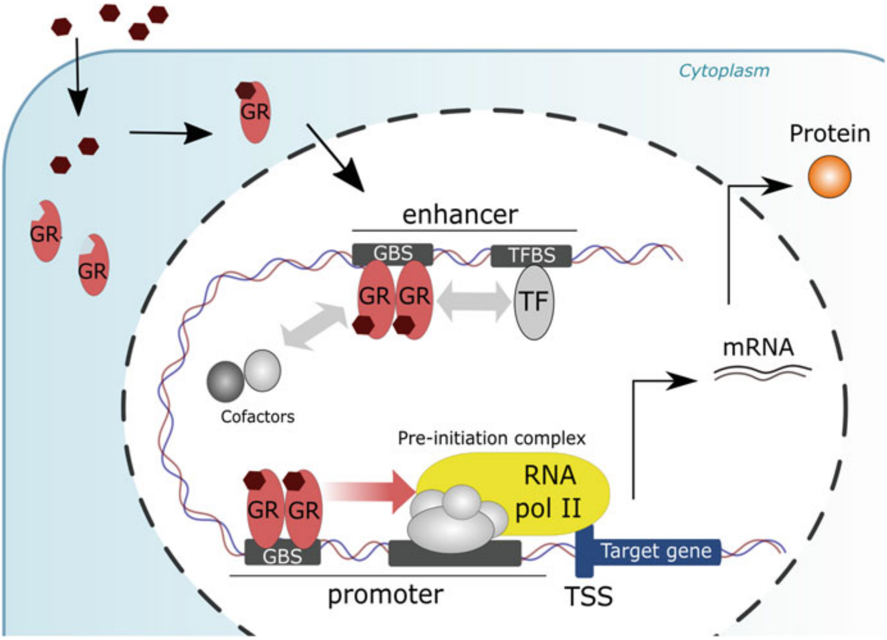


Fig. 1 Signaling pathway of the glucocorticoid receptor. Unbound glucocorticoid receptor (GR) resides in the cytoplasm and upon binding to its cognate steroid hormone (*dark red*) translocates to the nucleus where it interacts with GR binding sites (GBS) in the promoter and/or in enhancer regions. Genome-bound GR, together with cofactors and other transcription factors bound to transcription factor binding sites (TFBS), influences the recruitment and activity of RNA polymerase II to ultimately regulate the expression of its target gene

In this chapter we describe efforts to predict TF binding based on sequence and TF-dependent gene regulation based on their genome-wide binding pattern. As a model TF, we will often refer to findings for the glucocorticoid receptor (GR), a member of the nuclear steroid hormone receptor family. The reason for using this TF as a model is that GR's activity is strictly hormone-dependent. In the absence of hormone, GR resides in the cytoplasm, whereas upon hormone activation, GR translocates to the nucleus and binds to specific DNA sequences to regulate the expression of genes (Fig. 1). This hormonal on/off switch allows the relatively simple identification of putative target genes, by comparing the expression of genes between cells treated with hormone with untreated cells. Notably, depending on the response element bound, GR can either activate or repress the expression of genes [8]. Importantly, fundamental insights derived from studies using GR likely also apply to other TFs.

The aim of this chapter is to give the reader insight into mechanisms that specify where TFs bind in the genome and once bound, how they may, or may not, influence the expression of genes. First we will discuss how genomic TF binding can be predicted based on DNA sequence, but also depict the limitations of sequence-based predictions. Second, we will discuss attempts to link the binding of TFs to the regulation of genes, the role of the three-dimensional organization of the genome in the nucleus, and how the sequence of TFBSs influence how much of a target gene is expressed. Finally, we will present an outlook of how newly developed methods can contribute to our understanding of the role of TFs and TFBSs in orchestrating the expression of genes.

2 When $1 + 1 \neq 2$: The Prediction of Genomic Transcription Factor Binding Sites Based on DNA Sequence

One of the crucial steps in the regulation of gene expression is the binding of sequence-specific TFs to regulatory DNA sequences associated with their target genes. In principle, the binding of TFs can be predicted from sequence. In practice, however, sequence alone is a poor predictor of TF binding. This is in part a consequence of the fact that TFBSs are typically short and degenerate and thus potential binding sites are ubiquitously present in the genome and only a minority of these potential binding sites is actually bound by TFs. Furthermore, TF binding is often highly cell-type specific despite the fact that these cell types harbor the same genome.

Experimentally, *in vivo* genome-wide binding of TFs can be determined by chromatin immunoprecipitation (ChIP)-based techniques (Fig. 2a). As a first step of the ChIP procedure, formaldehyde is used to covalently cross-link TFs to their genomic binding sites. Subsequently, the cross-linked DNA is sheared into smaller fragments of approximately 200–300 base pairs in length and the resulting protein–DNA complexes are co-precipitated using an antibody specific for the TF of interest. Finally, either qPCR-based methods or DNA sequencing (for ChIP-seq) identifies the enrichment of DNA sequences that are occupied by a given TF (Fig. 2b). In the past decade, the advent of next generation sequencing methods resulted in a wealth of available genome-wide ChIP-seq data for different TFs from a wide variety of different cell types, tissues, and model organisms. From this data, the recognition sequence of a TF of interest can be derived using computational methods. These methods can uncover sequences that are over-represented in regions bound by a specific TF and can be used to generate a consensus motif. The consensus motif can be graphically displayed as a sequence logo to represent the position weight matrix (PWM) which describes the nucleotide preference at each nucleotide position within the motif (Fig. 2c) [9].

Conceivably, the PWM could now be used to directly predict TF binding to a given DNA sequence or even an entire genome. However, prediction of genome-wide binding based on the PWM typically fails for several reasons. First, not all DNA sequences that are bound in *vivo* match the consensus motif. This could be due to the fact that some TFs can bind to highly degenerate sequences, in which up to several base pairs can differ from its consensus sequence [10].

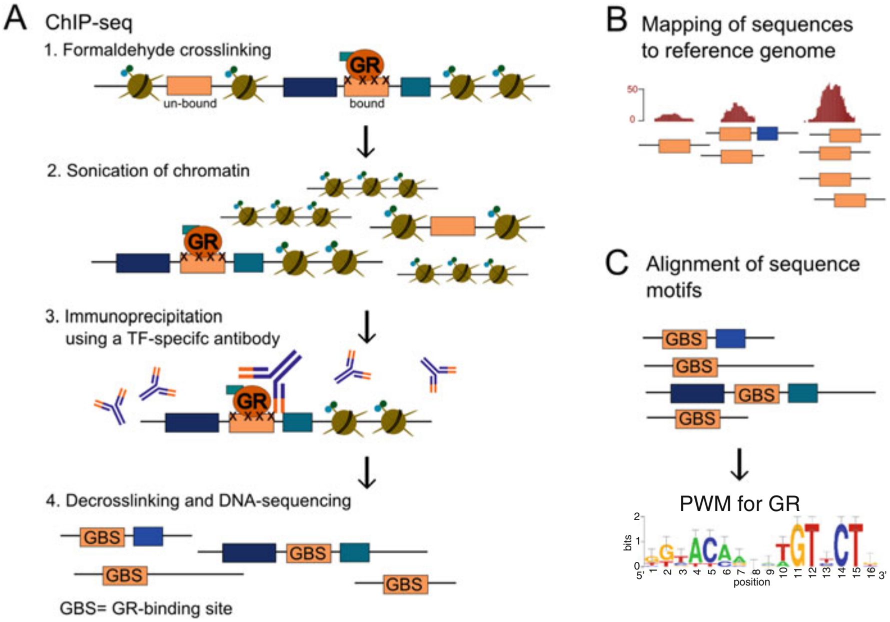

A ChIP-seq
1. Formaldehyde crosslinking
GR
un-bound
bound
2. Sonication of chromatin
GR
3. Immunoprecipitation using a TF-specifc antibody
GR
4. Decrosslinking and DNA-sequencing
GBS
GBS
GBS
GBS
GBS= GR-binding site
B Mapping of sequences to reference genome
50
0
C Alignment of sequence motifs
GBS
GBS
GBS
GBS
PWM for GR
bits
2
1
0
5'
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16
3'
position

Fig. 2 (**a**) Chromatin immunoprecipitation (ChIP)-sequencing for the identification of in *vivo* TFBSs. For ChIP-seq, protein-DNA interactions are fixed by the addition of formaldehyde. Next, the fixed chromatin is sheared into smaller fragments by sonication. DNA-fragments occupied by the TF of interest, here GR, are enriched by immunoprecipitation using a GR-specific antibody. (**b**) Generation of ChIP-seq tracks. The genome-wide location of TFBSs is analyzed by mapping and quantifying DNA-sequences obtained from ChIP-seq. (**c**) Generation of a position weight matrix (PWM). TFBSs can be represented by a PWM, describing the binding preferences of a given TF (here depicted for GR). To generate a PWM, DNA sequences obtained from ChIP-seq (or by other experimental approaches such as SELEX) are aligned and screened for TF binding motifs

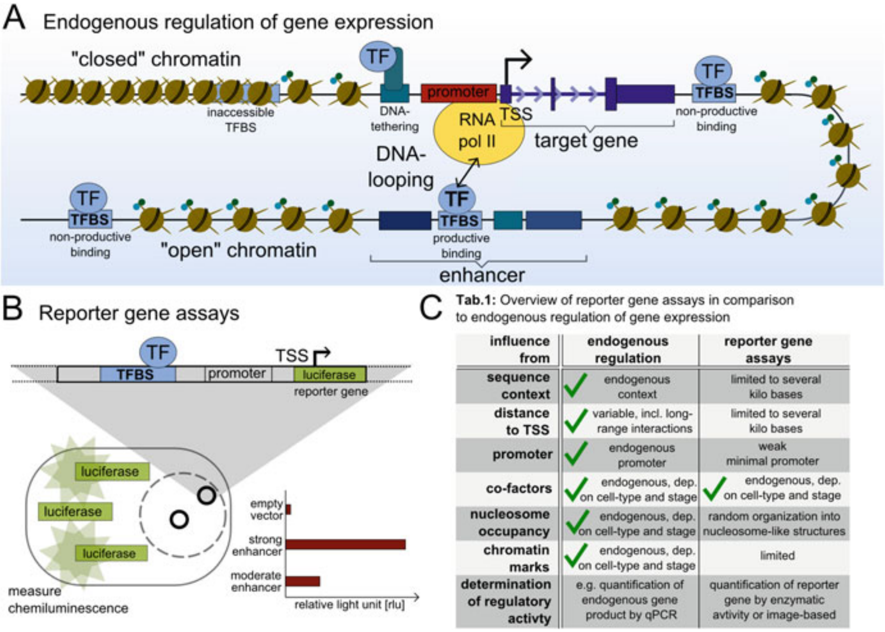
Cooperative binding with other TFs can turn such degenerate sequences into high affinity binding sites. For example, GR was reported to bind together with AP1 at composite regulatory sequences to cooperatively regulate *Notch4* gene expression [11]. Moreover, some TFs can bind without direct contact to the DNA by binding to other proteins, a mechanism referred to as DNA tethering. Hence, at tethered regions computational prediction of TFBSs using the PWM would miss indirect interactions mediated by other proteins. For example, studies using human cell lines have shown that GR binds at promoter regions of genes involved in mediating the immune-modulating actions of glucocorticoids that contain no obvious GR consensus motif. At these regions, GR-tethering to NFkB was shown to be an important mechanism responsible for GR-mediated gene regulation [12]. A second reason why PWMs fail to accurately predict genome-wide TF binding patterns is that not all computationally predicted DNA sequences are bound in *vivo*. In fact, only a minor fraction of all possible sequences matching the consensus motif of a TF are actually bound in *vivo*. For GR, the vast majority of genomic GR binding sites are located in the so-called open chromatin [13], arguing that chromatin accessibility (as assayed by DNase-I hypersensitivity assays) is a key player in specifying which of the potential binding sites encoded in the genome can be bound. Changes in chromatin accessibility, which can occur in response to environmental signals and during cellular differentiation [14, 15], can thus explain why TF occupancy can be highly cell-type [16, 17] and cell-stage specific [18].

Together, the computational prediction of genomic TFBSs suffers from two critical issues. First, false-negative predictions, when TFBSs are missed due to TF tethering by other DNA-binding factors or by binding to degenerate sequences. Notably, comparison of different computational models for the prediction of TF binding specificity showed that most often the best performing motifs were those with the highest nucleotide degeneracy [19]. Second, computational prediction of TFBSs may result in false-positive predictions for TFBSs that match the consensus but are not available for TF binding in *vivo*, e.g., due to their location in closed chromatin.

3 When $1 + 1 \neq 2$: The Prediction of Regulatory Activity Based on Genomic TF-DNA-Binding

TF binding and the regulation of nearby genes are clearly connected. However, this link is typically statistical rather than deterministic. For example, scanning a window of 300 kb around the TSS of genes showed that for all genes with a GR binding site in this window, only a fraction actually change their expression in response to GR binding (Meijsing lab unpublished results). Although the fraction of regulated genes is higher when only TFBSs in close proximity to the TSS are considered, the link between promoter-proximal GR binding and gene regulation remains far from deterministic. Similarly, ChIP-seq experiments typically uncover several thousands of peaks for an individual TF, whereas TF perturbations usually result in only a small number of affected genes [20, 21]. Consequently, TF binding is a poor predictor of gene regulation and understanding what distinguishes productive TF binding events (resulting in the regulation of a gene) from non-productive binding events remains a key challenge.

One additional signal that may help distinguish productive from non-productive binding events is the post-translational modification state of the histones located at the enhancer regions harboring TFBSs. For example, actively transcribed promoters and active enhancers show elevated levels of histone H3 lysine 27 acetylation (H3K27ac) [22, 23]. Thus, one possibility to computationally predict productive TF binding events is to combine information regarding TF binding with the occurrence of specific histone modifications. Such computational strategies were shown to be quite successful, especially when additional information such as sequence conservation, DNA accessibility, or gene expression data were also taken into account [24]. Testing if predicted enhancers are indeed capable of regulating the expression of genes is traditionally done using reporter gene assays. To test their activity, the predicted regulatory region is cloned in front of a minimal promoter sequence that drives the expression of a reporter gene, e.g., the expression of the luciferase gene (Fig. 3b). Next, the regulatory activity of a given TFBS can be analyzed in a heterologous context by measuring the amount of reporter gene activity. The presence of specific histone modifications at the enhancer region can serve as a good indicator of in *vivo* regulatory activity as detected by reporter gene assays [23]. However, the accuracy of the prediction is limited. For example, a high-throughput functional screen of enhancers computationally predicted based on their pattern of histone modifications, showed that only about one-fourth of all tested



Tab.1: Overview of reporter gene assays in comparison to endogenous regulation of gene expression

influence from	endogenous regulation	reporter gene assays
sequence context	✓ endogenous context	limited to several kilo bases
distance to TSS	✓ variable, incl. long-range interactions	limited to several kilo bases
promoter	✓ endogenous promoter	weak minimal promoter
co-factors	✓ endogenous, dep. on cell-type and stage	✓ endogenous, dep. on cell-type and stage
nucleosome occupancy	✓ endogenous, dep. on cell-type and stage	random organization into nucleosome-like structures
chromatin marks	✓ endogenous, dep. on cell-type and stage	limited
determination of regulatory activty	e.g. quantification of endogenous gene product by qPCR	quantification of reporter gene by enzymatic avtivity or image-based

Fig. 3 (**a**) Endogenous regulation of gene expression by enhancers. In *vivo*, bound TFBSs are mainly located in open chromatin regions, where they either bind directly to DNA or indirectly by tethering to other DNA-binding TFs. Productive TF binding can be influenced by the presence of associated chromatin marks or the occurrence of other co-factors. TFBSs can be located several thousands of kilo bases away from their target genes and can regulate gene expression by DNA-looping. To regulate gene expression, bound TFs influence the recruitment and activity of RNA polymerase II. (**b**) Reporter gene assays. To test the regulatory activity of a TFBS in reporter gene assays, the candidate regulatory region is cloned in front of a minimal promoter that drives the expression of a reporter gene. Upon transfection of the reporter plasmid into living cells, its regulatory activity can be analyzed by measuring the amount of generated gene product. In the depicted example, the regulatory activity of the tested regulatory region correlates with the level of luciferase activity. (**c**) Tab.1. Features influencing the regulation of gene expression in *vivo* in comparison to reporter gene assays

sequences was indeed active in reporter gene assays. Especially the classification into strong and weak enhancers based on their level of histone modifications did not have a great predictive value [25]. This could either mean that the predictions are wrong or that the reporter setting fails to recapitulate the complexity of gene regulation in the endogenous context. Regarding the latter, reporter genes differ from the endogenous genomic setting at which gene regulation takes place in a number of ways (Fig. 3). These differences include the fact that enhancers and TFBS are typically tested using a heterologous promoter and that reporters fail to recapitulate the endogenous sequence context or the chromatin environment of the investigated TFBS. Therefore, a regulatory sequence that is unable to drive reporter gene expression must not necessarily be inert in its natural genomic context. Conversely, the ability of an enhancer to activate the reporter gene does not proof that an enhancer region is capable of doing the same in the endogenous genomic context.

Notably, even when the function of putative enhancers is tested in their endogenous genomic context the results might be hard to interpret. For example, studies in Drosophila showed that the deletion of two enhancers linked to the expression of an important developmental gene resulted in only minor developmental defects when cultured under standard laboratory conditions. In contrast, at high or low temperatures the deletion of these obviously non-functional enhancers resulted in pronounced developmental defects [26]. This shows that the importance of enhancers might be context-dependent and only become apparent under specific environmental conditions. Furthermore, functional redundancy among enhancers might mask the functional importance of a specific enhancer when they are mutated individually [27].

In summary, although on a global scale the binding of TFs and the regulation of genes are clearly connected, if and how TF binding and gene regulation are linked at individual genes is typically unknown. Thus, unraveling the operating principles that specify which binding events are productive remains a major challenge.

4 When $1 + 1 \neq 2$: The Prediction of Target Genes by Incorporating the Three-Dimensional Genome Organization

In a classical view of transcriptional regulation, TFs bind to TFBSs located proximal to the promoter region of their target genes. Subsequently, bound TFBSs can serve as a binding platform to recruit other co-factors and RNA polymerase II to ultimately regulate gene expression. This classical view of gene regulation justifies approaches where target genes of a specific TFBS are predicted simply by assigning them to the gene whose TSS is closest. In support of this strategy, computational correlation of enhancer activity in reporter assays with gene expression profiles indicated that the majority of enhancers indeed act on the nearest gene. Nevertheless, up to 21 % of all enhancer candidates appeared to regulate more distal genes, suggesting that long-range regulation is a common phenomenon contributing to the complexity of transcriptional gene regulation [18]. Moreover, for many TFs, including GR, the majority of TFBSs identified by ChIP-seq are located distal from TSSs, suggesting that long-range enhancer–promoter interactions play a role in GR-mediated gene regulation [28].

For TFBSs that are localized at great linear distances from the TSS of genes, simply assigning them to the closest gene might be conceptually flawed due to the fact that based on their distance in linear space, the promoter and TFBS could in principle be located at opposite ends of the nucleus. In this case it would be unlikely that the promoter and the TFBS are functionally connected. In three-dimensional space, however, TFBSs that are remote in linear space might be in close proximity to promoters by looping of the flexible DNA polymer. Accordingly, imaging approaches have shown that remote enhancers can be in close proximity to the promoter of their target genes in three-dimensional space [29]. Furthermore, studies of the mammalian β-globin locus have uncovered that remote enhancers can regulate the expression of the closest gene, but also of other genes that are located further away in linear space [30, 31]. In three-dimensional space, however, these enhancers appear to be in close proximity to the promoters of several regulated genes within the cluster as assayed by chromosome conformation capture (3C), which maps the spatial organization of the genome in the nucleus [32, 33]. (3C)-based techniques such as 4C, 5C, or Hi-C (see [34] for a detailed overview of the different techniques) rely on the ability of formaldehyde to cross-link DNA-protein complexes and the assumption that loci in close spatial proximity have a higher probability to become part of the same cross-linked DNA-protein complex. The fixed chromatin is cut with a restriction enzyme and a subsequent ligation step joins DNA molecules that are in the same DNA-protein complex resulting in unique chimeric DNA-hybrids. Finally, the mapping of sequences obtained from DNA sequencing identifies pairs of loci that interact with a higher frequency than expected from random collision [34]. Genome-wide analysis of long-range interactions revealed that enhancer–promoter interactions predominantly take place within chromosomal units up to several megabases in size, referred to as topologically associating domains (TAD) [35]. How these chromosomal domains are established and maintained is largely unclear but two proteins, CTCF and cohesin, appear to be important for both DNA-looping and TAD establishment [36]. The resolution of 3C-based techniques, such as 4C, currently lies in the range of tens of kilobases (kbs). Of note, 3C-based methods generally suffer from a high background for regions close to the viewpoint and can therefore only give reliable information for long-distance enhancer–promoter interactions. Furthermore, it is important to keep in mind that 3C-based methods identify all physical interactions that are present in a given cell population [34]. In fact, evidence from single-cell Hi-C showed that the overall domain organization of TADs at megabase scale remains relatively stable among single cells. However, some individual chromosomal contacts could vary quite dramatically from cell to cell [37]. In addition, a high relative interaction frequency, as revealed by 3C, does not necessarily reflect a functional promoter–enhancer interaction but could also simply be a consequence of how the DNA is packaged in the nucleus or its co-localization to a distinct sub-nuclear structure [34].

Despite its limitations, 3C-based experiments have yielded important insights into the three-dimensional organization of the genome in the nucleus and its role in gene regulation. One of these studies, investigating Hi-C data across nine different cell types, revealed that the vast majority of DNA-looping interactions were highly conserved among cell types and even between different species. In addition to these invariant interactions, the study reported the occurrence of a relatively small number of cell-type specific enhancer–promoter interactions that correlated with distinct cell-type specific gene expression patterns [38]. In this context, emerging evidence suggests that the establishment of such cell-type specific DNA-loops depends on the expression of cell-type specific TFs [39, 40]. Conceivably, these TFs might either recruit other co-factors required for DNA-looping or self-assembly to efficiently bridge DNA interactions and promote cell-type specific DNA-looping [41]. In support of this hypothesis, several studies showed that a knock-out of tissue-specific TFs destabilized cell-type specific enhancer–promoter interactions [40, 42]. Another intriguing question is if, and how, these enhancer–promoter contacts might change in response to different environmental stimuli. Surprisingly, Hi-C data from human cells after TNF-alpha treatment showed no significant changes for the vast majority of DNA-looping contacts. Similarly, exposure to other stimuli such as IFN-gamma or estradiol resulted in only few changes in looping contacts. This finding suggests that most contacts are already preformed even in the absence of an activated signaling cascade [43] and that TFs that act upon these stimuli mostly use pre-established enhancer–promoter interactions. The prediction of target genes is further complicated by the possibility that promoters might interact with multiple enhancers. Given that genomic TFBSs by far outnumber the number of genes, it is indeed reasonable to assume that most genes are regulated by several enhancers. Indeed, 3C-based approaches revealed that TSS-viewpoints most often show contacts with multiple enhancer regions [38, 43]. Supporting this hypothesis, the effect of individual TF knock down on gene expression levels showed a negative correlation with the number of interacting TFBSs, suggesting that redundant TFs could rescue transcriptional outcome by integration of signals from several enhancers [21]. Furthermore, single nucleotide polymorphisms (SNP) in TFBSs at the population level rarely result in dramatic gene expression changes or disease phenotypes [44] and if effects were observed it required the presence of simultaneous SNPs in multiple enhancer regions [45]. Hence, cooperative regulation of gene expression by multiple TFBSs and the creation of regulatory hubs might be a common mechanism to ensure regulatory robustness by integrating regulatory signals from both remote contacts and promoter-proximal regions.

Notably, a recent study integrated information regarding DNA-looping from Hi-C data with genome-wide TF binding based on ChIP-seq experiments to predict TF-dependent gene regulation [43]. This study showed that genes associated with TFBSs looping to their promoters are more likely to be regulated than their counterparts that have TFBSs at the same distance without looping contacts.

Furthermore, this study found that the magnitude of gene expression changes increased with an increasing number of TFBSs that show long-range interactions [43]. This suggests that data from 3C-based approaches can help identify, or at least enrich, for productive TF binding events that result in the regulation of associated target genes. However, although the ability to predict changes in gene expression improves when taking long-range interactions into account, the connection is still far from deterministic. This might in part be due to the limited resolution of the Hi-C experiments (5–10 kb range), which could result in false-positive enhancer–promoter contacts and might thus improve further if technological advances improve the resolution of 3C-based methods.

5 When $1 + 1 \neq 2$: The Prediction of Gene Expression Level Based on Transcription Factor Binding Strength

So far we have discussed the regulatory activity of TFBSs and their associated target gene expression as an all-or-nothing event where genes are either regulated by a TF or not. However, in addition to expressing the right genes at the right time, getting the dosage of individual genes right is important for development and homeostasis. This fine-tuning of gene expression is a consequence of the integration of several signaling inputs that impinge on a gene. These inputs include the combinatorial interactions of TFs at response elements, post-translational modifications of DNA, RNA, and proteins, and processes that influence the stability of RNA once produced. Here, we will focus on one signaling input that can influence the level of expression of genes: the sequence of the TFBS and its sequence environment. One mechanism by which the sequence of a TFBS can influence transcriptional output is through differences in TF affinity, with high affinity binding sites resulting in more TF recruitment and consequently higher expression levels of associated target genes. However, in addition to affinity-driven differences in activity, TFBS sequence variants may also modulate transcriptional output by acting as allosteric ligands that influence the structure and activity of associated TFs towards their target genes.

The sequences of individual TFBSs bound by a TF typically differs between genomic loci and depending on the sequence, TFs can have higher or lower affinities for individual binding sites. In *vitro*, systematic evolution of ligands by exponential enrichment (SELEX) can be used to identify DNA sequences with the highest binding affinity for a specific TF. SELEX starts with a large initial library of random DNA oligonucleotides. From this library, high affinity binding sites are enriched by repeated cycles of TF binding followed by isolation and PCR amplification of bound sequences. The resulting pool of enriched DNA sequences can then be sequenced to identify sequences bound by the TF of interest [46] and to calculate relative TF affinities from the level of sequence enrichment [47]. In *vitro* approaches, such as SELEX, showed that the intrinsic DNA-binding affinity for a TF is in part determined by the base readout of the TF binding sequence as represented by its consensus motif. However, the base readout is not the only variable that contributes to the overall binding preference of a given TF. Evidence from structural biology showed that TF binding affinities are also influenced by the sequence-specific higher order conformation of DNA, resulting in specific bending of the DNA structure and altered protein-DNA interactions [48]. The consensus recognition motif derived from SELEX experiments captures which sequences are bound at high affinity by the TF investigated. In *vivo*, however, high affinity binding sites are not necessarily responsible for the biological consequences of TF signaling. In fact, the biological significance of low-affinity binding sites was confirmed for several TFs [49, 50]. For instance, it was shown that low-affinity binding sites of a Hox TF are responsible for the regulation of target genes in *vivo*. In addition, these low-affinity binding sites safeguard that only specific members of the HOX family of TFs can bind and activate transcription from these binding sites. Thus, low-affinity binding sites provide specificity among paralogous Hox TFs that was lost when these binding sites were changed to high affinity binding sites [50]. In support of the importance of low-affinity binding sites in gene regulation, computational modeling of enhancer evolution predicted that regulation by multiple low-affinity binding sites might be favored by evolutionary selection. A possible reason for this could be that multiple low-affinity binding sites offer more possibilities for the regulation of gene expression by changing multiple weak sites rather than one high affinity TFBS [51]. Furthermore, the usage of multiple low-affinity binding sites was suggested to enable efficient fine-tuning of gene expression in response to the integration of several signaling inputs [52]. Finally, enhancers containing multiple low-affinity binding sites for the same TF could maintain genetic redundancy and confer regulatory robustness [50].

If affinity is a major driver of transcriptional output levels, these levels can be calculated based on TFBS affinity [53, 54]. However, this occupancy hypothesis has recently been challenged by several studies showing that high affinity binding sites are not necessarily those with the highest activity [50, 55–57]. For example, the affinity of GR for different GR binding site variants determined in *vitro* does not correlate with in *vivo* transcriptional output as determined by reporter gene assays [56]. An alternative explanation for the binding site-specific activities could be that sequence variants induce distinct subtle structural changes in associated TFs which in turn influence their activity towards target genes [56, 58]. Although studying the role of the TFBS sequence on transcriptional output in isolation, where all other variables are kept the same, simplifies interpretation of the results, in reality, TFBSs are not an isolated linear stretch of DNA, but are embedded in a binding-site-specific context. Consequently, in *vivo*, additional factors contribute to the overall binding affinity and activity of a TFBS. For instance, the conformation of DNA is not only influenced by the core TFBS sequence but also by nucleotides flanking these sites [59]. Further, interactions between TFs binding at regions with multiple TFBSs modulate their interaction with the genome by direct physical interactions [60]. These interactions between TFs bound at regulatory regions can either be additive, synergistic, or antagonistic which can influence the level of transcriptional output. To complicate things even further, depending on the composition of the proteins binding at a single TFBS, GR can either act synergistically or antagonistically with these proteins [61].

Together, the multitude of mechanisms and signaling inputs that influence the expression level of genes provides the cell with a variety of mechanisms to fine-tune the expression of genes within individual cells or tissues. The effects of individual signaling inputs on gene expression may be context-specific and consequently, predicting expression levels from a limited number of features, for example, the affinity of a TF for its TFBS, is unlikely to achieve great levels of accuracy.

6 Conclusions and Future Directions

Efforts to predict TF binding based on sequence and to link TF binding to the expression of genes have failed to accurately describe the in *vivo* situation. This could be due to a lack of fundamental knowledge about the processes that specify where TFs bind and what determines if these binding events are productive in terms of resulting in gene expression changes. Recent advances have shown that adding additional knowledge can greatly advance our understanding. For example, adding information regarding chromatin accessibility helps to explain why TFs bind to only a subset of potential TFBSs, namely those that are accessible [62]. Similarly, adding knowledge regarding the three-dimensional organization of the genome improves the correlation between TF binding and gene regulation [43], which might further improve with increased resolution of the assays used to catalog the 3D genome organization. Collectively, 3C-based approaches have yielded important insights into the organization of enhancer–promoter interactions and for the prediction of candidate TFBS target genes. However, only few loci have been comprehensively investigated with respect to the relative contribution of individual enhancers to the overall expression of its target gene.

High-throughput functional testing of enhancer sequences is further facilitated by the development of several massive parallel enhancer analysis methods. For example, the STARR-seq method (Self Transcribing Regulatory Regions) [63] in which active enhancers drive their own expression. Next generation sequencing can subsequently reveal the sequence identity of active enhancers and quantitative information about their activity. This method can also be used to study how the combinatorial action of several TFBSs influences transcriptional output or the interplay between different core promoters and enhancers [64, 65]. Generally though, reporter gene assays differ in several fundamental ways from gene regulation in the endogenous genomic setting (Fig. 3c) and thus have a limited ability to uncover how gene regulation is orchestrated in *vivo*. This problem might be circumvented using genome-editing tools to study the role of enhancers in their endogenous genomic context. For example, the CRISPR/Cas9 system can be used for targeted disruption of enhancers to study their role in the regulation of genes [66]. This can help uncover if the effects of individual enhancers are additive, or if they act in a mutually redundant fashion. Further, by studying large numbers of enhancers, general principles that distinguish productive from non-productive binding events might become clear. Another interesting application of the CRISPR/Cas9 methodology is that usage of enzymatically inactive Cas9 enzymes fused to activators or inhibitors enables in *vivo* manipulation of enhancer activity [67]. Moreover, inactive Cas9 enzymes fused to chromatin modifying enzymes can be used to study the role of specific chromatin marks at individual loci [68].

Together, the combination of in *vivo*, in *vitro*, and in *silico* methods may ultimately provide the variables that are missing in our current equations and explain why currently it seems as if $1 + 1 \neq 2$ when we try to quantitatively and qualitatively model gene regulation. A greater understanding of the gene regulatory landscape could also be of therapeutic relevance as increasing evidence suggests that sequence variations in non-coding regions are a cause for several diseases including cancer, developmental, metabolic, immune, and neuropsychiatric disorders [69, 70].

References

1. Chiang, C., et al.: Manifestation of the limb prepattern: limb development in the absence of sonic hedgehog function. Dev. Biol. **236**(2), 421–435 (2001)
2. Struhl, G.: A homoeotic mutation transforming leg to antenna in Drosophila. Nature **292**(5824), 635–638 (1981)
3. Donehower, L.A., et al.: Mice deficient for p53 are developmentally normal but susceptible to spontaneous tumours. Nature **356**(6366), 215–221 (1992)
4. Consortium, E.P.: An integrated encyclopedia of DNA elements in the human genome. Nature **489**(7414), 57–74 (2012)
5. Bulger, M., Groudine, M.: Functional and mechanistic diversity of distal transcription enhancers. Cell **144**(3), 327–339 (2011)
6. de Laat, W., Duboule, D.: Topology of mammalian developmental enhancers and their regulatory landscapes. Nature **502**(7472), 499–506 (2013)
7. Calo, E., Wysocka, J.: Modification of enhancer chromatin: what, how, and why? Mol. Cell **49**(5), 825–837 (2013)
8. Meijsing, S.H.: Mechanisms of glucocorticoid-regulated gene transcription. Adv. Exp. Med. Biol. **872**, 59–81 (2015)
9. Zhang, Z., et al.: Evolutionary optimization of transcription factor binding motif detection. Adv. Exp. Med. Biol. **827**, 261–274 (2015)
10. Zhang, C., et al.: A clustering property of highly-degenerate transcription factor binding sites in the mammalian genome. Nucleic Acids Res. **34**(8), 2238–2246 (2006)
11. Wu, J., Bresnick, E.H.: Glucocorticoid and growth factor synergism requirement for Notch4 chromatin domain activation. Mol. Cell Biol. **27**(6), 2411–2422 (2007)
12. Rao, N.A., et al.: Coactivation of GR and NFKB alters the repertoire of their binding sites and target genes. Genome Res. **21**(9), 1404–1416 (2011)
13. Biddie, S.C., et al.: Transcription factor AP1 potentiates chromatin accessibility and glucocorticoid receptor binding. Mol. Cell **43**(1), 145–155 (2011)
14. West, J.A., et al.: Nucleosomal occupancy changes locally over key regulatory regions during cell differentiation and reprogramming. Nat. Commun. **5**, 4719 (2014)
15. He, H.H., et al.: Differential DNase I hypersensitivity reveals factor-dependent chromatin dynamics. Genome Res. **22**(6), 1015–1025 (2012)
16. Gertz, J., et al.: Distinct properties of cell-type-specific and shared transcription factor binding sites. Mol. Cell **52**(1), 25–36 (2013)
17. Morikawa, M., et al.: ChIP-seq reveals cell type-specific binding patterns of BMP-specific Smads and a novel binding motif. Nucleic Acids Res. **39**(20), 8712–8727 (2011)
18. Kvon, E.Z., et al.: Genome-scale functional characterization of Drosophila developmental enhancers in *vivo*. Nature **512**(7512), 91–95 (2014)

19. Weirauch, M.T., et al.: Evaluation of methods for modeling transcription factor sequence specificity. Nat. Biotechnol. **31**(2), 126–134 (2013)
20. Cusanovich, D.A., et al.: The functional consequences of variation in transcription factor binding. PLoS Genet. **10**(3), e1004226 (2014)
21. Gitter, A., et al.: Backup in gene regulatory networks explains differences between binding and knockout results. Mol. Syst. Biol. **5**, 276 (2009)
22. Creyghton, M.P., et al.: Histone H3K27ac separates active from poised enhancers and predicts developmental state. Proc. Natl. Acad. Sci. U. S. A. **107**(50), 21931–21936 (2010)
23. Heintzman, N.D., et al.: Distinct and predictive chromatin signatures of transcriptional promoters and enhancers in the human genome. Nat. Genet. **39**(3), 311–318 (2007)
24. Hardison, R.C., Taylor, J.: Genomic approaches towards finding cis-regulatory modules in animals. Nat. Rev. Genet. **13**(7), 469–483 (2012)
25. Kwasnieski, J.C., et al.: High-throughput functional testing of ENCODE segmentation predictions. Genome Res. **24**(10), 1595–1602 (2014)
26. Frankel, N., et al.: Phenotypic robustness conferred by apparently redundant transcriptional enhancers. Nature **466**(7305), 490–493 (2010)
27. Spivakov, M.: Spurious transcription factor binding: non-functional or genetically redundant? Bioessays **36**(8), 798–806 (2014)
28. So, A.Y., et al.: Determinants of cell- and gene-specific transcriptional regulation by the glucocorticoid receptor. PLoS Genet. **3**(6), e94 (2007)
29. Amano, T., et al.: Chromosomal dynamics at the Shh locus: limb bud-specific differential regulation of competence and active transcription. Dev. Cell **16**(1), 47–57 (2009)
30. Levings, P.P., Bungert, J.: The human beta-globin locus control region. Eur. J. Biochem. **269**(6), 1589–1599 (2002)
31. Hilton, I.B., et al.: Epigenome editing by a CRISPR-Cas9-based acetyltransferase activates genes from promoters and enhancers. Nat. Biotechnol. **33**(5), 510–517 (2015)
32. Tolhuis, B., et al.: Looping and interaction between hypersensitive sites in the active beta-globin locus. Mol. Cell **10**(6), 1453–1465 (2002)
33. Dekker, J., et al.: Capturing chromosome conformation. Science **295**(5558), 1306–1311 (2002)
34. Dekker, J., Marti-Renom, M.A., Mirny, L.A.: Exploring the three-dimensional organization of genomes: interpreting chromatin interaction data. Nat. Rev. Genet. **14**(6), 390–403 (2013)
35. Dixon, J.R., et al.: Topological domains in mammalian genomes identified by analysis of chromatin interactions. Nature **485**(7398), 376–380 (2012)
36. Zuin, J., et al.: Cohesin and CTCF differentially affect chromatin architecture and gene expression in human cells. Proc. Natl. Acad. Sci. **111**(3), 996–1001 (2014)
37. Nagano, T., et al.: Single-cell Hi-C reveals cell-to-cell variability in chromosome structure. Nature **502**(7469), 59–64 (2013)
38. Rao, S.S., et al.: A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping. Cell **159**(7), 1665–1680 (2014)
39. Fang, F., et al.: Coactivators p300 and CBP maintain the identity of mouse embryonic stem cells by mediating long-range chromatin structure. Stem Cells **32**(7), 1805–1816 (2014)
40. Drissen, R., et al.: The active spatial organization of the beta-globin locus requires the transcription factor EKLF. Genes Dev. **18**(20), 2485–2490 (2004)
41. Bouwman, B.A., de Laat, W.: Getting the genome in shape: the formation of loops, domains and compartments. Genome Biol. **16**, 154 (2015)
42. Vakoc, C.R., et al.: Proximity among distant regulatory elements at the beta-globin locus requires GATA-1 and FOG-1. Mol. Cell **17**(3), 453–462 (2005)
43. Jin, F., et al.: A high-resolution map of the three-dimensional chromatin interactome in human cells. Nature **503**(7475), 290–294 (2013)
44. Kilpinen, H., et al.: Coordinated effects of sequence variation on DNA binding, chromatin structure, and transcription. Science **342**(6159), 744–747 (2013)
45. Corradin, O., et al.: Combinatorial effects of multiple enhancer variants in linkage disequilibrium dictate levels of gene expression to confer susceptibility to common traits. Genome Res. **24**(1), 1–13 (2014)

46. Wang, J., et al.: In *vitro* DNA-binding profile of transcription factors: methods and new insights. J. Endocrinol. **210**(1), 15–27 (2011)
47. Slattery, M., et al.: Cofactor binding evokes latent differences in DNA binding specificity between Hox proteins. Cell **147**(6), 1270–1282 (2011)
48. Stella, S., Cascio, D., Johnson, R.C.: The shape of the DNA minor groove directs binding by the DNA-bending protein Fis. Genes Dev. **24**(8), 814–826 (2010)
49. Ramos, A.I., Barolo, S.: Low-affinity transcription factor binding sites shape morphogen responses and enhancer evolution. Philos. Trans. R. Soc. Lond. B Biol. Sci. **368**(1632), 20130018 (2013)
50. Crocker, J., et al.: Low affinity binding site clusters confer hox specificity and regulatory robustness. Cell **160**(1-2), 191–203 (2015)
51. He, X., Duque, T.S., Sinha, S.: Evolutionary origins of transcription factor binding site clusters. Mol. Biol. Evol. **29**(3), 1059–1070 (2012)
52. Gao, R., Stock, A.M.: Temporal hierarchy of gene expression mediated by transcription factor binding affinity and activation dynamics. mBio **6**(3), e00686-15 (2015)
53. Bain, D.L., et al.: Glucocorticoid receptor-DNA interactions: binding energetics are the primary determinant of sequence-specific transcriptional activity. J. Mol. Biol. **422**(1), 18–32 (2012)
54. Segal, E., et al.: Predicting expression patterns from regulatory sequence in Drosophila segmentation. Nature **451**(7178), 535–540 (2008)
55. Garcia, H.G., et al.: Operator sequence alters gene expression independently of transcription factor occupancy in bacteria. Cell Rep. **2**(1), 150–161 (2012)
56. Meijsing, S.H., et al.: DNA binding site sequence directs glucocorticoid receptor structure and activity. Science **324**(5925), 407–410 (2009)
57. Hammar, P., et al.: Direct measurement of transcription factor dissociation excludes a simple operator occupancy model for gene regulation. Nat. Genet. **46**(4), 405–408 (2014)
58. Zhang, J., et al.: DNA binding alters coactivator interaction surfaces of the intact VDR-RXR complex. Nat. Struct. Mol. Biol. **18**(5), 556–563 (2011)
59. Rohs, R., et al.: Nuance in the double-helix and its role in protein-DNA recognition. Curr. Opin. Struct. Biol. **19**(2), 171–177 (2009)
60. Meyer, M.B., Benkusky, N.A., Pike, J.W.: Selective distal enhancer control of the Mmp13 gene identified through clustered regularly interspaced short palindromic repeat (CRISPR) genomic deletions. J. Biol. Chem. **290**(17), 11093–11107 (2015)
61. Diamond, M.I., et al.: Transcription factor interactions: selectors of positive or negative regulation from a single DNA element. Science **249**(4974), 1266–1272 (1990)
62. John, S., et al.: Chromatin accessibility pre-determines glucocorticoid receptor binding patterns. Nat. Genet. **43**(3), 264–268 (2011)
63. Arnold, C.D., et al.: Genome-wide quantitative enhancer activity maps identified by STARR-seq. Science **339**(6123), 1074–1077 (2013)
64. Zabidi, M.A., et al.: Enhancer-core-promoter specificity separates developmental and housekeeping gene regulation. Nature **518**(7540), 556–559 (2015)
65. Dupin, C., et al.: Treatment of head and neck paragangliomas with external beam radiation therapy. Int. J. Radiat. Oncol. Biol. Phys. **89**(2), 353–359 (2014)
66. Korkmaz, G., et al.: Functional genetic screens for enhancer elements in the human genome using CRISPR-Cas9. Nat. Biotechnol. **34**(2), 192–198 (2016)
67. Maeder, M.L., et al.: CRISPR RNA-guided activation of endogenous human genes. Nat Methods **10**(10), 977–979 (2013)
68. Mendenhall, E.M., et al.: Locus-specific editing of histone modifications at endogenous enhancers. Nat. Biotechnol. **31**(12), 1133–1136 (2013)
69. Lupianez, D.G., et al.: Disruptions of topological chromatin domains cause pathogenic rewiring of gene-enhancer interactions. Cell **161**(5), 1012–1025 (2015)
70. Zhang, X., et al.: Identification of focally amplified lineage-specific super-enhancers in human epithelial cancers. Nat. Genet. **48**(2), 176–182 (2016)

Differential Network Analysis and Graph Classification: A Glocal Approach

Giuseppe Jurman, Michele Filosi, Samantha Riccadonna, Roberto Visintainer, and Cesare Furlanello

Abstract Based on the glocal HIM metric and its induced graph kernel, we propose a novel solution in differential network analysis that integrates network comparison and classification tasks. The HIM distance is defined as the one-parameter family of product metrics linearly combining the normalised Hamming distance H and the normalised Ipsen–Mikhailov spectral distance IM. The combination of the two components within a single metric allows overcoming their drawbacks and obtaining a measure that is simultaneously global and local. Furthermore, plugging the HIM kernel into a Support Vector Machine gives us a classification algorithm based on the HIM distance. First, we outline the theory underlying the metric construction. We introduce two diverse applications of the HIM distance and the HIM kernel to biological datasets. This versatility supports the adoption of the HIM family as a general tool for information extraction, quantifying difference among diverse instances of a complex system. An Open Source implementation of the HIM metrics is provided by the R package *nettools* and in its web interface ReNette.

Keywords Differential network • Network distance

1 Introduction

The paradigm shift towards complex systems science [3], stimulated by its recent theoretical and computational advances [4, 15], has paved the way for a parallel leap in computational biology by moving the focus from the differential gene expression analysis to differential network analysis (NetDA) [16, 25]. Due to the heterogeneity in the NetDA process and potential ill-posedness of some of the involved functional operations [1, 5, 38], a number of alternative approaches have appeared in the literature, with different strategies and aims [6, 7, 10, 16, 22, 23, 25, 41, 45, 50, 51].

G. Jurman (✉) • M. Filosi • R. Visintainer • C. Furlanello
Fondazione Bruno Kessler, via Sommarive 18 Povo, I-38122 Trento, Italy
e-mail: jurman@fbk.eu; filosi@fbk.eu; visintainer@fbk.eu; furlan@fbk.eu

S. Riccadonna
Centro Ricerca e Innovazione, Fondazione Edmund Mach, I-38010 San Michele all'Adige, Italy
e-mail: samantha.riccadonna@fmach.it

A. Rogato et al. (eds.), *Dynamics of Mathematical Models in Biology*,
DOI 10.1007/978-3-319-45723-9_2

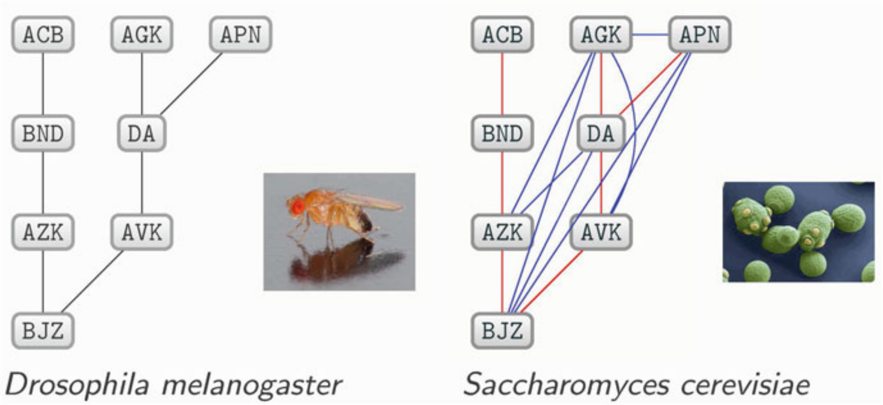


Fig. 1 A pair of similar subgraphs from a comparison of *D. melanogaster* and *S. cerevisiae* protein–protein interaction network as shown in [12]. *Blue links* are present only in the *S. cerevisiae* subnet

For example, NetDA can be used to compare networks corresponding to different organisms, phenotypes or conditions. The subgraph of the protein–protein interaction network shown in Fig. 1 (from [12]) is the same in terms of shared nodes for the fruit fly and the budding yeast. A group of links is shared by both instances of the subgraph, but the budding yeast network includes nine additional edges. Clearly, when graphs to compare have a more complex structure, more sophisticated quantitative indicators are needed also to ensure a reproducible analysis [26]. In general, the two key applications of NetDA are network comparison and network classification. Both can be framed in terms of similarity between graphs, which is best dealt with by defining a distance. However, non-metric alternatives can be used [17, 49], and even combinations of metric and statistical approaches [35, 43].

Here we propose to use the Hamming–Ipsen–Mikhailov (HIM) distance [31, 32] first as the underlying metric for the NetDA framework, and also to induce a kernel for classification purposes. The HIM metric linearly combines two distances, the Hamming [18, 24, 28, 40, 48] and the Ipsen–Mikhailov [27]; the first is an edit distance, while the latter is a spectral measure. These are the two most relevant families of graph distances: the edit distances are based on functions of insertion and deletion of matching links between the compared graphs, while the spectral measures are functions of the eigenvalues of one of the graph connectivity matrices. The Ipsen–Mikhailov distance was chosen after a comparative review [30], while Hamming was selected as the simplest member of the edit family. As a characterising feature, HIM is a glocal distance that overcomes the drawbacks of local (edit) and global (spectral) metrics when separately considered. In fact, local functions disregard the overall network structure, while spectral measures cannot distinguish isospectral graphs. Superiority of using the HIM distance over H or IM separately in practical applications is shown in the literature: NetDA based on the HIM distance has been used in metagenomics [52], MEG neuroimaging [21], liver high-throughput oncogenomics [20] and oncoimmunology [39]. In all cases, the findings derived by NetDA have been validated by matching the obtained quantitative outcomes with the qualitative biological knowledge reported in the literature. Moreover, the same method has found applicability also out of computational biology, e.g., socioeconomics [32] or even in multiplex network theory [29]. Here we present, after a brief summary of the main definitions, two novel application examples, in neurogenomics and in developmental functional genomics. In the first example, we highlight and quantify weighted network dissimilarities among gene expression of brain tissues with different phenotypes (location, sex and health status), while in the latter we describe the trajectory of the binary developmental gene network in fruit fly across its different life stages.

Finally, we describe the CRAN R package *nettools* and the web framework ReNette [19], which are available to implement NetDA projects.

2 The HIM Distance and Kernel

We recap hereafter the main definitions and results about the HIM metric and kernel. The synthesis is based on the notations of Table 1: a fully detailed description, including mathematical proofs, goes beyond the scope of the present chapter, and it is included in [31]. The (normalised) Hamming distance [18, 24, 28, 40, 48] is the (local) simplest edit metric, counting the presence/absence of matching links:

$$\mathrm{H}(\mathscr{N}_1,\mathscr{N}_2)=\frac{\mathrm{Hamming}(\mathscr{N}_1,\mathscr{N}_2)}{\mathrm{Hamming}(\mathscr{E}_N,\mathscr{F}_N)}=\frac{1}{N(N-1)}\sum_{1\leq i\neq j\leq N}|A^{(1)}_{ij}-A^{(2)}_{ij}|\ .$$

By definition, H ranges between 0 and 1, where

$$\mathrm{H}=0\text{ for }\ \mathrm{A}^{(1)}=\mathrm{A}^{(2)}\text{and }\mathrm{H}=1\text{ for }\ \mathrm{A}^{(1)}+\mathrm{A}^{(2)}=1_{\mathrm{N}}-\mathbb{I}_{\mathrm{N}}.$$

Note that, for H, all links are equivalent regardless of their position within the network: for instance, in Fig. 2, both networks B_1 and B_2 differ from A for just one link, and thus $H(A,B_1)=H(A,B_2)$, although B_1 is connected as A while B_2 is not.The Ipsen–Mikhailov distance [27] is the (global) L_2 integrated difference of the Laplacian spectral densities:

$$\mathrm{IM}(\mathscr{N}_1,\mathscr{N}_2)=\sqrt{\int_0^\infty[\rho_{\mathscr{N}_1}(\omega,\overline{\gamma})-\rho_{\mathscr{N}_2}(\omega,\overline{\gamma})]^2\,\mathrm{d}\omega}\ .$$

The definition of IM follows the dynamical interpretation of an N nodes network as an N molecules system connected by identical elastic strings, where the pattern of connections is defined by the adjacency matrix A of the corresponding network. The dynamics of the system is described by the set of N differential equations

Table 1 Notation and list of symbols

$\mathscr{N}_1, \mathscr{N}_2$	Simple networks on N nodes $\{z_i\}_{i=1}^n$
$A^{(1)}, A^{(2)}$	Corresponding adjacency matrices, with $a_{ij}^{(1)}, a_{ij}^{(2)} \in \mathscr{F}$
$\mathscr{F}$	Field $\mathbb{F}_2 = \{0, 1\}$ (unweighted case) or $[0, 1] \subseteq \mathbb{R}$ (weighted case)
$\mathbb{I}_N$	Identity matrix $\begin{pmatrix} 1 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 \\ & \vdots & & \\ 0 & 0 & \cdots & 1 \end{pmatrix}$
1_N	Unitary matrix $\begin{pmatrix} 1 & 1 & \cdots & 1 \\ 1 & 1 & \cdots & 1 \\ & \vdots & & \\ 1 & 1 & \cdots & 1 \end{pmatrix}$
0_N	Zero matrix $\begin{pmatrix} 0 & 0 & \cdots & 0 \\ 0 & 0 & \cdots & 0 \\ & \vdots & & \\ 0 & 0 & \cdots & 0 \end{pmatrix}$
$\mathscr{E}_N$	Empty network (adjacency matrix 0_N)
$\mathscr{F}_N$	Clique (adjacency matrix $1_N - \mathbb{I}_N$)
∂_g	Degree of node z_g, $\partial_g = \partial(z_g) = \sum_{j=1}^N A_{gj}$
D	Degree matrix $\begin{pmatrix} \partial_1 & 0 & \cdots & 0 \\ 0 & \partial_2 & \cdots & 0 \\ & \vdots & & \\ 0 & 0 & \cdots & \partial_n \end{pmatrix}$
L	Laplacian matrix $D - A$, positive and semidefinite [11]
Spec_L	Laplacian spectrum $\{0, \lambda_1, \lambda_2, \ldots, \lambda_N\}$, with $\lambda_1 \leq \ldots \leq \lambda_N$ eigenvalues
ω_i	Vibrational frequencies $\sqrt{\lambda_i}$, solution of the ODE system $\ddot{x}_i + \sum_{j=1}^N A_{ij} (x_i - x_j) = 0$ [27]
ρ	Spectral density as sum of Lorentz distributions $\rho(\omega, \gamma) = K \sum_{i=1}^{N-1} \frac{\gamma}{(\omega-\omega_i)^2+\gamma^2}$
K	Normalisation constant defined by $\int_0^\infty \rho(\omega, \gamma) \mathrm{d}\omega = 1$
γ	Half-width at half-maximum
$\overline{\gamma}$	Unique solution of $\int_0^\infty \left[\rho_{\mathscr{E}_N}(\omega, \gamma) - \rho_{\mathscr{F}_N}(\omega, \gamma)\right]^2 \mathrm{d}\omega = 1$ [31]

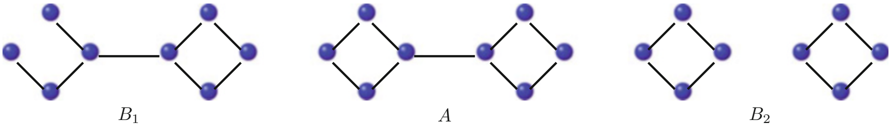


Fig. 2 Link equivalence for Hamming metric: $H(A, B_1) = H(A, B_2)$ although B_1 is connected while B_2 consists of two connected components

$$\ddot{x}_i + \sum_{j=1}^{N} A_{ij}(x_i - x_j) = 0 \quad \text{for } i = 0, \cdots, N-1 .$$

The vibrational frequencies ω_i for this network model are given by the square root of the eigenvalues of the Laplacian matrix of the network: $\lambda_i = \omega_i^2$, with $\lambda_0 = \omega_0 = 0$. The spectral density for a graph as the sum of Lorentz distributions is defined as

$$\rho(\omega,\gamma) = K\sum_{i=1}^{N-1} \frac{\gamma}{(\omega-\omega_i)^2+\gamma^2}\ ,$$

where γ is the common width and K is the normalisation constant defined by the condition $\displaystyle\int_0^\infty \rho(\omega,\gamma)\mathrm{d}\omega = 1$, and thus

$$K = \frac{1}{\displaystyle\gamma\sum_{i=1}^{N-1}\int_0^\infty \frac{\mathrm{d}\omega}{(\omega-\omega_i)^2+\gamma^2}}\ .$$

The scale parameter γ specifies the half-width at half-maximum, which is equal to half the interquartile range. Then the spectral distance ϵ_γ between two graphs $\mathcal{N}_1$ and $\mathcal{N}_2$ on N nodes with densities $\rho_{\mathcal{N}_1}(\omega,\gamma)$ and $\rho_{\mathcal{N}_2}(\omega,\gamma)$ can then be defined as

$$\epsilon_\gamma(\mathcal{N}_1,\mathcal{N}_2) = \sqrt{\int_0^\infty \left[\rho_{\mathcal{N}_1}(\omega,\gamma)-\rho_{\mathcal{N}_2}(\omega,\gamma)\right]^2\mathrm{d}\omega}\ .$$

The highest value of ϵ_γ is reached, for each N, when evaluating the distance between $\mathcal{E}_N$ and $\mathcal{F}_N$. Denote then by $\overline{\gamma}$ the unique solution of

$$\epsilon_\gamma(\mathcal{E}_N,\mathcal{F}_N) = 1\ .$$

Thus, by definition, IM too ranges between 0 and 1, where

$$\mathrm{IM} = 0 \text{ for } \mathrm{spec}(\mathrm{L}^{(1)}) = \mathrm{spec}(\mathrm{L}^{(2)}) \text{ and } \mathrm{IM} = 1 \text{ for } \{\mathcal{N}_1,\mathcal{N}_2\} = \{\mathcal{E}_{\mathrm{N}},\mathcal{F}_{\mathrm{N}}\}.$$

In fact, being a spectral measure, IM cannot distinguish isospectral (non-isomorphic) networks.

To overcome the drawbacks of both H and IM, we define their normalised cartesian product, the Hamming–Ipsen–Mikhailov distance:

$$\mathrm{HIM}_\xi(\mathcal{N}_1,\mathcal{N}_2) = \frac{1}{\sqrt{1+\xi}}\sqrt{\mathrm{H}^2(\mathcal{N}_1,\mathcal{N}_2)+\xi\cdot\mathrm{IM}^2(\mathcal{N}_1,\mathcal{N}_2)},$$

for $\xi \in [0,+\infty)$.

When ξ is not close to the bounds $\{0, +\infty\}$ (and one of the factors becomes dominant), the impact of ξ is minimal, and in general more relevant when HIM_ξ is used as a kernel [21]. Hereafter $\xi = 1$ will be assumed, and the subscript ξ omitted. Again, HIM is bounded between 0 and 1, with

$$\mathrm{HIM} = 0 \text{ for } \mathrm{A}^{(1)} = \mathrm{A}^{(2)} \text{ and } \mathrm{HIM} = 1 \text{ for } \{\mathscr{N}_1, \mathscr{N}_2\} = \{\mathscr{E}_\mathrm{N}, \mathscr{F}_\mathrm{N}\}.$$

The HIM distance can be naturally extended to directed networks, by transforming it into an undirected bipartite graph through the procedure shown in [36].

The HIM distance naturally induces a kernel via Gaussian (Radial Basis Function) map [9, 13] to be used standalone or in a Multi-Kernel Learning framework to increase performance and enhance interpretability [33]:

$$K(\mathscr{N}_1, \mathscr{N}_2) = e^{-\gamma \cdot \mathrm{HIM}^2_\xi(\mathscr{N}_1, \mathscr{N}_2)} \,,$$

for a positive real number γ.

Although the HIM kernel is not positively defined in general for all $\gamma \in \mathbb{R}^+_0$, by results in [44] it can be used in Support Vector Machines or other algorithms whenever K is positively defined for the given training data, which is the case for all the examples shown in what follows. In general, the range of suitable values for γ can be computed by imposing positiveness to all eigenvalues of the matrix $e^{-\gamma \cdot \mathrm{HIM}^2_\xi(x_i, x_j)}$ for x_i, x_j in the training set.

3 Application to -omic Studies

3.1 *The UKBEC Dataset*

The United Kingdom Brain Expression Consortium (UKBEC) hybridised on a Affymetrix Human Exon 1.0 ST Array (transcript version) 1213 human brain samples from ten diverse regions. Samples originated from 134 neurologically and neuropathologically normal individuals and were used in three studies aimed at better understanding gene expression differences [42, 46, 47]. Data details about sample stratification according to sex and tissue location are listed in Table 2(a). Here, this dataset[1] is used to build the absolute Pearson coexpression networks corresponding to different region/gender/age group defined on the 50

[1] Available as GEO46706 at http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE46706.

Table 2 Sample size of the UKBEC human brain dataset stratified by gender and tissue location (a) and by gender and age group (b)

(a)								(b)					
Region	Abbr.	M	F	Region	Abbr.	M	F	Age	M	F	Age	M	F
Cerebellar cortex	CB	95	35	Frontal cortex	FCX	93	34	< 32	86	39	58–62	117	20
Hippo campus	HC	92	30	Medulla	Med	88	31	32–44	130	19	62–68	72	29
Occipital cortex	OCC	94	35	Putamen	PUT	96	33	44–48	74	24	68–76	82	39
Substantia nigra	SN	73	28	Temporal cortex	TCX	86	33	48–53	109	27	76–83	66	56
Thalamus	Thal	91	33	White matter	WM	97	34	53–58	101	20	≥ 83	68	53

Region: the tissue location. Abbr.: abbreviation as in Fig. 3, M: number of samples from male individuals, F: number of samples from female individuals. $a \sim b$ means $a < x \leq b$

genes belonging to the BRAIN_DEVELOPMENT (GO:0007420, GSEA M7203) pathway,[2] corresponding to 1012 probes on the Affymetrix Human Exon 1.0 ST Array platform.[3] In detail, each network has 1012 nodes (one for each probe) and the weight of a link between two nodes is the absolute value of the Pearson correlation between the vectors collecting the expression levels of the corresponding probes for the samples belonging to the considered region/gender/age group.

First, we consider planar projections of all the mutual HIM distances between networks with shared nodes based on the metric multidimensional scaling (mMDS) [14, 37] in Fig. 3. The mMDS plot shows the mutual HIM distances with networks stratified for both sex and tissue location. Citing the authors, the study in [46] ‘provides unequivocal evidence that sex-biased gene expression in the adult human brain is widespread in terms of both the number of genes and range of brain regions involved’. In our analysis, the result is numerically confirmed by the major effect emerging at the gene coexpression level (Fig. 3): male and female networks can be linearly separated in the mMDS space, with large HIM distances between both inter- and intragender tissue locations. In particular, intragender HIM distances among different tissue regions are larger for the female samples (range [0.112,0.232], median 0.146) than for the male (range [0.077,0.200], median 0.118), with statistical significance (t-test p-value $1.9 \cdot 10^{-4}$).

In Fig. 4, we show instead the mMDS projections for the mutual HIM distances of the coexpression networks built separately for male and female subjects, partitioned in ten age groups: the sample size for each network is listed in Table 2(b). While the plot for the females does not show any global pattern, for males the first five groups (age $<$ 58 y) have small mutual HIM distances and they result clustered together. On the other hand, the five older male groups are both mutually distant and distant from the younger subjects cluster, too. In this dataset, the small sample size in the female subgroup may be a relevant source of noise for some of the age groups,

[2] Available at http://software.broadinstitute.org/gsea/msigdb/cards/BRAIN_DEVELOPMENT.

[3] The platform has no probes for the 51st gene of the pathway, VCX3A.

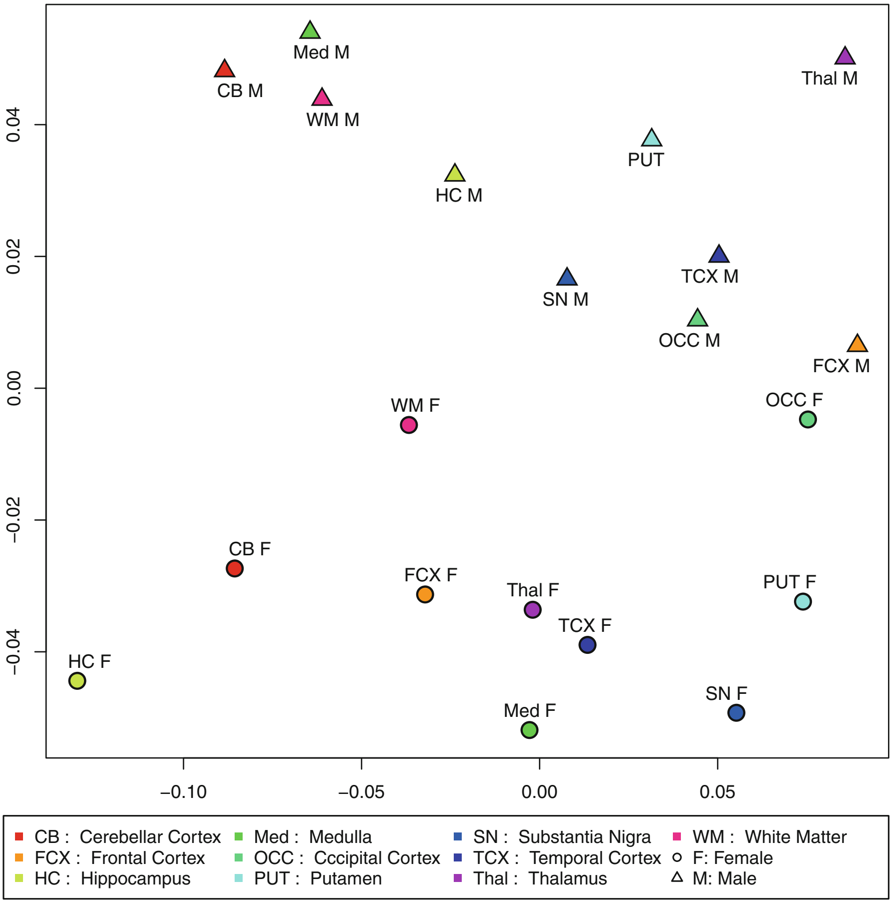


Fig. 3 Metric multidimensional scaling projection on two dimensions of all 190 mutual HIM distances between gene coexpression brain development networks stratified by gender and tissue locations

e.g., the 32–44. Our results are consistent with findings obtained with different data and methodology by Berchtold and colleagues in [8], suggesting the existence of a global pattern of gene expression change associated with brain aging, more evident from the sixth decade onward, with different evolutions between males and females, with larger variations in male subjects. Biologically, this is due to a wider global decrease in males in the catabolic and anabolic capacity with aging, mainly in genes linked to energy production and protein synthesis and transport [8].

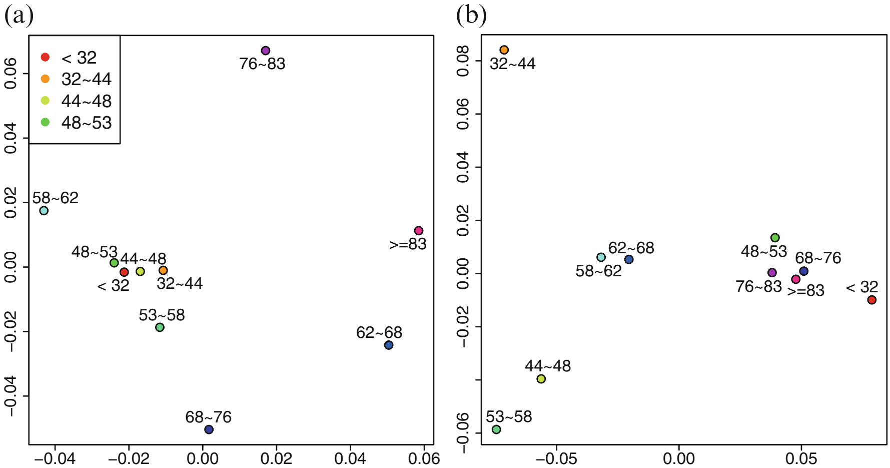


Fig. 4 Metric multidimensional scaling projection on two dimensions of all 45 mutual HIM distances between gene coexpression brain development networks stratified by age groups, separately for the male (**a**) and female (**b**) subjects. $a \sim b$ means $a < x \leq b$

3.2 *The D. melanogaster Development Dataset*

In [34], Kolar and colleagues applied the Keller algorithm to infer the gene regulatory networks of *Drosophila melanogaster* from a time series of gene expression data measured during its full life cycle, originally published in [2]. They followed the dynamics of 588 development genes along 66 time points spanning through four different stages (Embryonic—time points 1–30, Larval—t.p. 31–40, Pupal—t.p. 41–58, Adult—t.p. 59–66), constructing a time series of inferred networks N_i,[4] where a link between two nodes exists whenever the Keller algorithm detects a mutual inference between the corresponding genes at the given time point: in Fig. 5a we show four instances of the N_i networks, at different timing.

As a first step in the quantitative NetDA of this dataset, we measure the HIM distance between each N_i and the initial network N_1: the resulting distance time series is shown in Fig. 5b. The largest variations, both between consecutive terms and with respect to the initial network N_1, occur in the Embryonal stage (E). In particular, the HIM distance grows until time point 23; next networks get closer again to N_1, showing that the interactions of the selected 588 genes in the adult stage are more similar to the corresponding net of interaction in the Embryonal stage, rather than in the other two stages, consistently with the findings reported in the original reference [34]. Moreover, while the Hamming component ranges between 0 and 0.0223, the Ipsen–Mikhailov distance has 0.0851 as its maximum,

[4] Publicly available at http://cogito-b.ml.cmu.edu/keller/downloads.html.

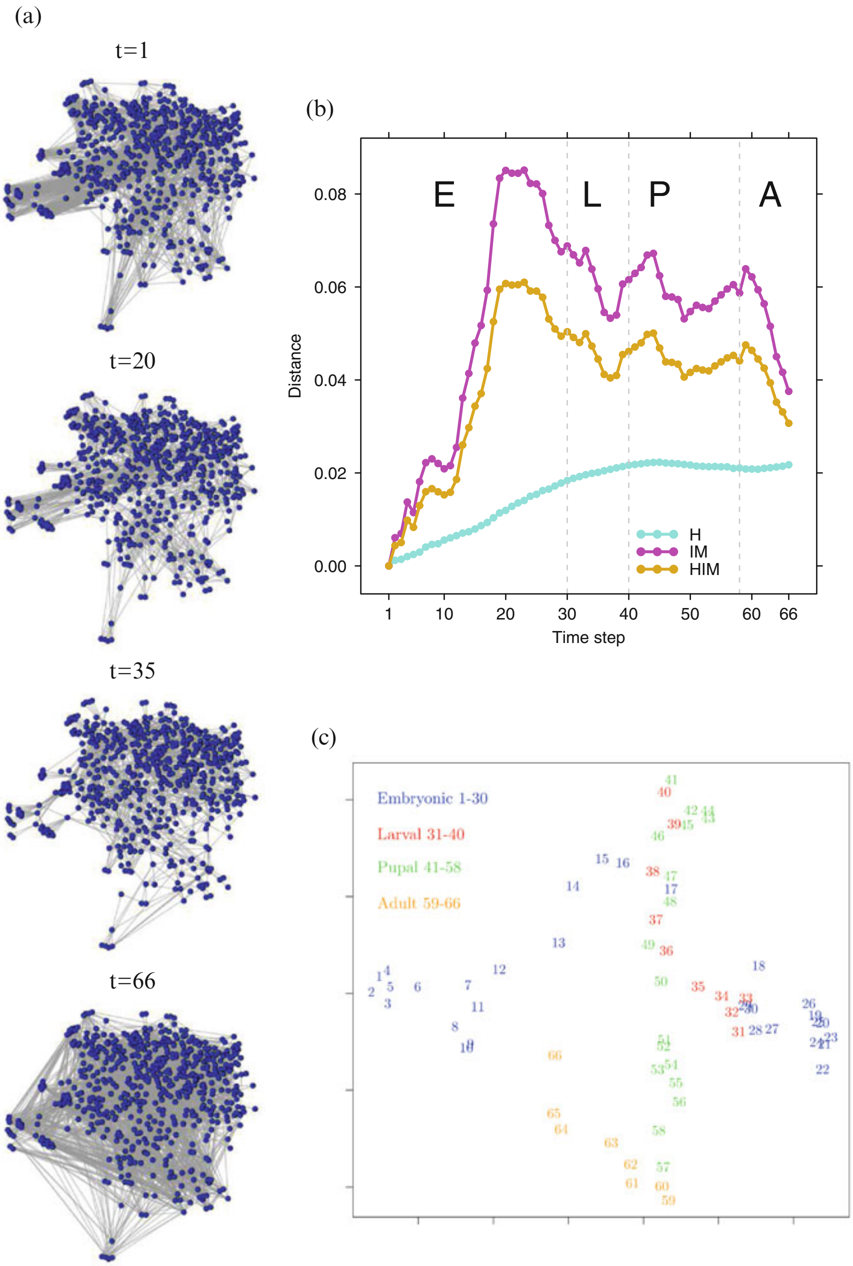


Fig. 5 *D. melanogaster* development network dataset. (**a**) Keller interaction network N_i for the *D. melanogaster* development genes at the time points $i = 1, 20, 35, 66$. (**b**) Evolution of H (*cyan*), IM (*magenta*) and HIM (*golden red*) distances network time series across 66 time points in the four stages Embryonic (E), Larval (L), Pupal (P) and Adult (A). (**c**) Metric multidimensional scaling planar projection of the mutual HIM distances between the 66 networks N_i, coloured according to the developmental stage Embryonic (*blue*), Larval (*red*), Pupal (*green*) and Adult (*orange*)

indicating an higher variability of the networks in terms of structure rather than matching links: in this case, in fact, the HIM distance is driven by the evolution of the IM component.

Then we computed all 2145 HIM distances $\text{HIM}(N_i, N_j)$, and we projected them on a 2D mMDS representation, shown in Fig. 5c. Interestingly, the networks for the Embryonal stage split into two clusters (before and after time points 17), and the Embryonal and Pupal stages are orthogonal in this representation.

Moreover, the Adult stage networks form a cluster well separated from the other nets, with the Larval stage graphs mixing with the Pupal and late Embryonal stages. Finally, a Support Vector Machine classifier with HIM kernel was developed with the *kernlab* package in R, with a five-fold cross validation with $\gamma = 10^3$ and $C = 1$. The classifier reached accuracy 0.97 in discriminating Embryonic and Adult networks from Larval and Pupal. Similarly, in the same setup, perfect separation is reached between Embryonic and Adult stages for all values of $\gamma > 10^3$.

4 Conclusion

The interest of the HIM metric is its global/local approach: by combining edit and spectral distance types, we overcome the drawbacks of the two distance components. The two presented applications in functional high-throughput -omics support the effectiveness of the approach. The strategy of a NetDA based on the HIM distance offers a reproducible method: the metric gives a completely quantitative assessment of the differences among networks (on shared nodes) as well as a scalar product for kernel learning machines.

Operatively, we provide an Open Source implementation of the HIM distance with the R package *nettools* available on CRAN and GitHub,[5] and in the web interface ReNette [19].[6] In particular, ReNette includes a complete pipeline for NetDA, integrating a comprehensive collection of tools for network inference, network comparison and network stability analysis [20] (a methodology for assessing the robustness of an inferred network w.r.t. data subsampling) through queue-based submission system and asynchronous task management. The software is already configured for usage on multicore workstations, on high performance computing clusters and on a cloud-based cluster, to deal with the extraction of the Laplacian spectrum, which represents the computational bottleneck of the algorithm.

[5] https://github.com/MPBA/nettools.git.

[6] http://renette.fbk.eu.

References

1. Angulo, M., Moreno, J., Barabási, A.L., Liu, Y.Y.: Fundamental limitations of network reconstruction (2015). arXiv:1508.03559
2. Arbeitman, M., Furlong, E., Imam, F., Johnson, E., Null, B., Baker, B., Krasnow, M., Scott, M., Davis, R., White, K.: Gene expression during the life cycle of *Drosophila melanogaster*. Science **297**(5590), 2270–2275. Erratum in Science **298**(5596), 1172 (2002)
3. Barabási, A.L.: The network takeover. Nat. Phys. **8**, 14–16 (2012)
4. Barabási, A.L.: Network science. Philos. Trans. R. Soc. A **371**(1987), 20120375 (2013)
5. Baralla, A., Mentzen, W., de la Fuente, A.: Inferring gene networks: dream or nightmare? Ann. N. Y. Acad. Sci. **1158**, 246–256 (2009)
6. Barla, A., Jurman, G., Visintainer, R., Squillario, M., Filosi, M., Riccadonna, S., Furlanello, C.: A machine learning pipeline for discriminant pathways identification. In: Biganzoli, E., Vellido, A., Ambrogi, F., Tagliaferri, R. (eds.) Computational Intelligence Methods for Bioinformatics and Biostatistics. Lecture Notes in Computer Science, vol. 7548, pp. 36–48. Springer, Berlin (2012)
7. Barla, A., Jurman, G., Visintainer, R., Squillario, M., Filosi, M., Riccadonna, S., Furlanello, C.: A Machine learning pipeline for discriminant pathways identification. In: Kasabov, N. (ed.) Springer Handbook of Bio-/Neuroinformatics, Chap. 53, p. 1200. Springer, Berlin (2013)
8. Berchtold, N., Cribbs, D., Coleman, P., Rogers, J., Head, E., Kim, R., Beach, T., Miller, C., Troncoso, J., Trojanowski, J., Zielke, H., Cotman, C.: Gene expression changes in the course of normal brain aging are sexually dimorphic. Proc. Natl. Acad. Sci. U. S. A. **105**(40), 15605–15610 (2008)
9. Bolla, M.: Spectral Clustering and Biclustering: Learning Large Graphs and Contingency Tables. Wiley, New York (2013)
10. Chuang, H.Y., Lee, E., Liu, Y.T., Lee, D., Ideker, T.: Network-based classification of breast cancer metastasis. Mol. Syst. Biol. **3**, 140 (2007)
11. Chung, F.: Spectral Graph Theory. CBMS Regional Conference Series in Mathematics, vol. 92. American Mathematical Society, Philadelphia (1997)
12. Cootes, A., Muggleton, S., Sternberg, M.: The identification of similarities between biological networks: application to the metabolome and interactome. J. Mol. Biol. **369**, 1126–1139 (2007)
13. Cortes, C., Haffner, P., Mohri, M.: Positive definite rational kernels. In: Learning Theory and Kernel Machines. Proceedings of COLT 2003. Lecture Notes on Computer Science, vol. 2777, pp. 41–56. Springer, Berlin (2003)
14. Cox, T., Cox, M.: Multidimensional Scaling. Chapman and Hall, Boca Raton (2001)
15. Csermely, P., Korcsmáros, T., Kiss, H., London, G., Nussinov, R.: Structure and dynamics of biological networks: a novel paradigm of drug discovery. A comprehensive review. Pharmacol. Ther. **138**, 333–408 (2013)
16. de la Fuente, A.: From 'differential expression' to 'differential networking' - identification of dysfunctional regulatory networks in diseases. Trends Genet. **26**(7), 326–333 (2010)
17. Dehmer, M., Mowshowitz, A.: The discrimination power of structural superindices. PLoS ONE **8**(7), e70551 (2013)
18. Dougherty, E.: Validation of gene regulatory networks: scientific and inferential. Brief. Bioinform. **12**(3), 245–252 (2010)
19. Filosi, M., Droghetti, S., Arbitrio, E., Visintainer, R., Riccadonna, S., Jurman, G., Furlanello, C.: ReNette: a web-infrastructure for reproducible network analysis (2014). bioRxiv-doi:10.1101/008433
20. Filosi, M., Visintainer, R., Riccadonna, S., Jurman, G., Furlanello, C.: Stability indicators in network reconstruction. PLoS ONE **9**(2), e89815 (2014)
21. Furlanello, T., Cristoforetti, M., Furlanello, C., Jurman, G.: Sparse predictive structure of deconvolved functional brain networks. High-Dimensional Statistical Inference in the Brain, NIPS 2013 Workshop (2013). arXiv:1310.6547[q-bio.NC]

22. Gill, R., Datta, S., Datta, S.: A statistical framework for differential network analysis from microarray data. BMC Bioinf. **11**(1), 1–10 (2010)
23. Ha, M., Baladandayuthapani, V., Do, K.A.: DINGO: differential network analysis in genomics. Bioinformatics **31**(21), 3413–3420 (2015)
24. Hamming, R.: Error detecting and error correcting codes. Bell Syst. Tech. J. **29**(2), 147–160 (1950)
25. Ideker, T., Krogan, N.: Differential network biology. Mol. Syst. Biol. **8**, 565 (2012)
26. Ioannidis, J., Allison, D., Ball, C., Coulibaly, I., Cui, X., Culhane, A.C., Falchi, M., Furlanello, C., Game, L., Jurman, G., Mehta, T., Mangion, J., Nitzberg, M., Page, G., Petretto, E., van Noort, V.: Repeatability of published microarray gene expression analyses. Nat. Genet. **41**(2), 499–505 (2009)
27. Ipsen, M., Mikhailov, A.: Evolutionary reconstruction of networks. Phys. Rev. E **66**, 046109 (2002). Erratum in Phys. Rev. E **67**, 039901 (2003)
28. Iwayama, K., Hirata, Y., Takahashi, K., Watanabe, K., Aihara, K., Suzuki, H.: Characterizing global evolutions of complex systems via intermediate network representations. Sci. Rep. **2**, 423 (2012)
29. Jurman, G.: Metric projections for dynamic multiplex networks (2016). arXiv:1601.01940
30. Jurman, G., Visintainer, R., Furlanello, C.: An introduction to spectral distances in networks. In: Apolloni, B., Bassis, S. (eds.) Proceedings of WIRN10, Frontiers in Artificial Intelligence and Applications, vol. 226, pp. 227–234. IOS Press, Amsterdam (2011)
31. Jurman, G., Visintainer, R., Riccadonna, S., Filosi, M., Furlanello, C.: The HIM glocal metric and kernel for network comparison and classification (2014). arXiv:1201.2931v3
32. Jurman, G., Visintainer, R., Filosi, M., Riccadonna, S., Furlanello, C.: The HIM glocal metric and kernel for network comparison and classification. In: Proceedings IEEE DSAA'15, vol. 36678, pp. 1–10. IEEE, New York (2015)
33. Kloft, M., Brefeld, U., Sonnenburg, S., Zien, A.: ℓ_p-norm multiple kernel learning. J. Mach. Learn. Res. **12**, 953–997 (2011)
34. Kolar, M., Song, L., Ahmed, A., Xing, E.: Estimating time-varying networks. Ann. Appl. Stat. **4**(1), 94–123 (2010)
35. Koutra, D., Vogelstein, J., Faloutsos, C.: DELTACON: a principled massive-graph similarity function. In: Proceedings of the 13th SIAM International Conference on Data Mining (SDM), pp. 162–170. SIAM, New York (2013)
36. Liu, Y.Y., Slotine, J.J., Barabási, A.L.: Controllability of complex networks. Nature **473**(7346), 167–173 (2011)
37. Mardia, K.: Some properties of classical multidimensional scaling. Commun. Stat. Theory Meth. **A7**, 1233–1241 (1978)
38. Meyer, P., Alexopoulos, L., Bonk, T., Califano, A., Cho, C., de la Fuente, A., de Graaf, D., Hartemink, A., Hoeng, J., Ivanov, N., Koeppl, H., Linding, R., Marbach, D., Norel, R., Peitsch, M., Rice, J., Royyuru, A., Schacherer, F., Sprengel, J., Stolle, K., Vitkup, D., Stolovitzky, G.: Verification of systems biology research in the age of collaborative competition. Nat. Biotechnol. **29**(9), 811–815 (2011)
39. Mina, M., Boldrini, R., Citti, A., Romania, P., D'Alicandro, V., De ioris, M., Castellano, A., Furlanello, C., Locatelli, F., Fruci, D.: Tumor-infiltrating T lymphocytes improve clinical outcome of therapy-resistant neuroblastoma. Oncoimmunology **4**(9), e1019981 (2015)
40. Morris, M., Handcock, M., Hunter, D.: Specification of exponential-family random graph models: terms and computational aspects. J. Stat. Softw. **24**(4), 1–24 (2008)
41. Pavlopoulos, G., Secrier, M., Moschopoulos, C., Soldatos, T., Kossida, S., Aerts, J., Schneider, R., Bagos, P.: Using graph theory to analyze biological networks. BioData Min. **4**(1), 10 (2011)
42. Ramasamyi, A., Trabzuni, D., Guelfi, S., Varghese, V., Smith, C., Walker, R., De, T., United Kingdom Brain Expression Consortium (UKBEC), North American Brain Expression Consortium, Coin, L., de Silva, R., Cookson, M., Singleton, A., Hardy, J., Ryten, M., Weale, M.: Genetic variability in the regulation of gene expression in ten regions of the human brain. Nat. Neurosci. **17**(10), 1418–1428 (2014)

43. Ruan, D., Young, A., Montana, G.: Differential analysis of biological networks. BMC Bioinf. **16**, 327 (2015)
44. Schölkopf, B.: Support Vector Learning. Oldenbourg, Munchen (1997)
45. Sharan, R., Ideker, T.: Modeling cellular machinery through biological network comparison. Nat. Biotechnol. **24**(4), 427–433 (2006)
46. Trabzuni, D., Ramasamy, A., Imran, S., Walker, R., Smith, C., Weale, M., Hardy, J., Ryten, M., North American brain expression consortium. Widespread sex differences in gene expression and splicing in the adult human brain. Nat. Commun. **4**, 2771 (2013)
47. Trabzuni, D.: United Kingdom Brain Expression Consortium (UKBEC), Thomson, P.: Analysis of gene expression data using a linear mixed model/finite mixture model approach: application to regional differences in the human brain. Bioinformatics **30**(11), 1555–1561 (2014)
48. Tun, K., Dhar, P., Palumbo, M., Giuliani, A.: Metabolic pathways variability and sequence/networks comparisons. BMC Bioinf. **7**(1), 24 (2006)
49. Xiao, Y., Dong, H., Wu, W., Xiong, M., Wang, W., Shi, B.: Structure-based graph distance measures of high degree of precision. Pattern Recogn. **41**(12), 3547–3561 (2008)
50. Yang, B., Zhang, J., Yin, Y., Zhang, Y.: Network-based inference framework for identifying cancer genes from gene expression data. BioMed. Res. Int. **2013**, 12pp. (2013). Article ID 401649
51. Yoon, B.J., Qian, X., Sahraeian, S.: Comparative analysis of biological networks. IEEE Signal Process. Mag. **29**(1), 22–34 (2012)
52. Zandoná, A., Chierici, M., Jurman, G., Furlanello, C., Cucchiara, S., Del Chierico, F., Putignani, L.: A metagenomic pipeline integrating predictive profiling methods and complex networks for the analysis of NGS microbiome data. NIPS Workshop - Machine Learning in Computational Biology (2014)

Structural vs Practical Identifiability of Nonlinear Differential Equation Models in Systems Biology

Maria Pia Saccomani and Karl Thomaseth

Abstract This paper reappraises two different viewpoints adopted for testing identifiability of nonlinear differential equation models. The aim is to take advantage through their joint use of the complementary information provided. The common objective is to assess whether model parameters can be estimated from specific input/output (I/O) experiments. The *structural identifiability* analysis investigates whether unknown model parameters can be identified uniquely, at all, with a particular I/O configuration. This is investigated using differential algebra, e.g., as implemented in the software DAISY (Differential Algebra for Identifiability of SYstems). In contrast, *practical identifiability* analysis is a data-based approach to assess the precision of parameter estimates obtainable from experimental data. It is based on simulated model outputs and their sensitivities with respect to parameters. The relevant novelty of using both methodologies together is that structural identifiability analysis allows a clearer understanding of the practical identifiability results. This result is shown in the identifiability analysis of a much quoted biological model describing the erythropoietin(Epo)-induced activation of the JAK-STAT signaling pathway, which is known to play a role in the regulation of cell proliferation, differentiation, chemotaxis, and apoptosis and is important for hematopoiesis, and immune development. This study shows that some results on practical identifiability tests can be proven in an analytical way by a differential algebra test and that this test can provide additional information helpful for the experiment design.

Keywords Biological systems • Identifiability • Parameter estimation

M.P. Saccomani (✉)
Department of Information Engineering, University of Padova, Via G Gradenigo 6a, 35131 Padova, Italy
e-mail: pia@dei.unipd.it

K. Thomaseth
Institute of Electronics, Computer and Telecommunication Engineering (IEIIT-CNR) c/o DEI, Via G Gradenigo 6b, 35131 Padova, Italy
e-mail: karl.thomaseth@ieiit.cnr.it

A. Rogato et al. (eds.), *Dynamics of Mathematical Models in Biology*,
DOI 10.1007/978-3-319-45723-9_3

1 Introduction

Mathematical modeling has become ubiquitous in quantitative molecular biology and biotechnology with applications ranging from metabolic engineering to cancer therapy. A large number of publications present mathematical models to investigate complex, dynamic, nonlinear interaction mechanisms in cellular processes like signal transduction pathways and metabolic networks. Typically these mechanisms are modeled according to physicochemical laws, such as mass or molar balance, and mathematical equations are introduced to describe the rates of reactions or transformations between different molecules. Different mathematical frameworks are available depending on the alternative approximations adopted to represent the biological system under study. Ordinary nonlinear differential equations (ODE) involving parameters such as reaction rates are commonly used. For example, the Michaelis–Menten equation is frequently used to describe the internal law governing the biochemistry of a system, assuming that diffusion is fast compared to reaction rates. The unknown parameters of ODE models contain key information that can be gathered in general only indirectly, as it is usually not possible to measure directly the dynamics of every portion of the system. The recovery of parameter values can then be only approached as an estimation problem of internal parameters to fit external input/output (I/O) measurements, where *input* represents external perturbations of a system and *output* the (measured) system response to this input.

In this context, the first relevant question is whether the model parameters can be (uniquely) determined, at least for suitable input functions, assuming that all observable variables are error free. This is a mathematical property of the model called a priori or *structural identifiability* [1, 4, 5, 8, 9] that can, and should, in principle be checked before collecting experimental data, or at the latest before trying to fit the model to data.

Concerning uniqueness, it is important to distinguish between *global*, i.e., one parameter solution, and *local*, i.e., a finite number of parameter solutions, identifiability [11]. It has been observed by several authors, see, e.g., [5, 10], that the weaker property of local identifiability about a parameter value may often be sufficient in practice. However, in biomedical applications that rely upon model parameter values to discriminate between classes, e.g., healthy versus pathological states, global identifiability should be a necessary request [6]. If the postulated model is neither globally nor locally identifiable, the parameter estimates that could still be obtained by some numerical optimization algorithms could be totally unreliable.

From a theoretical point of view only proven structural identifiability can guarantee, either globally or locally, the uniqueness of the parameter solution, which is a prerequisite for the parameter estimation problem to be well-posed. Obviously, although necessary, structural identifiability is not sufficient to guarantee an accurate identification of the model parameters from real I/O data.

Conversely, many currently studied models in systems biology are rather large networks containing many states and parameters such that checking structural identifiability may become prohibitively complex. Situations of this kind can be approached by semi-empirical techniques, which are essentially based on simulations and on the study of the level curves of a cost function which, once minimized, should yield parameter estimates that are at least unique around an optimal parameter value. This is called *practical identifiability* in the literature [10, 12]. Checking practical identifiability can be done on more realistic models, which may explicitly involve noise in the measurements. It should be kept in mind, however, that, since they are data-based (or simulation-based), practical identifiability methods cannot provide a mathematically rigorous answer to the uniqueness problem.

In the following sections we provide first the theoretical background for structural and practical identifiability analysis and then we use a typical benchmark biological model to show how the joint use of the two approaches appears the most promising because of the complementary information provided.

2 Structural vs Practical Identifiability

Consider a nonlinear dynamic system described in state-space form as

$$\dot{\mathbf{x}}(t) = \mathbf{f}(\mathbf{x}(t), \mathbf{u}(t), \boldsymbol{\theta}) \tag{1}$$

$$\mathbf{y}(t) = \mathbf{h}(\mathbf{x}(t), \mathbf{u}(t), \boldsymbol{\theta}) \tag{2}$$

with state $\mathbf{x}(t) \in \mathbb{R}^n$, input $\mathbf{u}(t) \in \mathbb{R}^q$ ranging on some vector space of differentiable functions, output $\mathbf{y}(t) \in \mathbb{R}^m$, and the constant unknown parameter vector $\boldsymbol{\theta}$ belonging to some open subset $\Theta \subseteq \mathbb{R}^p$. Whenever initial conditions are specified, the relevant equation $\mathbf{x}(0) = \mathbf{x}_0$ is added to the system. The functions $\mathbf{f}$ and $\mathbf{h}$ are vectors of *rational functions* in $\mathbf{x}$.

2.1 Structural Identifiability Analysis

We adopt the definition of identifiability proposed in [13]. Let $\mathbf{y} = \psi_{\mathbf{x}_0}(\boldsymbol{\theta}, \mathbf{u})$ be the I/O map of the system (1), (2) started at the initial state $\mathbf{x}_0$. The I/O map allows to calculate the output function $\mathbf{y}$ in terms of the input function $\mathbf{u}$, being $\mathbf{u}$ and $\mathbf{y}$ both known from the I/O configuration of system (1), (2). In particular, when starting with a system described by a state-space representation, as (1), (2), it is possible to calculate the solution $\mathbf{x}$ of the differential equations (1) and to define the I/O map in terms of the unknown parameters $\boldsymbol{\theta}$ only. The identifiability definitions are thus given on the basis of this I/O map.

Definition 1. The system (1), (2) is (a priori) globally (or uniquely) identifiable from I/O data if, for at least a generic set of points $\boldsymbol{\theta}^* \in \Theta$, there exists (at least) one input function of time, $\mathbf{u}(t)$, such that the equation

$$\psi_{\mathbf{x}_0}(\boldsymbol{\theta}, \mathbf{u}) = \psi_{\mathbf{x}_0}(\boldsymbol{\theta}^*, \mathbf{u}) \tag{3}$$

has only one solution $\boldsymbol{\theta} = \boldsymbol{\theta}^*$ for almost all initial states $\mathbf{x}_0 \in X \subseteq \mathbb{R}^n$.

Structural identifiability analysis addresses whether the system model (1), (2) is globally identifiable and, if not, which subsets of parameters are either globally (only one solution), locally (a finite number of solutions), or nonidentifiable (an infinite number of solutions).

For testing *structural identifiability*, different methods have been proposed to check models described by linear and nonlinear differential equations [4, 5, 7–9]. In this paper we focus on a structural identifiability test based on differential algebra and on the software DAISY (Differential Algebra for Identifiability of SYstems) [2]. The reader is referred to [1, 13] for a detailed documentation of the theory behind DAISY.

Briefly, this algorithm permits to find the *I/O relations* as a set of polynomial differential equations involving only the variables $(\mathbf{u}(t), \mathbf{y}(t))$ and their first and higher order time derivatives. The coefficients of these I/O relations are polynomials of the unknown parameter $\boldsymbol{\theta}$ that form the *exhaustive summary* of the model. Identifiability is tested by checking injectivity of the exhaustive summary on parameter $\boldsymbol{\theta}$. This is achieved by applying Buchberger's algorithm [3] to compute a Gröbner basis of the system. In particular, if (3) has one and only one solution $\boldsymbol{\theta}$, the Gröbner basis is of the following form:

$$\mathbf{G}(\boldsymbol{\theta}, \boldsymbol{\theta}^*) = \left[\{\boldsymbol{\theta}_1 - \boldsymbol{\theta}_1^*, \ldots, \boldsymbol{\theta}_p - \boldsymbol{\theta}_p^*\}\right] \tag{4}$$

showing that the model (1), (2) is globally identifiable. In any case, the Gröbner basis provides the unique parameterization of the model and allows to count the number of solutions, i.e., the number of distinct values of the unknown parameter $\boldsymbol{\theta}$ that solve the system of equations implied by (3).

In contrast, (3) has infinite many solutions if the basis $\mathbf{G}(\boldsymbol{\theta}, \boldsymbol{\theta}^*)$ has less components than the number of estimated parameters. This occurs either if one or more parameters disappear from the Gröbner basis or if the parameters satisfy a number of algebraic relations less than p. This means that the I/O map will be identical, thus non-distinguishable, for all values of the *hidden* parameters, and/or for specific, analytically known, *combinations* of parameters.

2.2 Practical Identifiability Analysis

Practical or data-based identifiability aims at assessing the (statistical) confidence with which model parameters are estimated from noisy measurements, typically to judge the reliability of results and to support consequent interpretations. For this purpose the model output equations (2) are normally revised by adding measurement noise, such as

$$\mathbf{y}(t) = \mathbf{h}(\mathbf{x}(t), \mathbf{u}(t), \boldsymbol{\theta}) + \boldsymbol{\varepsilon}(t) := \hat{\mathbf{y}}(t, \boldsymbol{\theta}) + \boldsymbol{\varepsilon}(t) \tag{5}$$

Prior assumptions on the statistical properties of the random noise $\boldsymbol{\varepsilon}(t) \in \mathbb{R}^m$, or actual knowledge of prediction error residuals after model fit to data, greatly influence calculations of confidence intervals. These latter are most often obtained using asymptotic results with local quadratic approximation of estimation criteria, e.g., log-likelihood, around point estimates or nominal parameter values, and first and second order statistics of measurement noise. While posterior reconstruction of confidence intervals of model parameter estimates, e.g., obtained by the *profile likelihood approach*, is more accurate than quadratic approximations, these latter are applicable even in an a priori setting and are often fully equivalent in practice, including as concerns their ability to detect structurally nonidentifiable parameters characterized by a completely flat profile likelihood function [10].

Assuming that a finite set of $N > m$ input–output measurements is available, the weighted sum of squared *prediction errors* is

$$V_N(\boldsymbol{\theta}) := \frac{1}{2} \sum_{k=1}^{N} [\mathbf{y}(t_k) - \hat{\mathbf{y}}(t_k, \boldsymbol{\theta})]^\top \mathbf{Q}_k, [\mathbf{y}(t_k) - \hat{\mathbf{y}}(t_k, \boldsymbol{\theta})] \tag{6}$$

where $\mathbf{Q}_k$ are positive semidefinite weights usually taken as the inverse of measurement noise variance, but without loss of generality, assumed in the following equal to the identity matrix, $\mathbf{Q}_k = \mathbf{I}$. Finally, with parameter estimates obtained as

$$\hat{\boldsymbol{\theta}} = \arg \min_{\boldsymbol{\theta}} V_N(\boldsymbol{\theta}) \ . \tag{7}$$

The model (1), (2), or parameter $\boldsymbol{\theta}$, can be defined *practically identifiable* if the minimum of $V_N(\boldsymbol{\theta})$ is well characterized in terms of necessary and sufficient conditions for a local minimum, i.e., a vanishing gradient: $\nabla_{\boldsymbol{\theta}} V_N(\hat{\boldsymbol{\theta}}) = 0$, and convexity in the neighborhood of $\hat{\boldsymbol{\theta}}$, i.e., with a positive definite Hessian matrix: $\nabla^2_{\boldsymbol{\theta}^2} V_N(\hat{\boldsymbol{\theta}}) > 0$.

Straightforward calculations under the simplifying a priori assumption of expected zero measurement noise yield the simplified Hessian matrix:

$$\nabla^2_{\boldsymbol{\theta}} V_N(\hat{\boldsymbol{\theta}}) = \sum_{k=1}^{N} \sum_{j=1}^{m} \nabla_{\boldsymbol{\theta}} \mathbf{y}_j(t_k, \hat{\boldsymbol{\theta}}) \nabla^\top_{\boldsymbol{\theta}} \mathbf{y}_j(t_k, \hat{\boldsymbol{\theta}}) = \mathbf{S}(\boldsymbol{\theta})^T \mathbf{S}(\boldsymbol{\theta}) \ , \tag{8}$$

where $\nabla_{\boldsymbol{\theta}} \mathbf{y}_j(t_k, \boldsymbol{\theta}) \in \mathbb{R}^{p \times 1}$ is the sensitivity with respect to all estimated parameters of the j-th output component measured at time t_k; $\mathbf{S}(\boldsymbol{\theta}) \in \mathbb{R}^{m \cdot N \times p}$, with $\mathbf{S}(\boldsymbol{\theta})^T = [\mathbf{S}_1(\boldsymbol{\theta})^T, \mathbf{S}_2(\boldsymbol{\theta})^T, \ldots, \mathbf{S}_m(\boldsymbol{\theta})^T]$ is the sensitivity matrix formed by all individual sensitivity matrices of measured model outputs, defined as $\mathbf{S}_j(\boldsymbol{\theta})^T = \left[\nabla_{\boldsymbol{\theta}} \mathbf{y}_j(t_1, \boldsymbol{\theta}), \ldots, \nabla_{\boldsymbol{\theta}} \mathbf{y}_j(t_N, \boldsymbol{\theta})\right]$.

It is reminded that positive *semi*-definite matrices, in contrast to positive definite ones, can be singular with one, or more, zero eigenvalues. Their corresponding eigenvectors provide, at the minimum, directions along which the cost function remains constant. This is the prerequisite for nonidentifiability. A more exhaustive statistical interpretation of $\nabla^2_{\boldsymbol{\theta}} V_N(\hat{\boldsymbol{\theta}})$, e.g., its relationship with the Fisher information matrix and the covariance matrix of parameter estimates, goes beyond the scope of this paper but can be found in standard textbooks, e.g., [15].

Parameter estimation based on prediction error minimization is therefore a well-posed problem that leads to (local) minima of the cost function (6) if the sensitivity matrix $\mathbf{S}(\boldsymbol{\theta})$, calculated at some point $\boldsymbol{\theta}$, has full rank. This must hold almost everywhere in the admissible parameter space.

Finally, the most dependable approach to assess the rank of a matrix is based on Singular Value Decomposition which provides the following factorization:

$$\mathbf{S}(\boldsymbol{\theta}) = \mathbf{U}\Sigma\mathbf{V}^T \tag{9}$$

where $\mathbf{U} \in \mathbb{R}^{m \cdot N \times m \cdot N}$ and $\mathbf{V} \in \mathbb{R}^{p \times p}$ are the orthonormal eigenvector matrices of $\mathbf{S}(\boldsymbol{\theta})\mathbf{S}(\boldsymbol{\theta})^T$ and $\mathbf{S}(\boldsymbol{\theta})^T\mathbf{S}(\boldsymbol{\theta})$, respectively, and $\Sigma \in \mathbb{R}^{m \cdot N \times p}$ is diagonal (referring to the top $p \times p$ submatrix) with sorted *singular values* $\sigma_1 \geq \sigma_2 \geq, \ldots \geq \sigma_p \geq 0$, which are also the square roots of the eigenvalues of the positive semidefinite matrix $\mathbf{S}(\boldsymbol{\theta})^T\mathbf{S}(\boldsymbol{\theta})$. The theoretical (vs practical) rank of $\mathbf{S}(\boldsymbol{\theta})$ is defined as the smallest $r \leq p$ at which $\sigma_{r+1} = 0$ (vs $\bar{\sigma} > \sigma_{r+1}$, with $\bar{\sigma}$ being a user-defined threshold) [16].

3 A Model of JAK-STAT Signaling Pathway

In this section we consider a dynamic model published in several journals [10, 14]. The aim of the model is to investigate the Epo-induced activation of the JAK-STAT signaling pathway that primarily consists of the cytoplasmic tyrosine kinase JAK and the latent transcription factor STAT. This pathway is known to play a role in the regulation of cell proliferation, differentiation, chemotaxis, and apoptosis and is important for hematopoiesis and immune development. The biochemical reactions of the JAK-STAT pathway are described by the following nonlinear ODE system:

$$\begin{cases} \dot{x}_1(t) = -k_1\, u(t)\, x_1(t) + 2\, k_4\, x_4(t-\tau) \\ \dot{x}_2(t) = -k_2\, {x_2}^2(t) + k_1\, u(t)\, x_1(t) \\ \dot{x}_3(t) = -k_3\, x_3(t) + k_2\, {x_2}^2(t)/2 \\ \dot{x}_4(t) = -k_4\, x_4(t) + k_3\, x_3(t) \\ y_1(t) = s_1\, (x_2(t) + 2\, x_3(t)) \\ y_2(t) = s_2\, (x_1(t) + x_2(t) + 2\, x_3(t)) \end{cases} \tag{10}$$

where $x_i(t)$ $i = 1, \ldots, 4$ denote the four involved molecular compounds, $u(t)$ is the input function, k_i $i = 1, \ldots, 4$ the kinetic rate constants, τ is a "delay reaction" which is obtained using a linear chain approximation with intermediate steps assumed equal to ($\tau = 10/k_4$), s_1, s_2 the scaling parameters, and $y_1(t), y_2(t)$ the experimentally observable quantities. The initial conditions are assumed to be zero, except for $x_1(0) = ic_1$ which needs to be estimated from the experimental data.

In the literature, practical identifiability of the model (10) has been analyzed using statistical criteria. In particular, in [10] the profile likelihood approach is used based on the idea of detecting flatness of the likelihood function $V_N(\boldsymbol{\theta})$ by exploring the parameter space in the direction of least increase of the objective function for each parameter component. This allows to experimentally observe the behavior of the function around a nominal parameter value.

In this way, the authors establish that parameters k_2, s_1, s_2 together with the initial condition ic_1 cannot be identifiable while the others are found to be (practically) identifiable except for one (k_3) where the minimum is so flat to be declared "practically nonidentifiable." They conclude that this structural nonidentifiability is a result of missing information about absolute concentration in the experimental setup. Thus, to get an identifiable model, they enrich the experiment and add a new measurement. In the more recent paper [14], the authors include the input into the parameters estimation process. In this case, to get the identifiability of the model, they fix the initial condition $x_1(0)$.

3.1 Identifiability Analysis of the JAK-STAT Model

Here we check structural identifiability of model (10) by using the software DAISY and practical identifiability at nominal parameter values by determining the rank of the sensitivity matrix. This comparative analysis seems to be done for the first time.

DAISY is applied initially to check the uniqueness of parameter estimates in the entire parameter space. Results, supported by the Gröbner bases computed analytically by the algorithm, show that parameters k_2, s_1, s_2, ic_1 are linked by algebraic constraints with one degree of freedom. This actually indicates that these four parameters are structurally nonidentifiable and that it is sufficient to know just one of them (not necessarily the scale factor parameter) to make the model structurally globally identifiable. This "flexibility" for recovering identifiability is an important issue because it allows for different choices in the design of the experiment, where many constraints exist especially in a biological experimental setup. The structural test also guarantees that the practical nonidentifiability of k_3 is actually only due to data problems and that it is sufficient to include only one constraint equation on the initial conditions to retrieve the structural identifiability of the model. Hence the structural identifiability analysis has essentially replicated by an analytical approach without experimental data nor assumptions on the initial conditions values, the results obtained about a nominal point in [10].

In order to integrate the analytical results provided by DAISY with the practical identifiability approach, the Gröbner basis determined for the JAK-STAT model (10)

Table 1 Gröber basis and Jacobian matrix

$G(\boldsymbol{\theta}, \boldsymbol{\theta}^*)$	k_1	k_2	k_3	k_4	s_1	s_2	ic_1
$20\,k_1 - 39$	20	0	0	0	0	0	0
$95\,k_2 - 11\,s_2$	0	95	0	0	0	−11	0
$25\,k_3 - 17$	0	0	25	0	0	0	0
$100\,k_4 - 149$	0	0	0	100	0	0	0
$19\,s_1 - 25\,s_2$	0	0	0	0	19	−25	0
$20\,s_2\,ic_1 - 19$	0	0	0	0	0	20	19

is recalculated for the same nominal parameter values used to assess practical identifiability [14]. Nominal parameter values are reported here as decimal as well as rational numbers, in parenthesis, being the latter used by DAISY for carrying out calculations on a ring with infinite precision: $k_1 = 1.95(39/20)$, $k_2 = 0.11(11/100)$, $k_3 = 0.68(17/25)$, $k_4 = 1.49(149/100)$, $s_1 = 1.25(5/4)$, $s_2 = 0.95(19/20)$, $ic_1 = 1(1)$ where k_3 was fixed to twice the lower limit.

The Gröbner basis reported in Table 1 (first column), and its Jacobian matrix formed by the partial derivatives of the Gröber basis with respect to the parameters, shown for an easier inspection in the right-hand side of Table 1, confirm the structural results already discussed. In particular, the first, third, and fourth rows and columns depend each upon one parameter only, and define thus uniquely the values of the structurally globally identifiable parameters k_1, k_3, and k_4. The second, fifth, and sixth rows involve the remaining nonidentifiable parameters, namely k_2, s_1, s_2, and ic_1, that are thus linked by algebraic constraints with one degree of freedom, i.e., three Gröbner basis equations in four unknowns.

In order to check practical identifiability, the model variables $y_1(t)$ and $y_2(t)$ are simulated between 0 and 60 min, together with all sensitivity equations, using the nominal parameter values mentioned previously. The model input, $u(t)$, was calculated between 0 and 30 min, as the positive half-cycle of the sine wave with total period 60 min. In Fig. 1 the time course of $y_1(t)$ and $y_2(t)$ are shown.

Virtually identical profiles (not shown) are obtained by changing the model nonidentifiable parameters according to the above Gröbner basis (Table 1).

In particular, the Gröbner basis polynomials equated to zero provide a globally identifiable parameterization of the model. It is easy to see that the most straightforward approach to reach it is to fix s_2 to an arbitrary numerical value and calculate the remaining parameters from the other equations: $k_1 = 39/20$, $k_2 = 11\,s_2/95$, $k_3 = 17/25$, $k_4 = 149/100$, $s_1 = 25\,s_2/19$, $ic_1 = 19/(20\,s_2)$, where fixed parameters are reported for completeness. By varying the value of s_2 one could observe that the trajectory of $y(t)$ is not affected.

Alternatively, as considered in [14], one can fix the initial condition ic_1, and calculate from Table 1, $s_2 = 19/(20\,ic_1)$, $s_1 = 5/(4\,ic_1)$, $k_2 = 209/(1900\,ic_1)$.

Obviously, assignment of any one of the parameters k_2, s_1, s_2, and ic_1 and recalculation of the other parameters is possible.

A final remark regards a geometric interpretation of the relationship between Gröbner basis and Jacobian matrix reported in Table 1, and the eigenvectors of Table 2 that form a basis for expressing output variations as functions of parameter

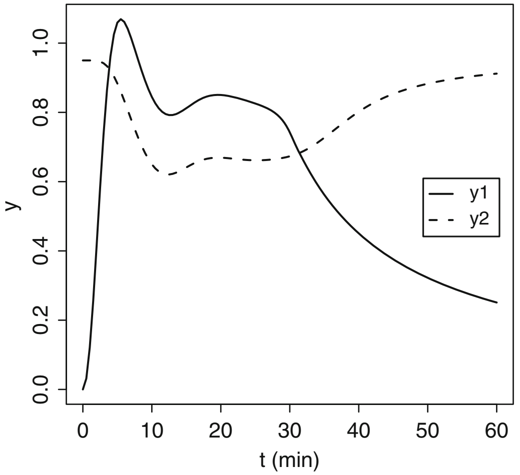


Fig. 1 Time course of model outputs using as forcing input $u(t) = \max(0, \sin(2\pi t/60))$

Table 2 Singular values σ of sensitivity matrix $\mathbf{S}(\boldsymbol{\theta})$ and relative right eigenvectors $\mathbf{V}$

σ	23.94	11.29	4.734	0.8312	0.5376	0.313	0
k_1	−0.00103	0.003642	−0.05622	0.06273	0.8462	−0.5261	0
k_2	0.9356	−0.2611	−0.2281	0.02152	−0.02083	−0.01021	0.05899
k_3	0.0228	0.000688	0.02984	0.2179	0.5047	0.8345	0
k_4	−0.04448	−0.04458	−0.01768	0.9721	−0.167	−0.151	0
s_1	−0.1928	0.07973	−0.7103	−0.01419	−0.01845	0.04546	0.6704
s_2	−0.1296	−0.7311	0.4328	−0.03032	0.02316	−0.01742	0.5095
ic_1	−0.2612	−0.6236	−0.5018	−0.04418	−0.003354	0.03915	−0.5363

variations: $\delta\mathbf{y}(t) = \mathbf{S}(\boldsymbol{\theta})\delta\boldsymbol{\theta}$. In particular, the last column of Table 2, $\mathbf{V}_7$, defines the null space for parameter perturbations, i.e., $\delta\mathbf{y}(t) = 0$ if $\delta\boldsymbol{\theta} \propto \mathbf{V}_7$, because $\sigma_7 = 0$. Interestingly, it can be verified that $\mathbf{V}_7$ generates also the null space for the Jacobian matrix in Table 1, i.e., $\nabla G(\boldsymbol{\theta}, \boldsymbol{\theta}^*)\mathbf{V}_7 = 0$ (up to roundoff errors). This result may be unexpected but not surprising since $G(\boldsymbol{\theta}, \boldsymbol{\theta}^*) = 0$ defines, for a fixed θ^*, the values of $\boldsymbol{\theta}$ that produce identical output trajectories. This is consistent with the fact that parameter variations, which do not modify output trajectories, do not change $G(\boldsymbol{\theta}^* + \delta\boldsymbol{\theta}, \boldsymbol{\theta}^*) \approx \nabla G(\boldsymbol{\theta}^*, \boldsymbol{\theta}^*)\delta\boldsymbol{\theta}$.

4 Conclusions

In this study we propose a unified viewpoint of the two identifiability analysis approaches, namely *structural* and *practical identifiability*, by motivating their joint use. These methodologies are traditionally regarded as disjoint because they are based, in turn, on differential algebraic manipulations or numerical simulation of systems equations. The former does not require experimental data and can be tested using differential algebra software, such as DAISY, on the model equations without assuming prior knowledge on parameter values, whereas the method based on sensitivity analysis requires "nominal" parameter values, obtainable from a simulated or a real experiment, and consists of procedures based on the analysis of the minima of a likelihood-type function depending on these data. These minima correspond to numerical parameter estimates. Thus an important difference consists in the inability of structural identifiability analysis to infer on the precision of parameter estimation, whereas practical identifiability analysis relies on simulated data that depend on assumed parameter values and on subjective thresholds to define a discrimination line between identifiable and nonidentifiable parameters. The relevant novelty of using both methodologies is that, in case of nonidentifiability, the structural analysis allows to integrate the practical identifiability results: one can calculate the analytical relations among the nonidentifiable parameters, described by the Gröbner basis of the model, at the "nominal" parameter values.

Furthermore, if the parameter turns out to be practically nonidentifiable from simulated or real data, without having performed the structural identifiability, it may be difficult to assess the causes from a practical identifiability test. Apparent nonidentifiability may be due either to structural nonidentifiability or to the paucity of information in the data or to an imprecise reconstruction, due to noise, of the level sets or of the minima of the function $V_N(\boldsymbol{\theta})$. This instead may be revealed, in analytic terms, by structural methods. By knowing for example that the model is structurally globally identifiable, the investigator knows that the problem is related to the simulated or real experimental data, for example, to the scarceness of measurement samples. Thus, the joint use of the two identifiability approaches can provide guidelines to avoid unfruitful studies and simulations of modified model structures.

References

1. Audoly, S., Bellu, G., D'Angiò, L., Saccomani, M.P., Cobelli, C.: Global identifiability of nonlinear models of biological systems. IEEE Trans. Biomed. Eng. **48**(1), 55–65 (2001)
2. Bellu, G., Saccomani, M.P., Audoly, S., D'Angiò, L.: DAISY: a new software tool to test global identifiability of biological and physiological systems. Comput. Methods Prog. Biomed. **88**, 52–61 (2007)
3. Buchberger, B.: Ph.D. thesis 1965: An algorithm for finding the basis elements of the residue class ring of a zero dimensional polynomial ideal. J. Symb. Comput. **41**(3), 475–511 (2006)
4. Chapman, M.J., Godfrey, K.R., Chappell, M.J., Evans, N.D.: Structural identifiability of non-linear systems using linear/non-linear splitting. Int. J. Control **76**(3), 209–216 (2003)
5. Chis, O., Banga, J.R., Balso-Canto, E.: Structural identifiability of systems biology models: a critical comparison of methods. PloS ONE **6**(11), e27755 (2011)
6. Cobelli, C., Saccomani, M.P.: Unappreciation of a priori identifiability in software packages causes ambiguities in numerical estimates. Letter to the editor. Am. J. Physiol. **21**, E1058–E1059 (1990)

7. Joly-Blanchard, G., Denis-Vidal, L.: Some remarks about identifiability of controlled and uncontrolled nonlinear systems. Automatica **34**, 1151–1152 (1998)
8. Ljung, L., Glad, S.T.: On global identifiability for arbitrary model parameterizations. Automatica **30**(2), 265–276 (1994)
9. Ollivier, F.: Le problème de l'identifiabilité structurelle globale: étude théorique, méthodes effectives et bornes de complexité. Thèse de Doctorat en Science, École Polytéchnique, Paris (1990)
10. Raue, A., Kreutz, C., Maiwald, T., Bachmann, J., Shilling, M., Klingmüller, U., Timmer, J.: Structural and practical identifiability analysis of partially observed dynamical models by exploiting the profile likelihood. Bioinformatics **25**, 1923–1929 (2009)
11. Raue, A., Karlsson, J., Saccomani, M.P., Jirstrand, M.M., Timmer, J.: Comparison of approaches for parameter identifiability analysis of biological systems. Bioinformatics **30**(10), 1440–1448 (2014)
12. Rodriguez-Fernandez, M., Rehberg, M., Kremling, A., Banga, J.R.: Simultaneous model discrimination and parameter estimation in dynamic models of cellular systems. BMC Syst. Biol. **7**, 76 (2013)
13. Saccomani, M.P., Audoly, S., D'Angiò, L.: Parameter identifiability of nonlinear systems: the role of initial conditions. Automatica **39**, 619–632 (2004)
14. Schelker, M., Raue, A., Timmer, J., Kreutz, C.: Comprehensive estimation of input signals and dynamics in biochemical reaction networks. Bioinformatics, ECCB **28**, i529–i534 (2012)
15. Seber, G.A., Wild, C.J.: Nonlinear Regression. Wiley, New York (1989)
16. Thomaseth, K., Batzel, J.J., Bachar, M., Furlan, R.: Parameter estimation of a model for Baroreflex control of unstressed volume. In: Mathematical Modeling and Validation in Physiology, 215–246. Springer, Berlin (2012)

Boolean Dynamics of Compound Regulatory circuits

Elisabeth Remy, Brigitte Mossé, and Denis Thieffry

Abstract In biological regulatory networks represented in terms of signed, directed graphs, topological motifs such as circuits are known to play key dynamical roles. After reviewing established results on the roles of simple motifs, we present novel results on the dynamical impact of the addition of a short-cut in a regulatory circuit. More precisely, based on a Boolean formalisation of regulatory graphs, we provide complete descriptions of the discrete dynamics of particular motifs, under the synchronous and asynchronous updating schemes. These motifs are made of a circuit of arbitrary length, combining positive and negative interactions in any sequence, and are including a short-cut, and hence a smaller embedded circuit.

Keywords Regulatory motifs • Boolean dynamics

1 Introduction

1.1 Motivations

Biological regulatory networks are often represented in terms of signed, directed graphs. In these graphs, topological motifs, such as elementary directed and signed circuits, also often called ‘feedback loops’, are known to play significant dynamical roles [19]. In particular, positive regulatory circuits (involving an even number of negative interactions) have been associated with multistability, and more generally with the occurrence of multiple attractors, which may account for biological differentiation phenomena. On the other hand, negative circuits have been associated with sustained periodic behaviour and/or homeostasis [18]. Necessary conditions

E. Remy (✉) • B. Mossé
Aix Marseille Université, CNRS, Centrale Marseille, I2M, UMR 7373, 13453 Marseille, France
e-mail: elisabeth.remy@univ-amu.fr; brigitte.mosse@univ-amu.fr

D. Thieffry
Computational Systems Biology team, Institut de Biologie de l’Ecole Normale Supérieure (IBENS), CNRS UMR8197, INSERM U1024, Ecole Normale Supérieure, PSL Research University, F-75005 Paris, France
e-mail: denis.thieffry@ens.fr

A. Rogato et al. (eds.), *Dynamics of Mathematical Models in Biology*,
DOI 10.1007/978-3-319-45723-9_4

relating the occurrence of such circuits with the corresponding dynamical properties have been defined and properly demonstrated in continuous and discrete frameworks [13, 14, 16, 17]. However, the dynamical properties of more complex regulatory motifs made of intertwined circuits still need to be clarified [6]. In this article, we rely on a Boolean modelling framework (introduced in Sect. 1.2) to review recent achievements associating simple or more complex regulatory motifs with specific dynamical properties, i.e. in terms of the number and type of attractors (Sect. 1.3). Next, we report novel results regarding the dynamical properties of *chorded circuits*, made of an elementary (positive or negative) circuit with a chord (Sect. 2). For sake of brevity, we introduce our main results here, leaving the details (theorems and proofs) for a forthcoming publication.

1.2 Boolean Formalism

The Boolean formalisation of (biological) regulatory networks relies on the delineation of two types of graphs, called regulatory graphs and state transition graphs [1] [5].

In a regulatory graph, each vertex represents a regulatory component and each arc (oriented, signed edge) represents a regulatory interaction (activation or inhibition) between two components. Here, each component is associated with a Boolean variable, meaning that it can take two possible levels, 0 or 1, denoting the absence/inactivation or the presence/activity of the modelled entity. A logical rule associated with each component specifies its target value depending on the presence/absence of its regulators. The dynamical behaviour of the resulting model can then be computed starting from any initial state, step by step, updating the current state according to the logical formulae (logical simulations).

The dynamics of a logical model can be represented in terms of a state transition graph (STG), in which vertices denote different states of the system (represented by a Boolean vector encompassing the values of all the components), whereas arcs represent enabled transitions between pairs of states.

In this work, we have considered the two main updating policies for the generation of STGs. According to the synchronous policy, all components are updated simultaneously at each step; consequently, each state has at most one

[1] Classical terms of graph theory can be found in [3]. Moreover, we use here the following terminology:

Isolated (elementary) circuit: a connected directed graph with every vertex of in-degree and out-degree equal to 1;
Circuit: a subgraph of a regulatory graph amounting to an isolated circuit;
Flower-graph: group of circuits sharing one single vertex;
Chorded circuit: circuit with a chord, possibly a self-loop;
Cycle: a subgraph of a state transition graph amounting to an isolated circuit.

successor. In contrast, according to the asynchronous policy, only one variable can be updated at each step and all the possible successors of a state are considered (non-deterministic, branching dynamics).

Of particular interest are the sets of states forming attractors, i.e., minimal groups of states from which the system cannot escape, which represent potential asymptotical behaviours. Attractors can be ranged into two main classes: stable states, corresponding here to fixed states (i.e. without successors), and cyclic attractors, corresponding here to terminal cycles or to more complex terminal strongly connected components comprising several intertwined cycles.

Several methods have been proposed to efficiently identify all stable states (see, e.g., [10]). However, other means are needed to assess the reachability of the stable states from specific initial states, or yet to identify cyclic attractors (see, e.g., [8]). Proper dynamical analyses often rely on the computation of the STG. As the size of the model increases, the size of the STG increases exponentially. To cope with this problem, one can reduce directly the model before simulation (model reduction), and/or compress the resulting STG into a hierarchical transition graph (HTG) [4]. Strongly connected components (SCCs) form a partition of the STG. They are trivial (constituted by a unique state) or complex (containing at least two states). The compression of an STG into a HTG is achieved by clustering the states of the complex SCCs, and gathering the trivial SCCs leading to the same complex SCC and attractors. The components grouping trivial SCCs are called *irreversible components*.

The HTG displays all the reachable attractors, and the other clusters of states leading to one single attractor or to specific subsets of attractors. HTG computation is done on the fly, without having to store the whole STG, which often enables strong memory and CPU usage shrinking [4]. Furthermore, this functionality eases the identification of the key commutations (change of component levels) underlying irreversible choices between the different reachable attractors. The HTG representation is very compact and very informative regarding the organisation of the original STG.

1.3 Asymptotical Properties of Simple Motifs

Figures 1 and 2 provide examples of different classes of motifs, each endowed with specific dynamical properties. The two first rows of Fig. 1 correspond to the two main classes of regulatory circuits, also often called feedback circuits or feedback loops, namely positive and negative circuits [19]. The third and fourth rows correspond to the two main (coherent versus incoherent) classes of feedforward motifs, also often called—somewhat improperly—feedforward loops. The dynamical properties for these simple motifs have been extensively documented [1, 13].

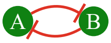
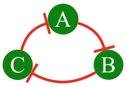
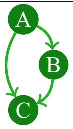
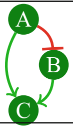
Type of motif	Definition/Topology	Properties (Boolean case)
	**Positive circuits** Circular sequence of signed interactions involving an even number of negative interactions	Multistability Differentiation
	**Negative circuits** Circular sequence of signed interactions involving an odd number of negative interactions	Periodic or homeostatic behaviour Biological cycles
	**Coherent FFM** Direct and indirect (via B) interactions from input A onto output C, with coherent (positive or negative) effects on output	Filtering of transient signal
	**Incoherent FFM** Direct and indirect (via B) interactions from input A onto output C, with incoherent effects on output	Generation of pulses

Fig. 1 Boolean dynamics of simple regulatory motifs: summary of previous results (notation FFM: feedforward motif)

Over the last years, a series of results has been obtained regarding the dynamical properties of compound regulatory circuits, in particular sets of circuits sharing one single vertex ('hub vertex'). In such cases, one can infer the dynamics of the whole system based on that of the hub vertex, as the hub vertex fully determines (directly or indirectly) the behaviour of the other vertices. Hence, these flower-graphs (as they are called in [7]) can give rise to 0, 1 or 2 stable states. Figure 2 gives six examples of such motifs; they together illustrate all possible situations in terms of attractors, i.e., regarding the potential occurrence of multiple stable states or of cyclic attractors. The first motif is associated with bistability (coexistence of two mirroring stable states), in the absence of cyclic attractor. The second motif has a unique, cyclic attractor. The fifth and sixth motifs have a unique stable state and no cyclic attractor. The third and fourth motifs correspond to a variety of dynamical situations depending on the logical rules associated with the hub vertex (AND, OR or XOR).

Note that each of these cases represents a large class of networks, encompassing potentially more vertices and interactions, but which can be formally reduced to these prototypic motifs without fundamental impact on the dynamics (i.e. regarding the number and types of attractors, see [11]). This suggests that the association of specific dynamical behaviours with the motifs listed in Figs. 1 and 2 could be extended to larger classes of motifs.

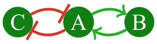
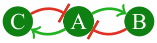
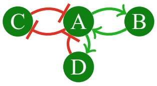
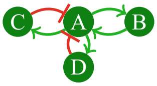
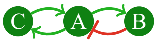
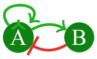
Type of motif	Definition/Topology	Properties
	Composition of positive circuits sharing one component (hub)	Multistability -Two stable states
	Composition of negative circuits sharing one component (hub)	Oscillatory behavior - No stable state
	Composition of circuits including one negative circuit and at least one positive circuit, all sharing one component (hub)	At least one stable state (depending on the logical rule associated with the hub)
	Composition of circuits including one positive circuit and at least one negative circuit, all sharing one component (hub)	At most one stable state (depending on the logical rule associated with the hub)
	Composition of one negative and one positive circuits sharing one component (hub)	One stable state - No attracting cycle
	Composition of a two-component negative circuit with one positive autoregulation	One stable state - No attracting cycle

Fig. 2 Boolean dynamics of simple regulatory motifs: summary of previous results

2 Boolean Dynamics of Circuits and Chorded Circuits

Most of the works on the regulatory motif listed in Fig. 2 focus only on their asymptotical behaviours (attractors, and even often only stable states). In the line of our previous study devoted to isolated circuits [13], we describe the whole synchronous and asynchronous STGs of regulatory motifs made of an isolated circuit with a unique chord—possibly a self-loop—(chorded circuits), and compare their dynamical properties to those of isolated circuits. Using combinatorics on specific abacus and analysis of recurrent sequences (not shown here), we emphasise that whatever the chosen updating rule, the STG depends on a small number of parameters.

We recall the structural properties of the synchronous and asynchronous STGs of isolated circuits in Sect. 2.1. Then, we present an outline of our new results concerning the synchronous and asynchronous dynamical structures of chorded circuits (Sect. 2.2).

2.1 Boolean Dynamics of Isolated Circuits

Whatever the updating policy, the STG of an isolated circuit is hierarchically organised in different levels, each encompassing states with identical numbers of updating calls (i.e. the number of genes called to change their expression level). At each level, the number of updating calls is always even in the case of a positive circuit, while it is odd in the case of a negative circuit.

In the synchronous case, the STG encompasses vertex-disjoint cycles involving states with the same number of updating calls.

In the asynchronous case, the STG of isolated circuits is connected. In the case of a positive isolated circuit, the STG is characterised by two stable states (no cyclic attractor), while in the case of a negative isolated circuit it is characterised by a single cyclic attractor (the STG generated by a 4-component positive circuit under asynchronous updating is shown in Figure 3(I), centre).

2.2 Boolean Dynamics of Chorded Circuits

Chorded circuits are made of a *long* circuit with a chord (additional short-cut interaction) between two components of the circuit (or amounting to a self-loop), thereby creating a *small* circuit (see Fig. 3(II), (III) and (IV) left). The chorded circuit is *coherent* if the signs of the two embedded circuits are identical; otherwise, the chorded circuit is *incoherent*. We compared the dynamics of chorded circuits with the dynamics of the long circuit. In any case, part of the states keep the same updating calls, while other states are sensitive to the presence of the short-cut, and called therefore hereafter *sensitive states*. Three cases for the logical rule have been considered, using the logical operators OR, AND and XOR. Note that using XOR amounts to define two dual interactions (i.e. with context sensitive signs) converging on a single vertex. The dynamics obtained with OR and AND rules are symmetrical: one can transform one of the resulting STGs into the other one by switching (ON or OFF) all component values. The topology of the STG, and thus the dynamical properties depend on the sign of the long circuit, and if it is a coherent chorded circuit or not. In contrast, the topology of the STG and the dynamics obtained with the XOR rule depends only on the number of genes involved, not on the signs of the two circuits.

2.2.1 Attractors of Chorded Circuits for the Synchronous Updating

In the cases of the OR and AND logical rules, the synchronous STG contains terminal cycles.

- If the long circuit is positive, these terminal cycles are found in the synchronous STG of the long circuit. If the chorded circuit is further coherent (positive small circuit), there are two stable states; if it is incoherent (negative small circuit), there is only one stable state.

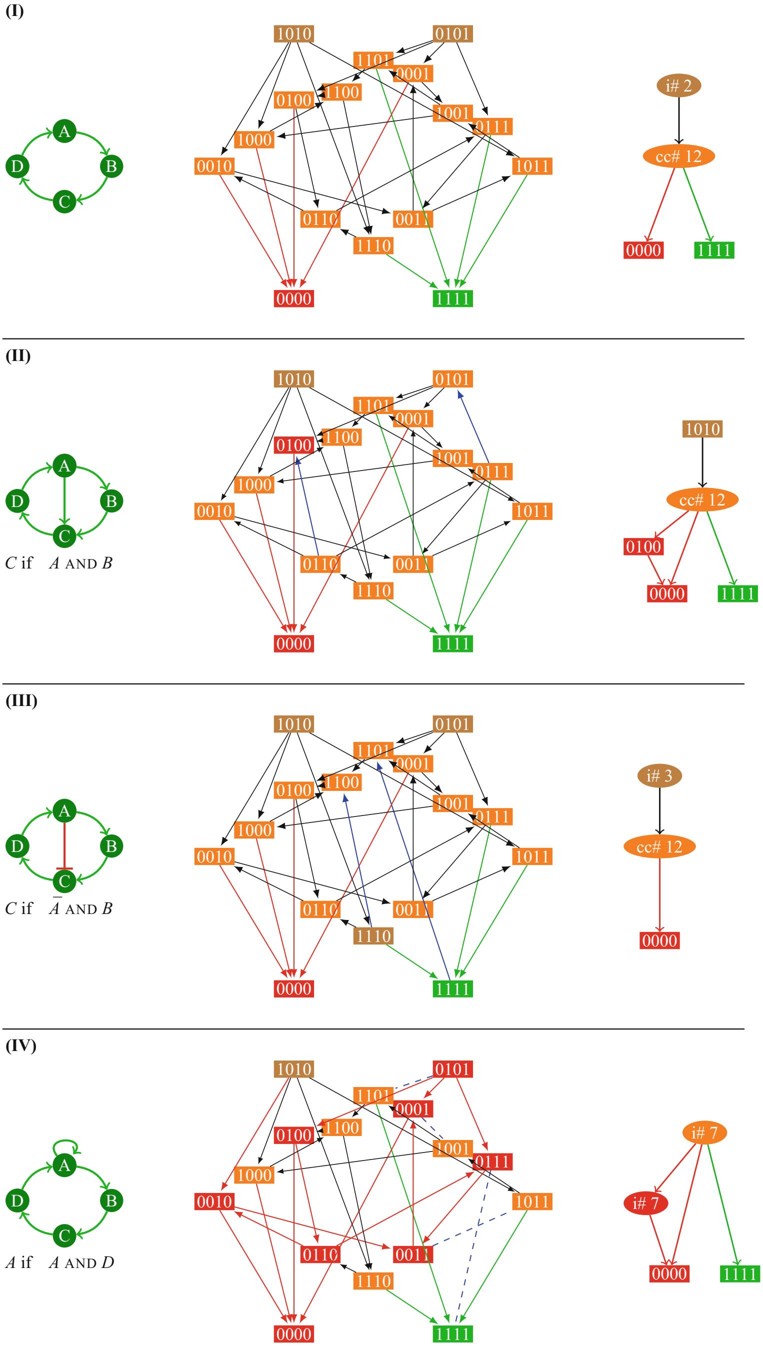


Fig. 3 Description of the asynchronous dynamics of: a 4-components isolated circuit **(I)**; a coherent chorded circuit **(II)**; an incoherent chorded circuit **(III)**; a circuit with a coherent self-regulation **(IV)**. From *left to right*: regulatory graph, state transition graph (STG) and its compression into a hierarchical transition graph (HTG). In the later, 'cc' and 'i' stand for cyclic and irreversible components, respectively, while the number written after '#' corresponds to the number of states encompassed by the component

- If the long circuit is negative, the terminal cycles differ from those obtained for the long circuit. If the chorded circuit is incoherent (positive small circuit), there is only one stable state; if it is coherent (negative small circuit), there is no stable state.

Accordingly, in the cases OR and AND, a coherent chorded circuit and its corresponding long circuit have the same number of stable states.

In the case of the XOR logical rule, the synchronous STG is constituted of vertex-disjoint cycles. It contains only one stable state and cycles with pseudo-random sequence of states, whatever the signs of the circuits.

2.2.2 Attractors of Chorded Circuits for the Asynchronous Updating

When the small circuit is not a self-loop, the asynchronous STG of the chorded circuit is obtained from that of the long circuit by changing the direction of edges between pairs of sensitive states that differ by the coordinate of the target component of the short-cut. When the small circuit consists in a self-loop, these edges are suppressed or created.

In the cases of the OR and AND logical rules, compare Fig. 3(II) and (III) with Fig. 3(I) if the small circuit is not a self-loop, and Fig. 3(IV) and (I) in the case of a self-loop. It can be demonstrated that a coherent chorded circuit and its corresponding long circuit have the same number and type of attractors, and in particular the same number of stable states. When the chorded circuit is incoherent, there is a unique attractor: a stable state. Moreover, the STG of an isolated circuit is always symmetrical by the transformation switching the component values (cf. Fig. 3(I) centre: the structure of the STG is conserved when switching all Os to 1 and *vice-versa*), and encompasses pairs of such symmetrical states at each level (a level is characterised by a constant number of updating calls) [13]. The introduction of a short-cut skews the dynamics. For example, in the case where both long and small circuits are positive, the basin of attraction of one of the stable states is increased at the expense of the other one (compare Fig. 3(II) with Fig. 3(I), right).

In the case of the XOR logical rule, the asynchronous STG of a chorded circuit encompasses a unique stable state as sole attractor. As using an XOR rule amounts to introduce dual regulations, this could be considered as a particular case of incoherent chorded motif.

3 Conclusion and Prospects

Figure 4 summarises our novel results regarding the dynamics of chorded circuits, focusing on the Boolean framework and the asynchronous updating scheme, and considering three different rules (AND, OR and XOR) for the vertex targeted by two regulations. These results can be generalised to a wide range of regulatory

	Isolated circuits	AND/OR **chorded circuits**	XOR **chorded circuits**
Synchronous STG	Vertex-disjoint cycles In each cycle, states with the same number of updating calls	Terminal cycles	Vertex-disjoint cycles Long pseudo-random cycles One stable state
	Positive circuits Even numbers of transitions Two stable states	**Positive long circuits** Terminal cycles from the synchronous STG of the long circuit • **Coherent** chorded circuit Two stable states • **Incoherent** chorded circuit One stable state	
	Negative circuits Odd numbers of transitions No stable state	**Negative long circuits** • **Coherent** chorded circuit No stable state • **Incoherent** chorded circuit One stable state	
Asynchronous STG	Connected level structure Levels form the SCCs (except perhaps for the two extremal levels), and gather states with the same number of successors	Deduced from the asynchronous STG of the long circuit ↪ deleting or creating edges if the short-circuit is a self-loop ↪ inverting edges otherwise	Deduced from the asynchronous STG of the long circuit ↪ deleting or creating edges if the small circuit is a self-loop ↪ inverting edges otherwise One stable state
	Positive circuits Even numbers of transitions Two stable states	**Positive long circuits** • **Coherent** chorded circuit Two stable states • **Incoherent** chorded circuit One stable state	
	Negative circuits: Odd numbers of transitions One terminal SCC	**Negative long circuits** • **Coherent** chorded circuit One terminal SCC • **Incoherent** chorded circuit One stable state	

Fig. 4 Boolean asynchronous dynamics of chorded circuits, compared to that of isolated circuits

motifs, e.g., involving longer short-cut paths, with the help of the reduction method described in [11]. However, simple and compound regulatory motifs are usually embedded in large, intricated networks. In this respect, it can be shown that motifs embedded in more complex networks may still display the associated properties in specific conditions, called ‘context of functionality’ in [6].

Noteworthy, recent developments in synthetic biology recurrently refer to regulatory motifs corresponding to the classes considered in this study, thereby demonstrating the potential practical impact of studies aiming at fully characterise the dynamics of simple regulatory motifs (for recent reviews on synthetic biological circuits, see [9, 12, 20]).

When facing a large and complex network, the enumeration and analysis of its constitutive motifs can lead to interesting insights about the network dynamics. For example, in the Boolean case, a bound on the number of attractors can be computed based on the number of positive regulatory circuits, taking into account potential (indirect) cross-interactions between them [2, 15]. Such results could be refined by considering recent results on the Boolean dynamics of more complex motifs, such as the flower-graphs [7], or yet the chorded circuits reported here.

More prospectively, the results obtained in the Boolean framework could serve as a guide to extend them to the multilevel logical framework, or even to transpose them into the differential framework, as it was the case with the delineation of theorems linking elementary positive and negative regulatory circuits with multistability and cyclic properties (see, e.g., [17, 19]).

References

1. Alon, U.: Network motifs: theory and experimental approaches. Nat. Rev. Genet. **8**(6), 450–461 (2007)
2. Aracena, J., Demongeot, J., Goles, E.: Positive and negative circuits in discrete neural networks. IEEE Trans. Neural Netw. **15**(1), 77–83 (2004)
3. Bang-Jensen, J., Gutin, G.: Digraphs, Theory, Algorithms, Applications. Springer, Berlin (2008)
4. Bérenguier, D., Chaouiya, C., Monteiro, P.T., Naldi, A., Remy, E., Thieffry, D., Tichit, L.: Dynamical modeling and analysis of large cellular regulatory networks. Chaos (Woodbury N.Y.) **23**(2), 025114 (2013)
5. Chaouiya, C., Remy, E., Mossé, B., Thieffry, D.: Qualitative analysis of regulatory graphs : a computational tool based on a discrete formal framework. In: Lecture Notes in Control and Information Science, vol. 294, pp. 119–26. Springer, Berlin (2003)
6. Comet, J.-P., Noual, M., Richard, A., Aracena, J., Calzone, L., Demongeot, J., Kaufman, M., Naldi, A., Snoussi, E.H., Thieffry, D.: On circuit functionality in boolean networks. Bull. Math. Biol. **75**(6), 906–919 (2013)
7. Didier, G., Remy, E.: Relations between gene regulatory networks and cell dynamics in Boolean models. Discret. Appl. Math. **160**(15), 2147–2157 (2012)
8. Garg, A., Dicara, A., Xenarios, I., Mendoza, L., De Micheli, G.: Synchronous vs. Asynchronous Modeling of Gene Regulatory Networks, Bioinformatics (Oxford, England) **24**(17), 1917–1925
9. Khalil, A.S., Collins, J.J.: Synthetic biology: applications come of age. Nat. Rev. Genet. **11**(5), 367–379 (2010)
10. Naldi, A., Thieffry, D., Chaouiya, C.: Decision diagrams for the representation and analysis of logical models of genetic networks. In: Computational Methods in Systems Biology. Lecture Notes in Computer Science, vol. 4695, pp. 233–47. Springer, Berlin (2007)
11. Naldi, A., Remy, E., Thieffry, D., Chaouiya, C.: Dynamically consistent reduction of logical regulatory graphs. Theor. Comput. Sci. **412**(21), 2207–2218 (2011)
12. Purcell, O., Savery, N.J., Grierson, Claire, S., di Bernardo, M.: A comparative analysis of synthetic genetic oscillators. J. R. Soc. Interface/R. Soc. **7**(52), 1503–1524 (2010)
13. Remy, E., Mossé, B., Chaouiya, C., Thieffry, D.: A description of dynamical graphs associated to elementary regulatory circuits. Bioinformatics (Oxford, England) **19**(Suppl. 2), 172–178 (2003)
14. Remy, E., Ruet, P., Thieffry, D.: Graphic requirements for multistability and attractive cycles in a Boolean dynamical framework. Adv. Appl. Math. **41**(3), 335–350 (2008)
15. Richard, A.: Positive circuits and maximal number of fixed points in discrete dynamical systems. Appl. Math. **157**(15), 3281–3288 (2009)

16. Richard, A., Comet, J.-P.: Necessary conditions for multistationarity in discrete dynamical systems. Discret. Appl. Math. **155**(18), 2403–2413 (2007)
17. Soulé, C.: Graphic requirements for multistationarity. Complexus **1**, 123–133 (2003)
18. Thomas, R.: On the relation between the logical structure of systems and their ability to generate multiple steady states or sustained oscillations. In: Numerical Methods in the Study of Critical Phenomena. Springer Series in Synergetics **9**, 180–193 (1981)
19. Thomas, R., D'Ari, R.: Biological Feedback. CRC Press, Boca Raton (1990)
20. Weber, W., Fussenegger, M.: Synthetic gene networks in mammalian cells. Curr. Opin. Biotechnol. **21**(5), 690–696 (2010)

A Differential Transcriptomic Approach to Compare Target Genes of Homologous Transcription Factors in Echinoderm Species

Elijah K. Lowe, Claudia Cuomo, and Maria I. Arnone

Abstract Embryonic development is controlled by differential gene expression throughout developmental time. The ParaHox genes, *Cdx* and *Xlox* have been shown to be involved in the formation of the properly functioning gut in the sea urchin *Strongylocentrotus purpuratus* and the sea star *Patiria miniata*. Several genes involved in the gene regulatory network (GRN) are known, however, the network is still incomplete. With the current state of sequencing technology, we are now able to expand the network and gain further insight into the process of gut development on a more global scale. Through the use of high-throughput sequencing technology and knockdown experiments we have further characterized the effects of *Cdx* and *Xlox* on the GRN involved in gut development at different developmental stages. Additionally, we have conducted a cross-species comparison to identify genes that are more likely to be evolutionarily important for the development of the echinoderm gut. Within those genes we found a number of transcription factors that could potentially have important roles in the formation of the echinoderm gut. Using both RNA-seq and gene homology, we have set the foundation for further studies of echinoderm gut and the ParaHox GRN downstream of *Xlox* and *Cdx*.

Keywords Differential transcriptomics • Gene regulatory network

1 Introduction

The developmental program of an organism and its phenotypic features are encoded into its DNA. The binding of transcription factors to specific DNA, which controls the expression of genes and ultimately the development of the embryo, is known as a gene regulatory network (GRN). Evolutionary conservation has provided us

E.K. Lowe (✉)
Stazione Zoologica Anton Dohrn, Naples, Italy

Beacon Center for Evolution in Action, Michigan State University, East Lansing, MI 48823, USA
e-mail: elijahkariem.lowe@szn.it

C. Cuomo • M.I. Arnone
Stazione Zoologica Anton Dohrn, Naples, Italy

A. Rogato et al. (eds.), *Dynamics of Mathematical Models in Biology*,
DOI 10.1007/978-3-319-45723-9_5

with a good tool to study the origins of phenotypic features and their developmental programs. With the advances in sequencing technology and the continued drop in prices, it has become more common to sequence an organism's transcriptome. This has facilitated the ability to examine organism on a genomic scale, allowing the study of all genes expressed at a giving time point in development, as well as for wild type versus experimental conditions. With transcriptomics we are able to better understand the complicity of evolution and increasing studies are taking advantage of this fact [1, 2].

In bilateria, homeobox-containing genes are important for the patterning of the anterior–posterior axis and mediate much of the embryonic development, with one of the most studied families being the Hox genes [3, 4]. Another important family of homeobox genes is the ParaHox family—*Gsx*, *Pdx* (*Xlox* in echinoderms), and *Cdx*, which are thought to be the ancient sister group to Hox genes and to have emerged from the ProtoHox cluster [5]. The ParaHox genes have been shown to be involved in gut development in vertebrates [6, 7] and also in the echinoderms [8, 9]. It appeared from the examination of the sea urchins *Strongylocentrotus purpuratus* that echinoderm had lost some chordate-like features in their function of Xlox and Cdx [10]. However, through the use of another echinoderm, the bat star *Patiria miniata*, it was discovered that these features appear to only have been lost in echinoids, while being retained in asteroids [11]. This shows the necessity to continue to study new organisms in order to gain a more complete evolutionary picture. The embryonic guts of both *S. purpuratus* and *P. miniata* first form a tube like structure with no sections known as the archenteron, then later divide into three sections, the foregut, the midgut, and the hindgut, which become in the larva the esophagus, the stomach, and the intestine, respectively.

Portions of the GRN for gut development in echinoderms have already been formed, but the network downstream of Xlox and Cdx has yet to be assembled. In *S. purpuratus*, Xlox morpholino antisense oligo (MASO) RNA-seq experiments have been conducted looking at known genes in the network [9], but have not been studied in-depth. Here we present the groundwork for reconstructing the GRN for gut development downstream of Xlox and Cdx in both *S. purpuratus* and *P. miniata*. Through the analysis of these MASO RNA-seq experiments we will identify direct and indirect targets of Xlox and Cdx in both species. Secondly, looking at the overlap in these two networks at homologous stages, and will better define the genes needed for the developing gut to form and properly section.

2 Results and Discussion

2.1 Gene Orthology

Prior to understanding or reconstructing the gut GRN for *S. purpuratus* and *P. miniata*, we must first understand the homology relationship between the two species. Proteomes for *S. purpuratus*, *P. miniata*, and *Xenopus tropicalis* were used

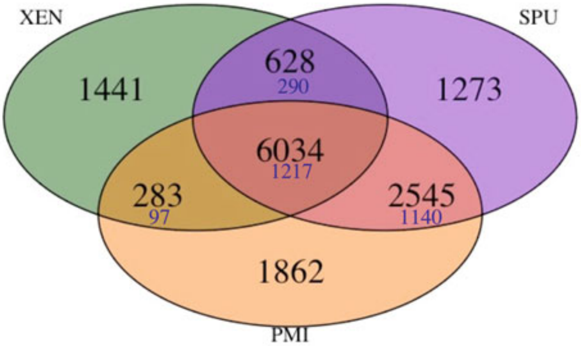


Fig. 1 Gene ortholog relationship between *S. purpuratus* (SPU), *P. miniata* (PMI), and *X. tropicalis* (XEN). Each *circle* represents one of the species and their overlap represents the orthologous groups that are in common. The numbers in the *larger black print* are the total number of orthologous groups and in the *smaller blue print* are the number of single copy orthologous groups

to construct orthologous groups and examine the gut GRN on an evolutionary scale. There were 29,805, 29,129, and 22,718 protein sequences in each proteome, respectively. Five sequences were removed from the *P. miniata*'*s* proteome during the filtering process, they were all eight base pairs or less in length. The three proteomes clustered into 14,066 homologous groups, being composed of 22,576 *S. purpuratus* proteins in 10,480 groups, 22,252 *P. miniata* proteins in 10,724 groups, and 20,813 *X. tropicalis* proteins in 8386 groups. Of these orthologous groups there were 6034 conserved amongst all three organisms, with 1217 (20 %) being single copy orthologs. The echinoderms had the largest number of orthologs as expected with 2545 orthologous groups and 45 % of the groups being single copy orthologs (Fig. 1).

2.2 Differential Expression Analysis

To identify genes downstream of *Xlox* and *Cdx* we analyzed both *S. purpuratus* and *P. miniata* embryos that were separately injected with MASO designed to block translation of *Xlox* or *Cdx* in each species. Time points for transcriptomic sequencing were selected based on QPCR expression trends from earlier studies [9, 11]; the midpoint of expression was chosen for each gene in their respective species. In *S. purpuratus* the time points selected are 48 hpf (late gastrula) and 72 hpf (pluteus larva) for *Xlox* and 66 hpf (prism) for *Cdx*. In *P. miniata* the time points are 66 hpf (late gastrula) for *Xlox* and 90 hpf (early bipinnaria larva) for *Cdx*, which are homologous stages to those of *S. purpuratus*. Looking at the morphology of the embryos, the sea urchin and the sea star late gastrula represent the stages when the gut is only an elongated tube without any constrictions, instead the sea urchin

Table 1 Homology of differential expressed transcripts

	Splox 48 h (294)	Splox 72 h (2384)	Spcdx 66 h (723)	Pmlox 66 h (108)	Pmcdx 90 h (693)
Orthologous groups	183	1457	470	70	404
Proteins in core	97 (33 %)	929 (39 %)	289 (40 %)	39 (36 %)	270 (39 %)
SCO in all	16 (5 %)	145 (6 %)	39 (5 %)	5 (5 %)	34 (5 %)
SCO in echino	23 (8 %)	150 (6 %)	48 (7 %)	10 (9 %)	31 (4 %)
Total proteins	207 (70 %)	1659 (70 %)	529 (73 %)	78 (72 %)	450 (65 %)

In parenthesis is the number of differential expressed genes for the given MASO RNA-seq experiment. Orthologous groups refer to the number of groups the total number of proteins were clustered into, while "Total proteins" refers to the total number of proteins that were clustered into orthologous groups. Proteins in core are the number of proteins found in *S. purpuratus*, *P. miniata*, and *X. tropicalis*. SCO in all are the number of single copy orthologous found in *S. purpuratus*, *P. miniata*, and *X. tropicalis*, while SCO in echino are the number of single copy orthologs found in only *S. purpuratus* and *P. miniata*

prism and the sea star early bipinnaria larva have already a tripartite shaped gut. The pluteus larva is an extra time point we chose for the sea urchin in which the gut is now complete with its cardiac and pyloric sphincters visible.

Differential expressed transcripts were identified using DESeq2 with a threshold of $\log_2\text{fc} > \pm 0.5$ and adjusted p-value of <0.05. In *S. purpuratus*, as time progressed, the knockdowns had a larger effect of more transcripts. There were only a couple of hundreds (294) transcripts affected by the Sp-Lox MASO at 48 hpf, compared to 723 transcripts effected by Sp-Cdx at 66 h, and 2384 at 72 h. Fifty-seven percent (167) of transcripts affected by the Sp-Lox MASO at 48 hpf were also affected at 72 hpf, showing similarities in the GRN as gut transitions from a tube like structure to a trisectioned gut.

When examining the *P. miniata* Xlox MASO RNA-seq we did not find a large number of transcripts to be differentially expressed at the late gastrula stage, with there being only 109 transcripts differentially expressed. However, when examining Pm-Cdx MASO RNA-seq at the early larva stage we observed many more genes being affected, 693, 450 (65 %) of which had a homologous relation to *S. purpuratus* and/or *X. tropicalis*.

Across all species at least 65 % of the transcripts were clustered into homologous groups, meaning that 30–35 % of the transcripts from each experiment were species specific or fell below our threshold (Table 1). Further analysis including phylogenic trees is necessary to better understand the relationship of these two species, but is currently out of the scope of this paper.

2.3 *Evolutionary Conserved Elements in S. purpuratus and P. miniata Gut GRNs*

Through the use of our orthology analysis and our MASO differential expression analysis we are able to discover conserved components in the downstream networks of both *Xlox* and *Cdx* in *S. purpuratus* and *P. miniata*. In the Cdx MASO RNA-seq there was the largest overlap between the species, 129 transcripts were found in both networks. Ninety-one out of these 129 genes were found in the "core" orthology group, meaning that at least one gene from *S. purpuratus*, *P. miniata*, and *X. tropicalis* was present in the orthologous group, and 11 (9 %) of those genes were identified as transcription factors that belong to the bzip, bHLH, C2H2, hmg, p53, and zf-C4 families. Late gastrula in *S. purpuratus* and *P. miniata* occurs at 48 hpf and 66 hpf, respectively, with 15 genes shared in their network, 67 % (10) of which were transcription factors. Although 48 hpf in *S. purpuratus* and 66 hpf in *P. miniata* are more morphologically similar, the overlaps in affected genes were stronger at 72 hpf in *S. purpuratus* and 66 hpf in *P. miniata*, with an additional 10 genes (25 in total) compared to the earlier stage, which also included the same group of transcription factors. Without the use of ChIP or other technologies such as ATAC-seq we are not able to determine the connectivity of these GRNs. Although we are not able to distinguish direct versus indirect targets in this study, identifying key components in the way of transcription factors is essential and will provide a foundation for future studies.

3 Conclusion

Here we present the foundation for studying the downstream GRN for gut development in *S. purpuratus* and *P. miniata* through the use of a MASO RNA-seq analysis. Seeing that RNA-seq can yield hundreds to thousands of potential genes we used the correlation between *S. purpuratus* and *P. miniata* to identify a subset of genes to be examined in future studies. Moreover the genes identified in our study as transcription factors will be the starting points for ATAC-seq and ChIP analyses. This study provides evidence that a genome-wide approach to study GRNs in development and evolution is feasible in echinoderms.

4 Methods

All computational analyses were conducted on the high performance-computing cluster at Michigan State University. Scripts for all the performed analyses are readily available for use and can be found in the following github repository https://github.com/elijahlowe/paraHox_analysis. Snippets of code were generated with the help of biostar and seq-answer online forums [12, 13]. RNA-seq reads will be stored in EBI database.

4.1 Animal Handling and Microinjection Procedures

Adults *S. purpuratus* and *P. miniata* have been obtained from Patrick Leahy (Kerchoff Marine Laboratory, California Institute of Technology, Pasadena, CA, USA), housed in circulating seawater aquaria in the Stazione Zoologica Anton Dohrn of Naples and kept in large tanks of seawater at 15–16 °C.

Microinjection was performed as described in Annunziata and Arnone [9], for sea urchin, and in Cheatle Jarvela and Hinman [14], for sea star. MASOs were obtained from Gene Tools (Corvallis) and injected at the following concentration: 150 μM, for Sp-Lox and Sp-Cdx translation MASOs (sequences as reported in [8] and [9]); 700 μM, for the Pm-Lox translation MASO (sequence 5′-CCAGGGTCATCATGTTCATGTTGGT-3′), and for the Pm-Cdx splicing MASO (sequence 5′-TTGACCTGTAGTTGAAATATGAGAA-3′). For each experiment and for each MASO, 600 zygotes were injected in sea urchin and 50 zygotes in sea star and each experiment was repeated three times with different batches of embryos to obtain three independent biological replicas. As a negative control, the same number of eggs was injected with 100 μM of the standard control morpholino (Gene Tools) and compared side-by-side with uninjected and MASO-injected embryos.

4.2 Embryos Collection, RNA Extraction, and Sequencing

Injected and uninjected sea urchin and sea star fertilized eggs have been allowed to develop until the desired stage at 15 °C in filtered seawater and then collected for the RNA extraction. The embryos have been collected in a tube and centrifuged at 3000 rpm for 2–3 min to remove all the seawater. RNA extraction has been carried out using the RNAqueous-Micro Kit (Ambion). Integrity and quantification of RNA has been checked before the sequencing using the Agilent Bioanalyzer 2100 with the RNA 6000 Pico kit for total eukaryote RNA. cDNA libraries have been prepared with 1 μg of starting total RNA and using the Illumina TruSeq RNA Sample Preparation Kit (Illumina), according to TruSeq protocol. Each library has been diluted to 2 nM and denaturated; 8 pM of each library has been loaded onto cBot (Illumina) for cluster generation with cBot Paired End Cluster Generation Kit (Illumina) and sequenced using the Illumina HiSeq 1500 with 100 bp paired-end reads in triplicate, obtaining 31–38 million reads for replicate. The sequencing service has been provided by the Laboratory of Molecular Medicine and Genomics (http://www.labmedmolge.unisa.it) at the University of Salerno, Italy.

4.3 Quality Control, Mapping, and Differential Expression

Reads were first trimmed using Trimmomatic (v0.33) with the scripts **trim_pm.qsub**, **trim_spcdx.qsub**, and **trim_splox.qsub** [15]. The parameters for trimming were chosen to efficiently remove erroneous reads while maximizing the information within the reads [16]. *S. purpuratus* reads were mapped to Genome sequence (V3.1) [17] and *P. miniata* reads were mapped to the genome sequence (V1.0) Scaffolds [18] using Bowtie2 (2.2.6) and Tophat (2.0.8b) [19, 20]. After mapping, reads were sorted using SamTools (v1.2) [21] and counts were extracted using

HTSeq (v0.6.1) [22]. The gff3 from Build 7 was used for generating exon-based transcript counts for *S. purpuratus* which is more informative seeing than DESeq2 does not use length-based count normalization [23, 24]. The following scripts were used **sp_cdx.qsub**, **sp_lox48.qsub**, and **sp_lox72.qsub** for *Sp*, while **pm_cdx.qsub** and **pm_lox.qsub** were used for *Pm*.

Differentially expressed genes were identified using DESeq2 [23], transcripts not meeting the threshold of 10 counts for at least one of the samples were removed. DESeq2 provides two methods of hypothesis testing: Wald test and likelihood ratio test (LRT). To account for the batch effect across different animals we used LRT, with the full model being batch + condition and the reduced model being batch. After, the differentially expressed genes using extracted information from Echinobase [18] for both species, which are in the **data/** directory, using annot_sp.py and annot_pm.py scripts.

4.4 Identification and Clustering of Orthologs

The proteomes for *S. purpuratus* (SPU_peptide sequence) and *P. miniata* (PMI_protein sequence) were downloaded from echinobase (http://www.echinobase.org/Echinobase/SpDownloads and http://www.echinobase.org/ Echinobase/PmDownload) while *Xenopus tropicalis* proteome (release 83) was downloaded from Ensembl (ftp://ftp.ensembl.org/pub/release-83/fasta/xenopus_tropicalis/pep/) in fasta format [18, 25]. Orthology was determined using orthoMCL [26]. Sequences of the three proteomes were concatenated into one file and transformed into orthoMCL format, so an all-vs-all protein blast search was conducted using the blastp program in the BLAST+ (v2.2.30) suite [27]. Prior to the blast search, sequences with stop codons and of a length shorter than 20 amino acids were removed. A protein blast (blastp) was performed using the concatenated fasta as the query and database. Blast results were then parsed, loaded into a MySQL database, and then proteins with at least 50 % similarity were clustered through the use of orthoMCL programs. The steps for orthoMCL are in the script ortho.qsub.

4.5 Transcription Factor Identification

Both the *S. purpuratus* and *P. miniata* proteomes were searched against the Pfam database [28] using HMMER/3.1b2 hmmscan [29]. These commands were executed using the following scripts hmmer_pm_tf.qsub and hmmer_spur_tf.qsub. The grep program was then used to search for the following term homeobox, Pax, bzip, hmg, sox, hlh, PF00104.25 (nuclear receptor), t-box, mh2 (smad), b-box, f-box, fork_head, ets, phd-finger, zf-C2H2 within –tblout output. Additionally, Pfam ids were extracted from the DBD Transcription Factor prediction database [30] and then grep against the –tblout output, combined filtered for redundancy. The list of Pfam ids can be found in the data directory in the github repository along with the TF we identified for *S. purpuratus* and *P. miniata*.

Acknowledgements This work was supported in part by Michigan State University through computational resources provided by the Institute for Cyber-Enabled Research and by MIUR (premiale PANTRAC to MIA). C.C. has been supported by a SZN PhD fellowship.

References

1. Wang, Z., Dai, M., Wang, Y., Cooper, K.L., et al.: Unique expression patterns of multiple key genes associated with the evolution of mammalian flight. Proc. Biol. Sci. **281**(1783), 20133133 (2014)
2. Lmanna, F., Kirschbaum, F., Waurick, I., Dieterich, C., Tiedemann, R.: Cross-tissue and cross-species analysis of gene expression in skeletal muscle and electric organ of African weakly-electric fish (Teleostei; Mormyridae). BMC Genomics **16**, 668 (2015). doi:10.1186/s12864-015-1858-9
3. Finnerty, J.R.: The origins of axial patterning in the metazoa: how old is bilateral symmetry? Int. J. Dev. Biol. **47**(7–8), 523–529 (2003)
4. Mallo, M., Alonso, C.R.: The regulation of hox gene expression during animal. Development **140**(19), 3951–3963 (2013)
5. Brooke, N.M., Garcia-Fernandez, J., Holland, P.W.: The ParaHox gene cluster is an evolutionary sister of the Hox gene cluster. Nature **392**, 920–922 (1998)
6. Wright, C.V., Cho, K.W., Oliver, G., De Robertis, E.M.: Vertebrate homeodomain proteins: families of region-specific transcription factors. Trends Biochem. Sci. **14**, 52–56 (1989)
7. Young, T., Deschamps, J.: Hox, Cdx, and anteroposterior patterning in the mouse embryo. Curr. Top. Dev. Biol. **88**, 235–255 (2009)
8. Cole, A.G., Rizzo, F., Martinez, P., Fernandez-Serra, M., Arnone, M.I.: Two ParaHox genes, SpLox and SpCdx, interact to partition the posterior endoderm in the formation of a functional gut. Development **136**, 541–549 (2009)
9. Annunziata, R., Arnone, M.I.: A dynamic regulatory network explains ParaHox gene control of gut patterning in the sea urchin. Development **141**(12), 2462–2472 (2014). doi:10.1242/dev.105775
10. Arnone, M.I., Rizzo, F., Annunciata, R., Cameron, R.A., Peterson, K.J., Martínez, P.: Genetic organization and embryonic expression of the ParaHox genes in the sea urchin S. purpuratus: insights into the relationship between clustering and collinearity. Dev. Biol. **300**, 63–73 (2006)
11. Annunziata, R., Martinez, P., Arnone, M.I.: Intact cluster and chordate-like expression of ParaHox genes in a sea star. BMC Biol. **11**, 68 (2013). http://www.biomedcentral.com/1741-7007/11/68
12. Parnell, L.D., Lindenbaum, P., Shameer, K., Dall'Olio, G.M., Swan, D.C., et al.: BioStar: an online question & answer resource for the bioinformatics community. PLoS Comput. Biol. **7**(10), e1002216 (2011)
13. Li, J.W., Schmieder, R., Ward, R.M., Delenick, J., Olivares, E.C., Mittelman, D.: SEQanswers: an open access community for collaboratively decoding genomes. Bioinformatics **28**(9), 1272–1273 (2012)
14. Cheatle Jarvela, A.M., Hinman, V.: A method for microinjection of Patiria miniata zygotes. J. Vis. Exp. (91), e51913 (2014). doi:10.3791/51913
15. Bolger, A.M., Lohse, M., Usadel, B.: Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics **30**, 2114–2120 (2014)

16. MacManes, M.D.: On the optimal trimming of high-throughput mRNAseq data. bioRxiv (2014). doi: 10.1101/000422
17. Sodergren, E., Weinstock, G.M., Davidson, E.H., Cameron, R.A., Gibbs, R.A., Angerer, R.C., Coffman, J.A.: The genome of the sea urchin Strongylocentrotus purpuratus. Science **314**(5801), 941–952 (2006)
18. Cameron, R.A., Samanta, M., Yuan, A., He, D., Davidson, E.: SpBase: the sea urchin genome database and web site. Nucleic Acids Res. **37**, D750–D754 (2009)
19. Langmead, B., Salzberg, S.L.: Fast gapped-read alignment with Bowtie 2. Nat. Methods **9**(4), 357–359 (2012)
20. Kim, D., Pertea, G., Trapnell, C., Pimentel, H., Kelley, R., Salzberg, S.L.: TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. Genome Biol. **14**(4), R36 (2013)
21. Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., Durbin, R., 1000 Genome Project Data Processing Subgroup: The Sequence alignment/map (SAM) format and SAMtools. Bioinformatics **25**, 2078–2079 (2009)
22. Anders, S., Pyl, P.T., Huber, W.: HTSeq — a Python framework to work with high-throughput sequencing data. Bioinformatics **31**, 166–169 (2014). doi:10.1093/bioinformatics/btu638
23. Love, M.I., Huber, W., Anders, S.: Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol. **15**, 550 (2014). doi:10.1186/s13059-014-0550-8
24. Zhao, S., Xi, L., Zhang, B.: Union exon based approach for RNA-seq gene quantification: to be or not to be? PLOS One (2015). doi:10.1371/journal.pone.0141910
25. Cunningham, F., Amode, M.R., Barrell, D., Beal, K., Billis, K., Brent, S., Carvalho-Silva, D., Clapham, P., Coates, G., Fitzgerald, S., Gil, L., Girón, C.G., Gordon, L., Hourlier, T., Hunt, S.E., Janacek, S.H., Johnson, N., Juettemann, T., Kähäri, A.K., Keenan, S., Martin, F.J., Maurel, T., McLaren, W., Murphy, D.N., Nag, R., Overduin, B., Parker, A., Patricio, M., Perry, E., Pignatelli, M., Riat, H.S., Sheppard, D., Taylor, K., Thormann, A., Vullo, A., Wilder, S.P., Zadissa, A., Aken, B.L., Birney, E., Harrow, J., Kinsella, R., Muffato, M., Ruffier, M., Searle, S.M.J., Spudich, G., Trevanion, S.J., Yates, A., Zerbino, D.R., Flicek, P.: Ensembl 2015. Nucleic Acids Res. **43**(Database issue), D662–D669 (2015). doi:10.1093/nar/gku1010
26. Fischer, S., Brunk, B.P., Chen, F., Gao, X., Harb, O.S., Iodice, J.B., Shanmugam, D., Roos, D.S., Stoeckert Jr., C.J.: Using OrthoMCL to assign proteins to OrthoMCL-DB groups or to cluster proteomes into new ortholog groups. Curr. Protoc. Bioinformatics. Chapter 6:Unit 6.12.1–19 (2011)
27. Camacho, C., Coulouris, G., Avagyan, V., Ma, N., Papadopoulos, J., Bealer, K., Madden, T.L.: BLAST+: architecture and applications. BMC Bioinformatics **10**, 421 (2008)
28. Finn, R.D.: Pfam: the protein families database. Encyclopedia of Genetics, Genomics, Proteomics and Bioinformatics (2012)
29. Durbin, R., Eddy, S.R., Krogh, A., Mitchison, G.: Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids. Cambridge University Press (1998). ISBN 0-521-62971-3
30. Wilson, D., Charoensawan, V., Kummerfeld, S.K., Teichmann, S.A.: DBD - taxonomically broad transcription factor predictions: new content and functionality. Nucleic Acids Res. **36**, D88–D92 (2008). doi:10.1093/nar/gkm964

Reconstructing a Genetic Network from Gene Perturbations in Secretory Pathway of Cancer Cell Lines

Marina Piccirillo, Kumar Parijat Tripathi, Sonali Gopichand Chavan, Seetharaman Parashuraman, Alessandra Varavallo, and Mario Guarracino

Abstract Gene perturbation studies play an important role in the reconstruction of genetic networks and in determining the influence of genes on each other activities. According to this hypothesis, we planned to develop new analysis methods, based on novel algorithms, to reconstruct genetic networks by incorporating gene expression datasets, containing profiles of cell lines that have been exposed to genetic perturbations. In the present work, we focus on a list of genes, localized in secretory pathway. These genes and their products are responsible for the delivery of different kind of proteins from their site of synthesis to their proper cellular location and they are essential for cellular functions and multicellular development. Using data from high-throughput experiments, gene expression profiles are collected from 33 genes perturbations (knockdown and over-expressed) experiments in four cancer cell lines. Data have been downloaded from the Library of Integrated Network-Based Cellular Signatures. We characterized gene regulatory networks of secretory pathway, and we provided some empirical results of the network modular organization. The interesting observation is that all these regulatory genes are also connected with each other through hub nodes. It means that interactions do not have a separate entity and are not regulated by independent behavior of perturbed genes, but probably, there is a global effect of all these perturbations on all subnetworks present in an interaction network.

Keywords Perturbation • Genetic network • Algorithm • Secretory pathway

M. Piccirillo (✉) • K.P. Tripathi • M. Guarracino
Laboratory for Genomics, Transcriptomics and Proteomics (Lab-GTP), High Performance Computing and Networking Institute (ICAR), National Research Council (CNR), Via Pietro Castellino 111, Naples, Italy
e-mail: marina.piccirillo@icar.cnr.it

S.G. Chavan • S. Parashuraman • A. Varavallo
Institute of Protein Biochemistry, National Research Council (CNR-IBP), Via Pietro Castellino 111, Naples, Italy

A. Rogato et al. (eds.), *Dynamics of Mathematical Models in Biology*,
DOI 10.1007/978-3-319-45723-9_6

1 Introduction

Genetic perturbations are experimental alterations of gene activity, by manipulating either the gene itself or its products. Such perturbations include point mutations, gene deletions, over-expression, or any other interference with the activity of the genes or their product. They can be used in conjunction with a reverse engineering algorithm to reconstruct and reveal the architecture of a gene regulatory network (GRN), by analyzing the steady-state changes in gene expression of a particular node in the network [9]. GRNs are the most important abstract organizational level in the cell, because they symbolize the signals that cells receive, in terms of activation and inhibition of genes [1], as shown in Fig. 1 (image extracted 14 May 2007 from the http://genomics.energy.gov website from the U.S. Department of Energy Genome Programs). GRNs are represented as graphs, in which nodes are genes, proteins, or metabolites and edges are the relations between the nodes; therefore it is possible to understand the molecular mechanism of each gene by identifying their interactions within the GRNs [6]. Our aim is to study GRNs from gene expression data, to identify direct and indirect interactions among genes, and to reconstruct the characteristics of the secretory pathway [2]. We chose to study secretory pathway because experimental evidence indicates that endoplasmic reticulum (ER) and Golgi apparatus can activate both survival mechanisms and cell suicide programs if the stress-signaling threshold is exceeded. Furthermore it is possible that the fragile balance of protein trafficking between various subcellular compartments provides a good therapeutic opportunity [11]. Several techniques have been proposed to analyze large data sets from whole-genome networks, such as cluster analysis and enrichment analysis, but they typically provide only indirect information about network structure [7].

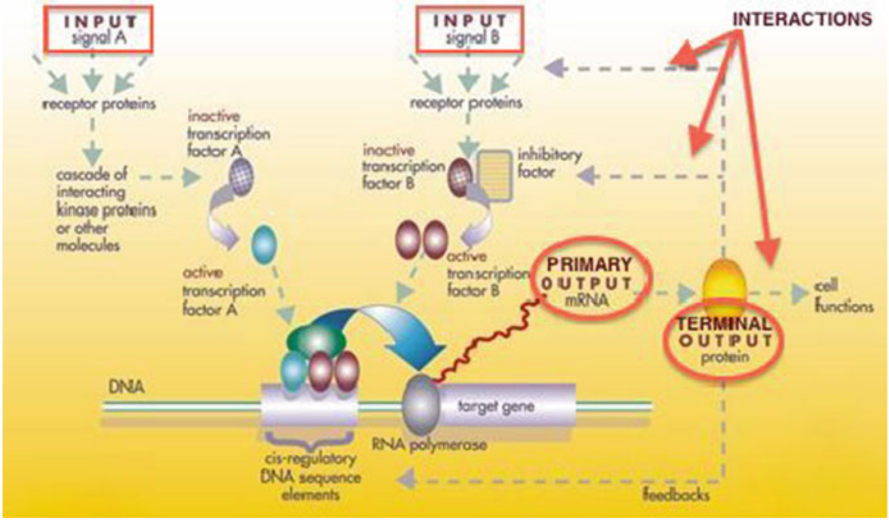

INPUT signal A
receptor proteins
cascade of interacting kinase proteins or other molecules
inactive transcription factor A
active transcription factor A
INPUT signal B
receptor proteins
inactive transcription factor B
inhibitory factor
active transcription factor B
INTERACTIONS
PRIMARY OUTPUT mRNA
TERMINAL OUTPUT protein
cell functions
DNA
target gene
RNA polymerase
cis-regulatory DNA sequence elements
feedbacks

Fig. 1 A gene regulatory network

2 Material and Methods

2.1 Data Retrieval

The secretory pathway has evolved to facilitate the transfer of cargo molecules to internal and cell-surface membranes [8]. Its study and characterization are a challenge, that the use of high-throughput experiments and network analysis tools have enabled to outdo. In this work, we try to reconstruct the regulatory networks of secretory pathways starting from 476,251 signatures and 22,268 probes present in the LINCS website (http://www.lincscloud.org/); selecting the gene expression profile data related to 33 gene perturbation experiments carried out in four cancer cell lines (A549, HA1E, HEPG2, and PC3). In Mitocheck database (http://www.mitocheck.org/), these latter are classified as mild or strong inhibitors of secretory cargo proteins from ER to plasma membrane and they are involved in the morphological alterations of COPII and/or COPI vesicular coat complexes. Then, for each expression profile we collected two or more technical replicates at different time points, considering profiles of differentially expressed genes computed by robust z-scores for each profile relative to population control. The genes are shown in Fig. 2, where we see the 33 perturbations with all the functions in which they are involved.

2.2 Reconstruction of Regulatory Interactions

To reconstruct regulatory interactions in gene networks from gene perturbation experiments, we developed a Python computational pipeline, which is divided into four steps:

1. For each cell line and for each perturbation, we computed the mean among the technical replicates of each biological replicate.

   Let $c_i \in C$ a set of cell lines, with i $= 1, 2, 3, 4$. Perturbation experiments $p^i_j \in P$ with j $= p1, p2, \dots p33$ have biological replicate represented as $b^{ij}_l \in B$ with l $= 1, \dots L_{ij}$. In turn, each biological replicate has technical replicate t^{ijl}_k with k $= 1, ..k_{ijl}$. The mean m for each perturbation p for a given cell line i is calculated as:

   - $b^{ij}_l = \sum_{k=1}^{k} t^{ijl}_k / k_{pijl}$, where k is the number of technical replicates for perturbation experiments.

2. In the second step, we created a matrix M; whose columns are the biological replicates, and then we calculated the first principal component pc_{ij}, which is the linear combination of x-variables along the direction of maximum variance.

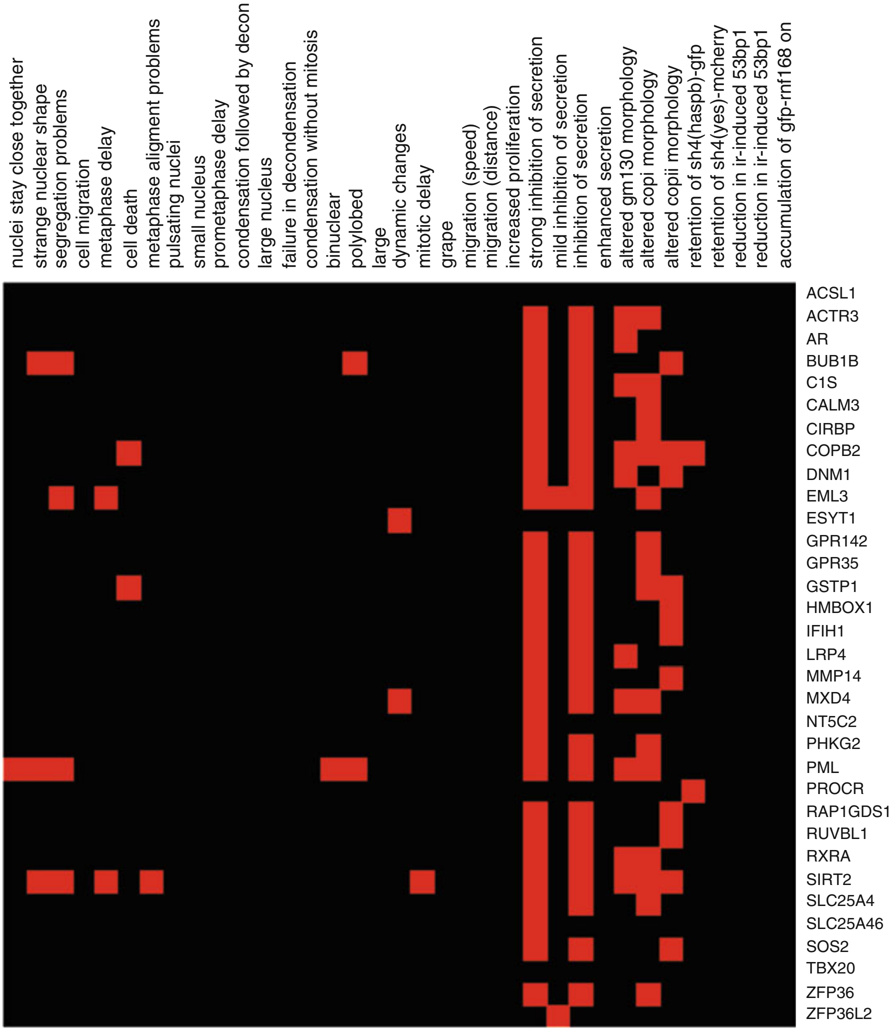


Fig. 2 List of 33 secretion regulators

3. For each cell line and for each perturbation, we determined the biological replicate $\bar{b}_l^{ij}$ among b_l^{ij} with maximum correlation, and then we constructed a matrix with them. The matrix $A^i = \bar{b}^{ij}$ represents the influence of perturbation on the expression values of all the probes in the experiments.
4. For each A^i and for each perturbation, we selected only those probes, for which we observed a fold change greater than 4 in case of over expression genes and decrease more than 4 in under expression genes, in at least one $\bar{b}^{ij}$.

2.3 Networks Reconstructions and Enrichment Analysis

The molecular interaction networks, for each cell line, were studied using the network visualization software Cytoscape [5]. To reconstruct the characteristics of the secretory pathway, we performed an enrichment analysis using the Molecular Signature Database (MsigDB) (http://www.broadinstitute.org/gsea/downloads.jsp).[1] We used Kyoto Encyclopedia of Genes and Genomes (KEGG) and BIOCARTA pathway gene set to study the enriched pathways in the networks. These tools reduce the complexity of analysis by grouping long lists of individual genes into smaller sets of related genes or proteins, that are involved in the same biological processes, components, or structures [4].

3 Results

3.1 Networks Reconstructions

From the reconstruction described in the previous section, we obtained four networks as shown in Figs. 3, 4, 5, and 6.

In A375 cancer cell line network there are 679 nodes and 1094 edges, while A549 cancer cell line contains 523 nodes and 978 edges. Instead we can distinguish 1140 nodes and 1609 edges in HA1E and 397 nodes with 665 edges in HEPG2 network, respectively. The obtained networks portray the direct and indirect interactions among genes, as well as the regulatory effects that perturbations have with other interaction partner genes. As we can see some perturbations, don't have independent behavior, but a combinatorial effect on transcriptional regulation.

3.2 Role of Secretion Regulators in Different Cellular Processes

Over the last years with the new high-throughput imaging-based methods, and more recently, with RNA interference (RNAi)-mediated gene knockdown experiments, a significant number of regulators associated with the secretory pathway have been revealed. In this study, we used a computational approach to try to find these regulators of secretory pathway. Comparing our networks with a list of secretion inhibitors involved in cell death, cell division, and motility; which we selected from a precedent study [8], we found that some perturbations regulate several of these inhibitors in each network as shown in Tables 1 and 2. An enrichment

[1] http://software.broadinstitute.org/gsea/msigdb/annotate.jsp.

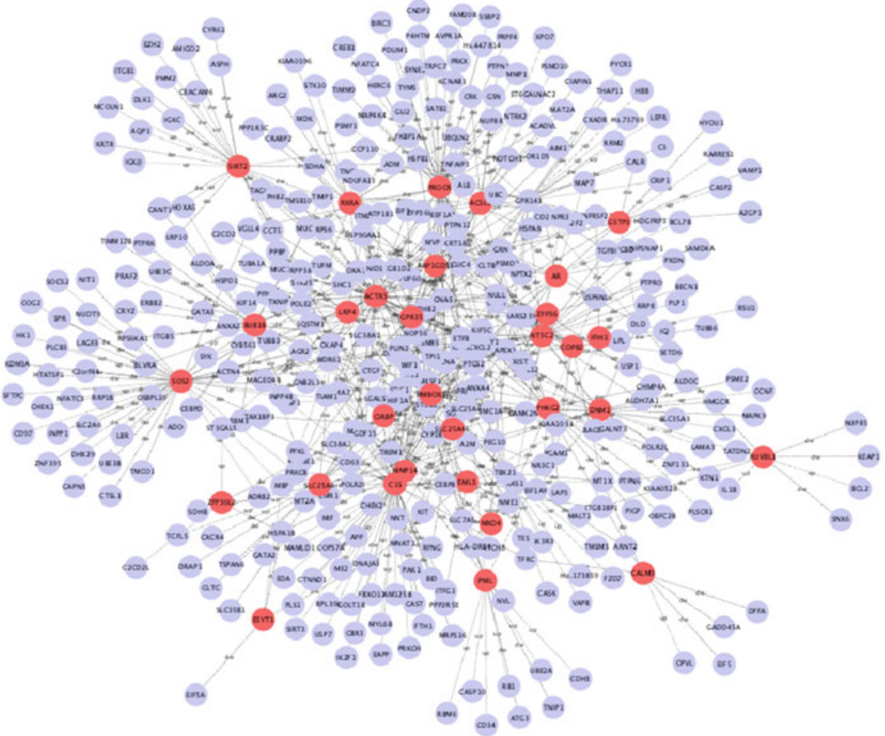


Fig. 3 Regulatory network in A375 cancer cell line. In *red* are depicted the 33 perturbations

analysis of the genes showed that more than 70 % of them participate in fundamental cellular processes such as transcription, explaining how their knockdown, causes cell death and therefore the transport inhibition. In particular, we put our focus on the membrane traffic regulators, such as COPB2, which encode for subunit beta of the Golgi coatomer complex and whose depletion causes cell death, thus underlining the importance of the secretory pathway for general cell health. This is a perturbation that is indirectly down-regulated from other perturbations in A375, A549, and HA1E networks. For example, in A549 cancer cell line, COPB2 is indirectly connected with PML and EML3 perturbations. See Fig. 7.

A crucial observation is that in HA1E network, COPB2 is connected directly with RUVBL1, which is located in Golgi apparatus and vesicles, and both regulate some common genes. Membrane traffic pathways are also regulated through the activities of kinases and phosphatases, and some of these are involved in ER–Golgi recycling. Overlapping, our regulatory interactions with a list of 48 genes, which are scored as secretion inhibitors in the study of Farhan et al. [3], we found that, in each network, there are some of these genes that are regulated from the perturbations (see Table 3). With respect to HEPG2 and A549 networks, two perturbations are connected indirectly through EPBHB2 of ER receptor tyrosine kinase, involved in axon guidance.

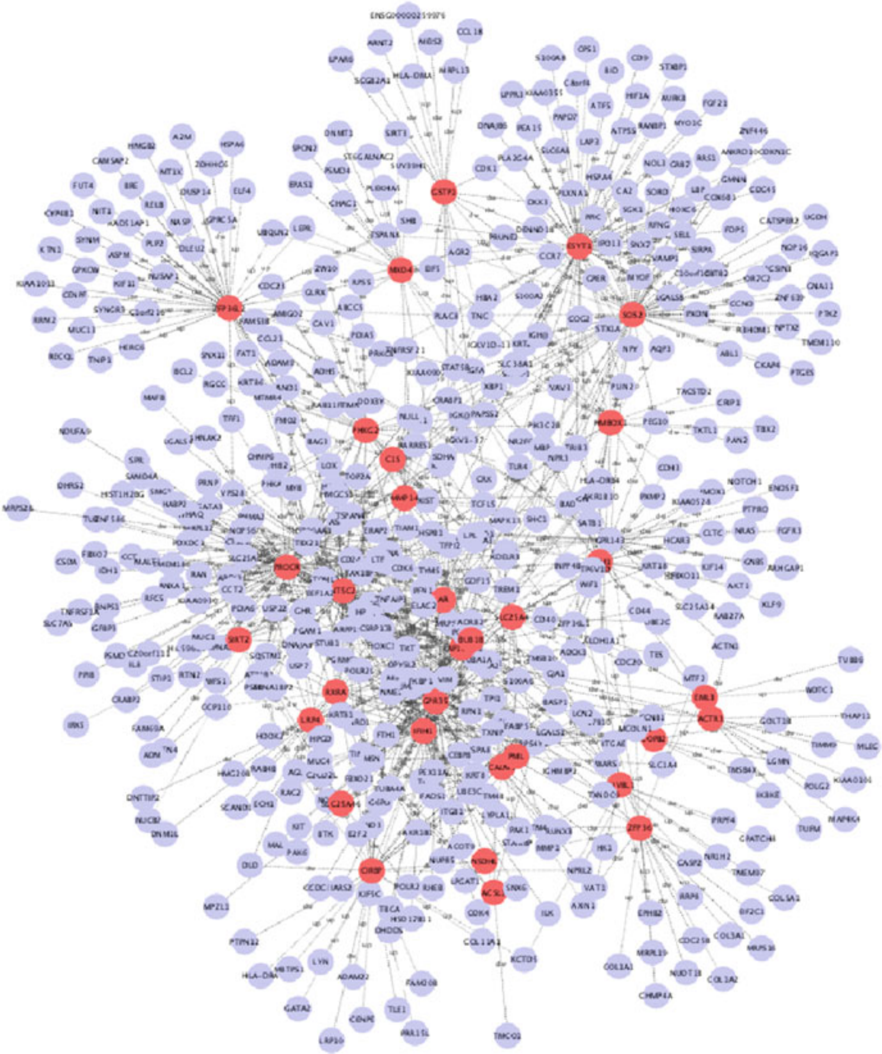


Fig. 4 Regulatory network in A549 cancer cell line

4 Discussions and Conclusions

GRNs represent a combination of diverse regulation and interaction mechanisms operating in different conditions and time scales. We integrated such data to try to describe with a computational approach the characteristics of secretory pathway. We applied here the proposed algorithm to gene expression's data of 33 perturbed genes in four cancer cell lines, and we reconstructed GRNs, providing also better understanding of cellular response towards chemical and genetic perturbations. A

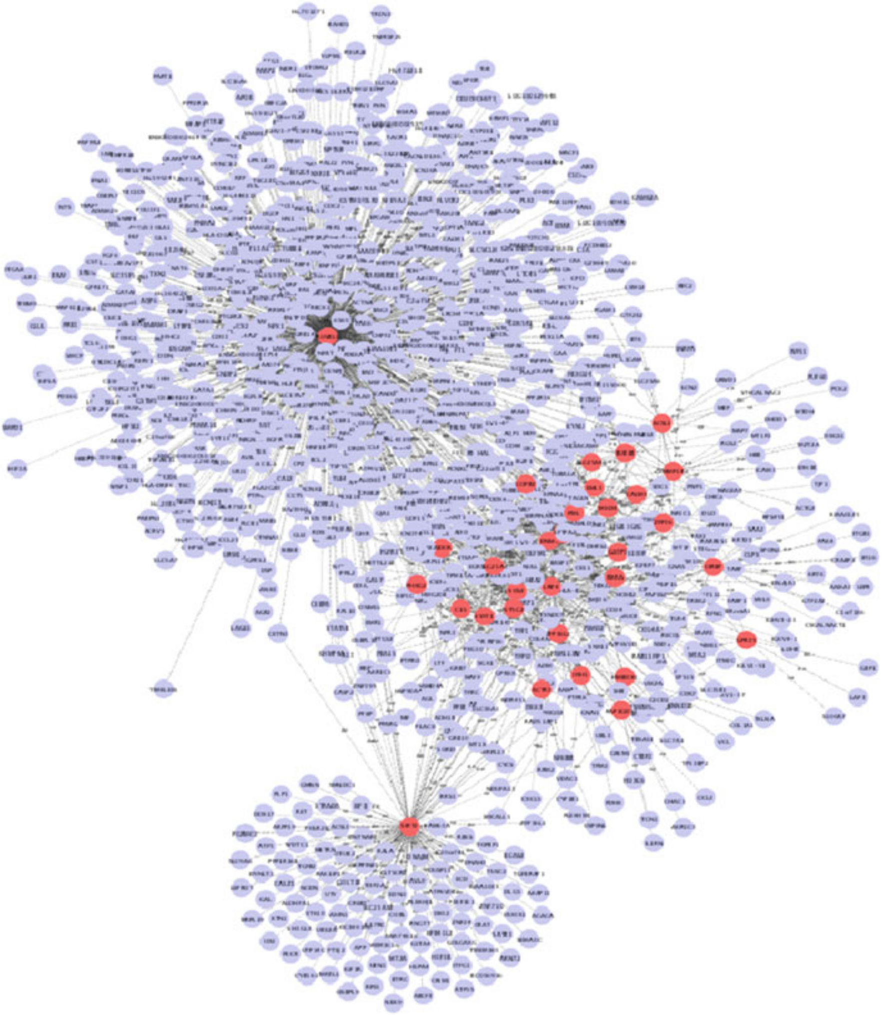


Fig. 5 Regulatory network HA1E cancer cell line

main result of the proposed procedure is that all these regulatory genes are also connected with each other through hub nodes. It means that the transcriptional response with respect to each perturbations does not have independent behavior, but somehow these perturbations put a combinatorial effect on transcriptional regulation, perhaps there is a global effect of all these perturbations on all subnetworks present in an interaction network. Our analyses indicate that the perturbations control some genes, which are involved in several processes of secretory pathway regulations. For example, depletion of the Golgi coatomer complex, COPB2 which is essential for Golgi budding and vesicular trafficking, caused cell death. Furthermore, we

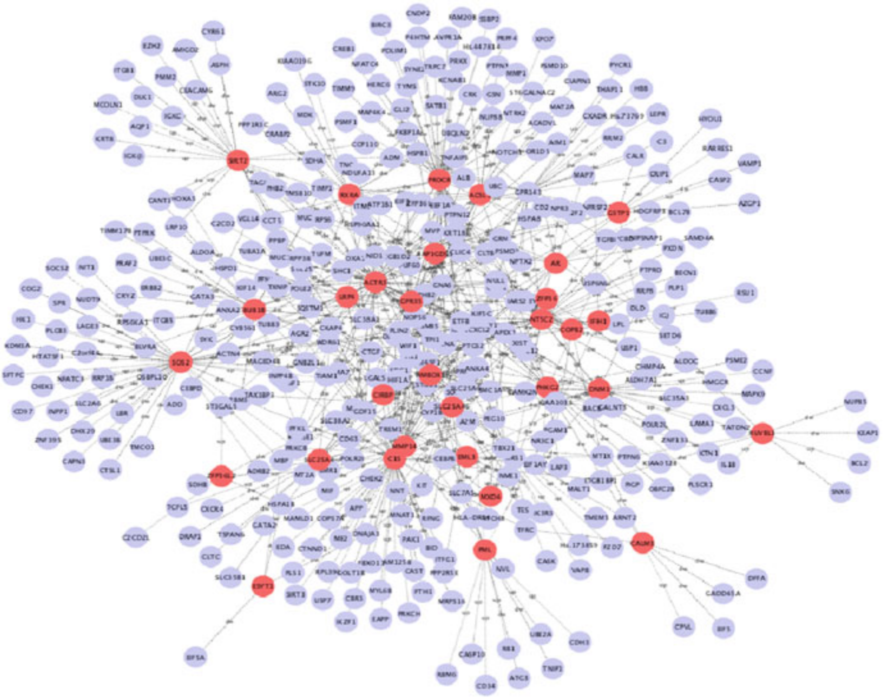


Fig. 6 Regulatory network HEPG2 cancer cell line

also found kinases and phosphatases, that can regulate membrane traffic pathways. In particular we found that the secretion inhibitors EPHB2 indirectly connect two perturbations in HEPG2 and A549 network. Together, these results imply that the mammalian cells have a highly sophisticated signaling and feedback system that allows them to modulate their secretory activity in response to external signals and their local environment. So our algorithm may help answer a multitude of questions about the genetic architecture of organisms. What is the structure of genetic networks? How do patterns of interactions genes change in different developmental stages, in different physiological states, in different environmental conditions, or in different cell types? Are there many genes that do not affect the activity of other genes? This approach could be useful also in determining the potential drug targets in case of aggressive human tumors. Furthermore, it is possible to cluster the mechanisms of action rather than the gene expression pattern; and moreover our algorithm is very scalable in term of the number of experiments used to a given network. For the future work by taking into account of existing algorithms and inference methods for reverse engineering of gene networks from large scale gene expression data, we plan to deepen the study of these techniques and overcome their limits; indeed our next goal will be to implement Wagner's algorithm [10], which is important to find short cycles and loops in the network.

Table 1 List of secretion inhibitors involved in cell death, cell division, and motility in A375 and A549 cells

A375			A549		
Gene symbol	Ensembl ID	Cell division phenotypes	Gene symbol	Ensembl ID	Cell division phenotypes
ACADVL	ENSG00000072778	Mitosis, cell death	AKT1	ENSG00000142208	Other phenotypes
BMPR1A	ENSG00000107779	Mitosis, Other phenotypes	BUB1B	ENSG00000156970	Mitosis
BUB1B	ENSG00000156970	Mitosis	CDC23	ENSG00000094880	Mitosis
CDK4	ENSG00000135446	Other phenotypes	CDK4	ENSG00000135446	Other phenotypes
CHN1	ENSG00000128656	Other phenotypes	COPB2	ENSG00000184432	Cell death
CLIC4	ENSG00000169504	Other phenotypes	EML3	ENSG00000149499	Mitosis
COPB2	ENSG00000184432	Cell death	GSTP1	ENSG00000084207	Cell death
DNAL4	ENSG00000100246	Other phenotypes	IDH1	ENSG00000138413	Mitosis
EML3	ENSG00000149499	Mitosis	MXD4	ENSG00000123933	Other phenotypes
GJB3	ENSG00000188910	Mitosis	NOL3	ENSG00000140939	Mitosis
GRWD1	ENSG00000105447	Cell death	PML	ENSG00000140464	Mitosis
GSTP1	ENSG00000084207	Cell death	SAMD4A	ENSG00000020577	Other phenotypes
KIF2C	ENSG00000142945	Mitosis	SIRT2	ENSG00000068903	Mitosis
MXD4	ENSG00000123933	Other phenotypes	ST6GALNAC2	ENSG00000122912	Other phenotypes
NFKBIE	ENSG00000146232	Mitosis	TBCA	ENSG00000171530	Cell death
NUP93	ENSG00000102900	Cell death	TRIB3	ENSG00000101255	Migration
PLCB2	ENSG00000137841	Mitosis	TXNDC9	ENSG00000115514	Mitosis
PML	ENSG00000140464	Mitosis	TYMS	ENSG00000176890	Cell death
PPOX	ENSG00000143224	Other phenotypes			
PRSS23	ENSG00000150687	Other phenotypes			
RPA1	ENSG00000132383	Mitosis, cell death			
SCYL3	ENSG00000000457	Other phenotypes			
SIRT2	ENSG00000068903	Mitosis			
SLC16A3	ENSG00000141526	Other phenotypes			
SLC25A16	ENSG00000122912	Other phenotypes			
ST6GALNAC2	ENSG00000122912	Other phenotypes			
TXNDC9	ENSG00000115514	Mitosis			

Table 2 List of secretion inhibitors involved in cell death, cell division, and motility in HA1E and HEPG2 cells

HA1E			HEPG2		
Gene symbol	Ensembl ID	Cell division phenotypes	Gene symbol	Ensembl ID	Cell division phenotypes
AMHR2	ENSG00000135409	Other phenotypes	ACAD VL	ENSG00000072778	Mitosis, cell death
BUB1B	ENSG00000156970	Mitosis	ALDOA	ENSG00000149925	Mitosis
C10orf68	ENSG00000150076	Other phenotypes	BUB1B	ENSG00000156970	Mitosis
CDK5R1	ENSG00000176749	Mitosis, migration, other phenotypes	CLIC4	ENSG00000169504	Other phenotypes
CER1	ENSG00000147869	Cell death	COPB2	ENSG00000184432	Cell death
CLCNKB	ENSG00000184908	Mitosis	EML3	ENSG00000149499	Mitosis
CLIC4	ENSG00000169504	Other phenotypes	GSTP1	ENSG00000084207	Cell death
COPB2	ENSG00000184432	Cell death	IDH1	ENSG00000138413	Mitosis
ECD	ENSG00000122882	Mitosis	ITGB5	ENSG00000082781	Migration, other phenotypes
EEF1E1	ENSG00000124802	Migration	MXD4	ENSG00000123933	Other phenotypes
EML3	ENSG00000149499	Mitosis	MYL6B	ENSG00000196465	Mitosis
GRWD1	ENSG00000105447	Cell death	PML	ENSG00000140464	Mitosis
GSTP1	ENSG00000084207	Cell death	SAMD 4A	ENSG00000020577	Other phenotypes
KCNQ4	ENSG00000117013	Other phenotypes	SIRT2	ENSG00000068903	Mitosis
MXD4	ENSG00000123933	Other phenotypes	ST6GAL NAC2	ENSG00000122912	Other phenotypes
NBR1	ENSG00000188554	Cell death	TAGLN	ENSG00000149591	Cell death
OGG1	ENSG00000114026	Mitosis	TYMS	ENSG00000176890	Cell death
PLA2G3	ENSG00000138308	Cell death	USP1	ENSG00000162607	Mitosis, migration
PML	ENSG00000140464	Mitosis			
PPP2R1A	ENSG00000105568	Mitosis			
ROS1	ENSG00000047936	Other phenotypes			
RRM1	ENSG00000167325	Other phenotypes			
SAMD4A	ENSG00000020577	Other phenotypes			
SCN5A	ENSG00000183873	Mitosis			
SIRT2	ENSG00000068903	Mitosis			
ST6GAL NAC2	ENSG00000122912				
TAGLN	ENSG00000149591	Cell death			
TRIB3	ENSG00000101255	Migration			
TXNDC9	ENSG00000115514	Mitosis			
TYMS	ENSG00000176890	Cell death			

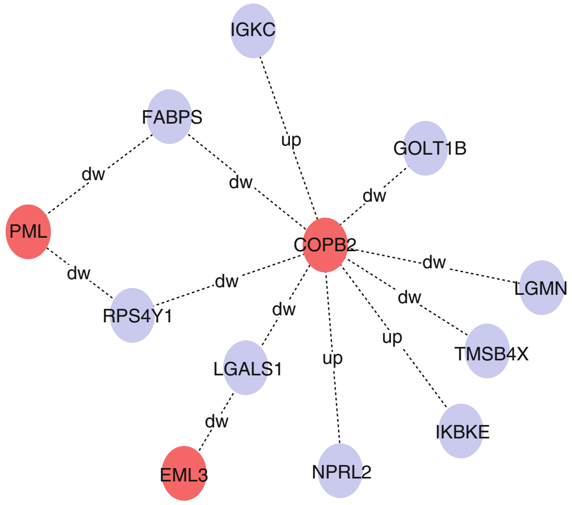


Fig. 7 Sub-network of A549 cancer cell line, in which we can see all the first neighbors nodes of COPB2 and its indirect interactions with other perturbations depicted in *red*

Table 3 List of kinases and phosphatases which are regulated from our perturbations in each network

Farhan et al. Class	Gene symbol	Gene bank	Description
A375			
Golgi	ABL1	NM005157	v-Abl Abelson murine leukemia viral oncogene homolog 1
Golgi	AURKB	NM004217	Aurora kinase B
Golgi	CDK4	NM000075	Cyclin-dependent kinase 4
A549			
Golgi	ABL1	NM005157	v-Abl Abelson murine leukemia viral oncogene homolog 1,
Golgi	AURKB	NM004217	Aurora kinase B
Golgi	CDK4	NM000075	Cyclin-dependent kinase 4
ER	EPHB2	NM017449	EPH receptor B2
Golgi	KIT	NM000222	v-Kit Hardy-Zuckerman 4 feline sarcoma viral oncogene homolog
HA1E			
Golgi	ABL1	NM005157	v-Abl Abelson murine leukemia viral oncogene homolog 1
ER	EGFR	NG007726	Epidermal growth factor receptor
ER	IKBKB	NM001556	Inhibitor of kappa light polypeptide gene enhancer in B-cells, kinase beta
Golgi	KIT	NM000222	v-Kit Hardy-Zuckerman 4 feline sarcoma viral oncogene homolog
HEPG2			
ER	EPHB2	NM017449	EPH receptor B2
Golgi	KIT	NM000222	v-Kit Hardy-Zuckerman 4 feline sarcoma viral oncogene homolog

References

1. Crombach, A., Hogeweg, P.: Evolution of evolvability in gene regulatory networks. PLoS Comput. Biol. **4**, e1000112 (2008)
2. Farhan, H., Rabouille, C.: Signalling to and from the secretory pathway J. Cell Sci. **124**, 171–180 (2011)
3. Farhan, H., et al.: MAPK signaling to the early secretory pathway revealed by kinase/phosphatase functional screening. J. Cell Biol. **189**, 997–1011 (2010)
4. Khatri, P., Sirota M., Butte A.J.: Ten years of pathway analysis: current approaches and outstanding challenges. PLoS Comput. Biol. **8**, e1002375 (2012)
5. Kohl, M., Wiese, S., Warscheid, B.: Cytoscape: software for visualization and analysis of biological networks. Methods Mol. Biol. **696**, 291–303 (2011)
6. Liu, L.-Z., Wu, F.-X., Zhang, W.-J.: Reverse engineering of gene regulatory networks from biological data. WIREs Data Min. Knowl. Discovery **2**, 365–385 (2012)
7. Marbach, D., Costello, J.C., Küffner, R., et al.: Wisdom of crowds for robust gene network inference. Nat. Methods **9**, 796–804 (2012)
8. Simpson, J.C., Joggerst, B., Laketa, V., et al.: Genome-wide RNAi screening identifies human proteins with a regulatory function in the early secretory pathway. Nat. Cell Biol. **14**, 764–774 (2012)
9. Tegner, J., Yeung, M.K., Hasty, J., Collins, J.J.: Reverse engineering gene networks: integrating genetic perturbations with dynamical modeling. Proc. Natl. Acad. Sci. U S A **100**, 5944–5949 (2003)
10. Wagner, A.: How to reconstruct a large genetic network from n gene perturbations in fewer than n2 easy steps. Bioinformatics **17**, 1183–1197 (2001)
11. Wlodkowic, D., Skommer, J., McGuinness, D., Hillier, C., Darzynkiewicz, Z.: ER-Golgi network–a future target for anti-cancer therapy. Leuk. Res. **33**, 1440–1447 (2009)

Dissecting the Functions of the Secretory Pathway by Transcriptional Profiling

Sonali Gopichand Chavan, Kumar Parijat Tripathi, Marina Piccirilo, Prathyush Deepth Roy, Mario Guarracino, Alberto Luini, and Seetharaman Parashuraman

Abstract The secretory pathway is responsible for biosynthesis, processing, sorting, and delivery of variety of proteins encoded in the human genome to their proper cellular location through a series of controlled events. Each step along the secretory pathway is controlled by regulatory modules that maintain the homeostasis of the system. Impairment in the functioning of the secretory pathway forms the basis of several pathologies. Nevertheless, the modules that interact with the secretory pathway and the underlying molecular circuits remain under explored, especially in the mammalian system. In order to identify and characterize these circuits we are implementing an approach based on the deconvolution of the transcriptional profiles resulting from the perturbation of the secretory pathway. Such analysis will help to detect the cellular modules that interact with secretory pathway and thus provide insights into the regulatory pathways coordinating their activities. Preliminary observations from these analyses indicate an interaction between the secretory pathway and the DNA replication/repair module, an interaction that can have potential implications for cancer.

Keywords Secretory pathway • DNA repair • GSEA

S.G. Chavan • P.D. Roy • S. Parashuraman (✉)
Institute of Protein Biochemistry at National Research Council (CNR), Naples, Italy
e-mail: r.parashuraman@ibp.cnr.it

K.P. Tripathi • M. Piccirilo • M. Guarracino
Laboratory for Genomics, Transcriptomics and Proteomics (LAB-GTP), High Performance Computing and Networking Institute (ICAR) at National Research Council (CNR), Naples, Italy

A. Luini
Institute of Protein Biochemistry at National Research Council (CNR), Naples, Italy

Istituto di Ricovero e Cura a Carattere Scientifico SDN, Via Emanuele Gianturco, 113, 80143
Naples, Italy

A. Rogato et al. (eds.), *Dynamics of Mathematical Models in Biology*,
DOI 10.1007/978-3-319-45723-9_7

1 Introduction

The secretory pathway is responsible for delivery of a large variety of proteins to their proper cellular location and is essential for cellular function and multicellular development [1]. It is composed of a series of compartments that include endoplasmic reticulum (ER), Golgi apparatus, and trans Golgi network (TGN), through which the cargoes (protein or lipid) are transported in an orderly fashion starting from the ER where the biosynthesis of cargoes is initiated. This is followed by processing of cargoes at the Golgi apparatus by addition of glycan groups and they are then sorted to their appropriate sites at the TGN [2]. At all these levels, each step including the anterograde transport (transport from ER to the PM via Golgi) and retrograde transport (transport in the reverse direction, but here refers mainly to the transport from Golgi to ER) is controlled by regulatory modules that maintain the homeostasis of the system [3]. Like every other module of the cell [4], the secretory pathway does not work in isolation but interacts with other cellular modules. Co-ordination circuits regulate the activities of these interacting modules so as to maintain homeostasis of the cell.

The functions and interactions of the secretory pathway have been studied by genome-wide RNA-mediated interference screens in Drosophila cell lines [5], in cultured human cells [6], and also in *Saccharomyces cerevisiae* [7]. Altered functioning of the secretory pathway has been associated with several pathologies, like the involvement of GOLPH3 in cancer [8] or the involvement of secretory pathway localized proteins in genetic diseases [9]. Nevertheless, the modules that interact with the secretory pathway and the underlying molecular circuits remain under explored, especially in the mammalian system. In order to identify and characterize these circuits we are implementing an approach based on the deconvolution of the transcriptional profiles resulting from the perturbation [10] of the secretory pathway. Such a study would help identify modules that interact with secretory pathway module and also provide insights into the regulatory pathways that coordinate their activities.

2 Materials and Methods

2.1 Strategy

To identify the modules that interact with the secretory pathway, we have perturbed its functioning by knocking down secretory pathway localized genes (using shRNAs), followed by an analysis of the changes in gene expression. Pathways or functions that were modulated under these conditions were identified using gene set enrichment analysis (GSEA). Following this, the transcription factors (TF) that might potentially regulate these pathways or functions were identified. The TFs can then be used to predict upstream signaling pathways that respond to the original perturbation (knockdown of secretory pathway localized genes). This analysis

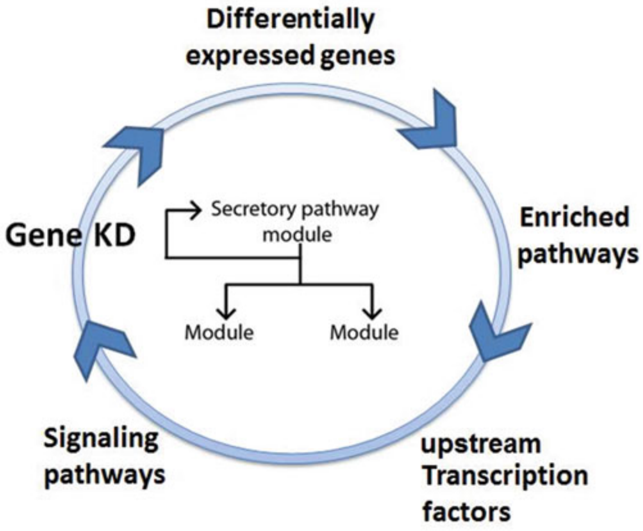


Fig. 1 Strategy to identify interacting modules of secretory pathway and the underling molecular circuit. Gene expression profiles obtained from cells, where the secretory pathway is perturbed by shRNA mediated silencing of secretory pathway localized genes, will be analyzed using GSEA to obtain pathways that are modulated by the perturbation. Then putative upstream transcription factors that can regulate the genes associated with these pathways will be predicted and validated. Then, literature mining coupled to experimental validation will be used to dissect the signaling pathways that modulate the TF activity under these conditions so to build the molecular circuit connecting perturbation to the gene expression changes

will help to map the molecular pathway connecting the perturbation of secretory pathway function to the modulation of other specific functions of the cells. This connection between cause and effect would help reveal the underlying molecular circuit regulating the interacting module. This general strategy is represented in Fig. 1.

2.2 *Data Collection and Processing*

The micro-array profiles following knockdown of secretory pathway genes were obtained from Library of Integrated Cellular Signatures (LINCS) which is an NIH program (http://www.lincscloud.org/perturbagens/) that funds the generation of perturbation profiles across multiple cell and perturbation types. Details about the cell types used in the study can be found here (http://www.lincscloud.org/cell-types/). It comprises of gene expression signatures produced by using L1000 technology which is a bead-based, high-throughput gene expression array. Only 1000 landmarks genes were experimentally measured and rest were computationally inferred. The transcriptional response was studied following the genetic and chemical perturbations. This data is computationally processed; wherein raw fluorescence intensity is converted to differential gene expression signatures. Data at different level of processing is available. Level 4 data was used in this study which represents signatures with differentially expressed genes computed by robust z-scores for each profile relative to population control. LINCS provides gene expression profiles for every perturbation obtained from different cell lines with multiple biological and technical replicates. The profiles for each perturbation of interest were subjected to pre-processing steps wherein the cell line dependent effect was removed by converting expression values to rank and then merging all of them into a single Prototype Ranked List (PRL) for each perturbation. This conversion and merging of data into PRL was done using "Gene Expression Signature Package" from Bioconductor in R [11]. This R package uses built-in "krubor" function to carry out rank merging process. It comprises of two steps (1) a distance is measured between two ranked list using Spearman's foot rule and two or more ranked lists are merged using Borda Merging method; (2) a single ranked list is obtained in a hierarchical way using Kruskal algorithm. Finally, all the PRLs representing the individual state (perturbation experiments) were generated as one input for the downstream analysis.

In this study, only small set of perturbation profiles were considered in order to standardize the system. This test set represents expression profiles following knockdown of a selected set of secretory pathway localized genes—ARF1, COPA, COPB2, COPZ1, COG2, COG4, COG7, M6PR, BLFZ1, GOLGA5, PLA2G4A, YKT6, RAB1B, SAR1B, TMED7, TMED9, TMED10, SEC24B, SEC24C, SEC24D, and BNIP1. The localization of these proteins across secretory pathway is represented in Fig. 2.

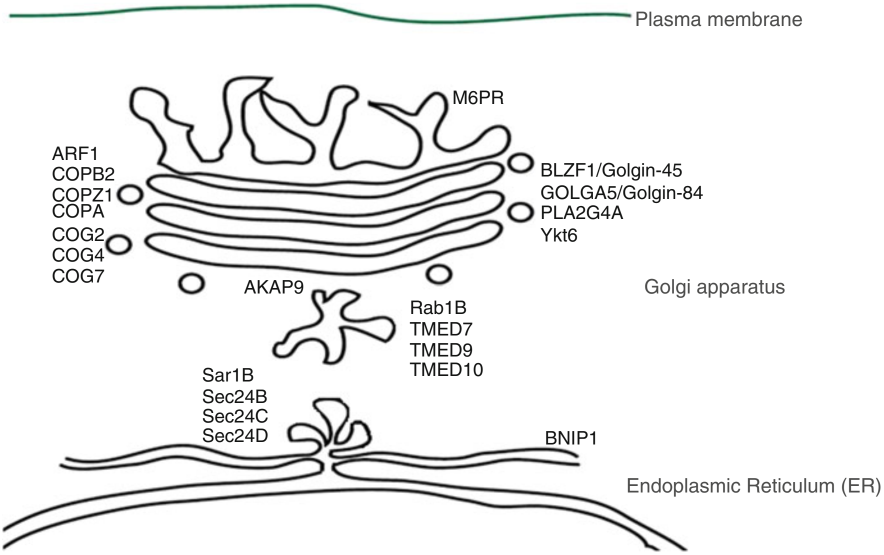


Fig. 2 Localization of the genes, whose expression was perturbed, to the compartments of the secretory pathway is represented

2.3 Gene Set Enrichment Analysis

The PRLs generated from gene expression profiles were subjected to GSEA using a java desktop application available at Molecular Signature Database (MsigDB; http://www.broadinstitute.org/gsea/down-loads.jsp). Given a set of a priori annotated set of genes (based on Gene ontology classifications, KEGG (Kyoto Encyclopedia of Genes and Genomes) pathways, or others), GSEA determines whether this set of genes shows statistically significant differences between two biological states viz. perturbation vs control that are being analyzed [12]. MsigDB has a collection of annotated gene sets (curated gene set, motif gene set, GO gene set, oncogenic signature, immunologic signature, etc.) for use with GSEA software. In order to study for the enrichment of the pathways, we used KEGG pathway gene set. Using GSEA, a number of enriched pathways were predicted across all the 22 PRLs. Only those pathways with the False Discovery Rate (FDR) cutoff 0.25 were taken into account. It has been suggested that given the lack of coherence in most expression datasets and the relatively small number of gene sets being analyzed, a FDR cutoff 0.25 is appropriate for the purposes of hypothesis generation [12]. We noted that many predicted pathways had significantly overlapping set of genes. So in order to streamline the results, the enriched pathways were consolidated into one group if they have more than 50 % of the genes overlapping.

2.4 Transcription Factor Prediction

The upstream transcription factors that can potentially regulate the expression of the genes belonging to the enriched pathways were predicted using the online resources TransFind (http://transfind.sys-bio.net/) and Locamo Finder (https://sysimm.ifrec.osakau.ac.jp/tfbs/locamo/) and HTRIDB (http://www.lbbc.ibb.unesp.br/htri). TransFind and Locamo Finder predict the TFs based on their affinity towards the putative promoters of the genes on interest. These affinities have been pre-calculated based on the available positional frequency matrices for the transcription factors [13]. On the other hand, prediction by HTRIDB is based on experimentally verified human transcriptional regulation interactions. Among the TFs obtained, only those that were commonly predicted by all these tools were selected for further analysis.

3 Result and Discussion

The gene expression profiles following shRNA mediated knockdown (KD) of selected secretory pathway genes were downloaded from LINCS database and processed to obtain PRLs (see methods). These PRLs were then subjected to GSEA analysis to obtain enriched pathways/functions that were modulated under

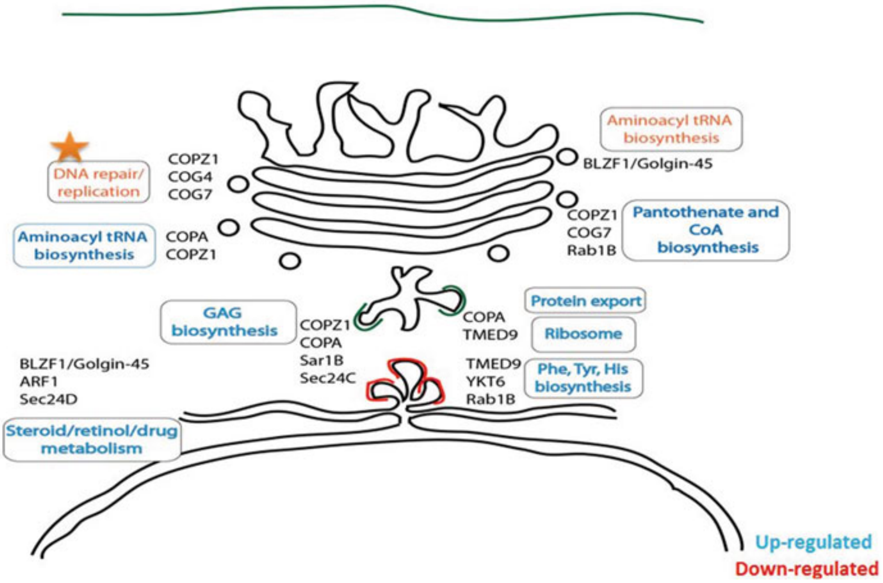


Fig. 3 The perturbed genes were grouped based on the common enriched pathways. The color code refers to downregulated pathways (in *red*) and upregulated pathways (in *blue*). Module unrelated to secretory pathway is marked by asterisk (*orange*)

the perturbation conditions. The perturbations were then grouped on the basis of the pathways that were modulated in common (Fig. 3). Most of these groups were related to secretory pathway module viz. glycosaminoglycan (GAG) biosynthesis, protein export, ribosome, Pantothenate and CoA biosynthesis, aminoacyl tRNA biosynthesis, and Phe, Tyr, His biosynthesis pathway, as expected. However, the group of COPZ1, COG4, and COG7 gene KDs was associated with the downregulation of DNA repair and replication pathway, which is a function not known to be related to the secretory pathway module. Thus GSEA analysis reveals both expected and unexpected modules that are modulated in response to a perturbation of the secretory pathway. COPZ1, COG4, and COG7 share a common known function of retrograde transport from Golgi to ER. This association suggests a possible interaction between the Golgi retrograde transport and the DNA repair response.

We then experimentally tested whether the DNA repair pathway is indeed regulated by the secretory pathway localized genes. To this end, we have downregulated COPZ1, COG4, or COG7 using siRNAs in HeLa cells, and then measured the increase of the DNA damage by studying the changes in the levels of phospho histone H3, a marker of the sites of DNA double strand breaks. Only the downregulation of COPZ1 showed increased levels of DNA damage as shown by an increase in the levels of phospho histone H3. Moreover, knockdown of coatomer proteins (COPA, COPZ1) has already been showed to increase DNA damage [14]. These findings suggest that the interaction between the modules of the secretory pathway

Table 1 Transcription factors predicted for DNA repair cluster

Cluster	Gene KD	Transcription factors
DNA repair/replication	COPZ1, COG4, COG7	E2F TF's (E2F1 and E2F4)
		HIF1A (Hypoxia inducible factor 1 alpha)
		NRF (Nuclear respiratory factor 1)
		RFX1 (MHC class II regulatory factor RFX1)
		Nkx2-5 (Homeobox protein Nkx-2-5)

Table 2 Position of predicted TF for DNA repair cluster

TF's	Probe ids	Position in profile
E2F1	2028_s_at	18146
	204947_at	22001
E2F4	202248_at	19056
HIF1A	200989_at	3507
NRF1	204651_at	19224
	204652_s_at	21124
	211279_at	16940
	211280_s_at	10201
RFX1	206321_at	3081

and DNA repair that we identified is probably a true interaction and moreover validates our strategy for identification of modules interacting with the secretory pathway.

We then analyzed the genes belonging to the DNA repair pathway that is modulated by the perturbation of Golgi retrograde transport (COPZ1, COG4, or COG7 KD), to identify the putative TFs that can regulate their expression. The enriched TFs obtained for this DNA repair group are listed in Table 1. Among these, E2F1, E2F4, NRF1, and RFX1 are known to be involved in regulation of DNA repair pathway genes of which E2F4 and RFX1 act as repressors [15, 16].

Since transcription factors are usually co-expressed along with their target genes, their position across the PRL (rank in the PRL) associated with COPZ1 KD was analyzed (Table 2), in order to restrict the TFs to those that are more likely to be the true effectors under our perturbation conditions. This analysis revealed that transcription factors E2F1, NRF1, and E2F4 are probably downregulated and HIF1A and RFX1 are probably upregulated. However, only the behavior of E2F1 (activator), NRF1 (activator), and RFX1 (repressor) are in concordant with the observed effect of the target genes, i.e., downregulation of DNA repair pathways (Fig. 4). This TF information can be used to map the upstream signaling pathways that connect DNA repair pathway to perturbation of COPZ1 expression (or impaired retrograde transport) by further analysis using online resources as well as experimental validation studies.

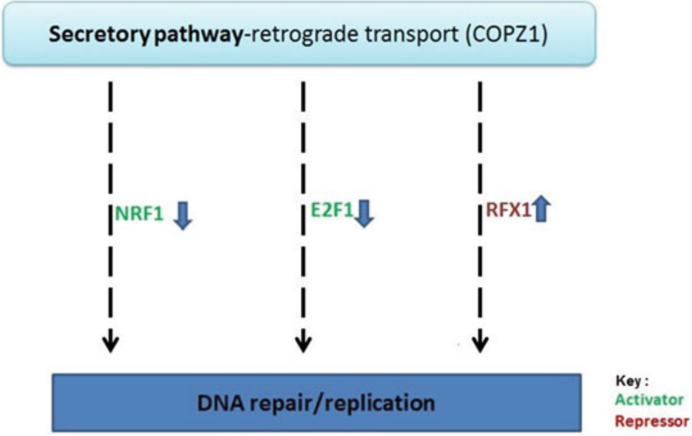


Fig. 4 Hypothetical model of Golgi retrograde transport mediated by COPZ1 possibly regulating the DNA repair pathway. The predicted TFs that might be involved in this regulation are indicated and the direction of their modulation (up- or downregulation) under conditions of COPZ1 KD is indicated by *colored arrows*. The known activity of the TF as a transcriptional activator or repressor is indicated by the color coding of the text. (Refer to the key for details)

3.1 Conclusion

The transcriptional profiling following the perturbation of genes localized to the secretory pathway shows a modulation in the levels of genes associated with pathways involved in the secretory transport as well as those unrelated to the secretory pathway. This study revealed an interesting possibility wherein the perturbation of secretory pathway function, particularly Golgi retrograde transport, might lead to the downregulation of DNA repair pathway. However, these predictions need to be experimentally validated and characterized. Since DNA repair pathways are associated with genome stability, understanding such interactions would be important for cancer studies. Moreover, we wish to extend this strategy of analysis for all the genes of secretory pathway in order to identify their novel functions.

References

1. Zhong, W.: Golgi during development. Cold Spring Harb. Perspect. Biol. **3**(9), a005363 (2011)
2. Kelly, R.B.: Pathways of protein secretion in eukaryotes. Science **230**(4721), 25–32 (1985)
3. Luini, A., Mavelli, G., Jung, J., Cancino, J.: Control systems and coordination protocols of the secretory pathway. F1000prime reports 6 (2014)
4. Costanzo, M., Baryshnikova, A., Bellay, J., Kim, Y., Spear, E.D., Sevier, C.S., Ding, H., Koh, J.L., Toufighi, K., Mostafavi, S., Prinz, J.: The genetic landscape of a cell. Science **327**(5964), 425–431 (2010)

5. Bard, F., Casano, L., Mallabiabarrena, A., Wallace, E., Saito, K., Kitayama, H., Guizzunti, G., Hu, Y., Wendler, F., DasGupta, R., Perrimon, N.: Functional genomics reveals genes involved in protein secretion and Golgi organization. Nature **439**(7076), 604–607 (2006)
6. Simpson, J.C., Joggerst, B., Laketa, V., Verissimo, F., Cetin, C., Erfle, H., Bexiga, M.G., Singan, V.R., Hrich, J.K., Neumann, B., Mateos, A.: Genome-wide RNAi screening identifies human proteins with a regulatory function in the early secretory pathway. Nat. Cell Biol. **14**(7), 764–774 (2012)
7. Jonikas, M.C., Collins, S.R., Denic, V., Oh, E., Quan, E.M., Schmid, V., Weibezahn, J., Schwappach, B., Walter, P., Weissman, J.S., Schuldiner, M.: Comprehensive characterization of genes required for protein folding in the endoplasmic reticulum. Science **323**(5922), 1693–1697 (2009)
8. Scott, K.L., Kabbarah, O., Liang, M.C., Ivanova, E., Anagnostou, V., Wu, J., Dhakal, S., Wu, M., Chen, S., Feinberg, T., Huang, J.: GOLPH3 modulates mTOR signalling and rapamycin sensitivity in cancer. Nature **459**(7250), 1085–1090 (2009)
9. De Matteis, M.A., Luini, A.: Mendelian disorders of membrane trafficking. N. Engl. J. Med. **365**(10), 927–938 (2011)
10. Ideker, T., Thorsson, V., Ranish, J.A., Christmas, R., Buhler, J., Eng, J.K., Bumgarner, R., Goodlett, D.R., Aebersold, R., Hood, L.: Integrated genomic and proteomic analyses of a systematically perturbed metabolic network. Science **292**(5518), 929–934 (2001)
11. Li, F., Cao, Y., Han, L., Cui, X., Xie, D., Wang, S., Bo, X.: GeneExpressionSignature: an R package for discovering functional connections using gene expression signatures. OMICS **17**(2), 116–118 (2013)
12. Subramanian, A., Tamayo, P., Mootha, V.K., Mukherjee, S., Ebert, B.L., Gillette, M.A., Paulovich, A., Pomeroy, S.L., Golub, T.R., Lander, E.S., Mesirov, J.P.: Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. Proc. Natl. Acad. Sci. U. S. A. **102**(43), 15545–15550 (2005)
13. Roider, H.G., Kanhere, A., Manke, T., Vingron, M.: Predicting transcription factor affinities to DNA from a biophysical model. Bioinformatics **23**(2), 134–141 (2007)
14. Paulsen, R.D., Soni, D.V., Wollman, R., Hahn, A.T., Yee, M.C., Guan, A., Hesley, J.A., Miller, S.C., Cromwell, E.F., Solow-Cordero, D.E., Meyer, T.: A genome-wide siRNA screen reveals diverse cellular processes and pathways that mediate genome stability. Mol. Cell **35**(2), 228–239 (2009)
15. Verona, R., Moberg, K., Estes, S., Starz, M., Vernon, J.P., Lees, J.A.: E2F activity is regulated by cell cycle-dependent changes in subcellular localization. Mol. Cell. Biol. **17**(12), 7268–7282 (1997)
16. Lubelsky, Y., Reuven, N., Shaul, Y.: Autorepression of rfx1 gene expression: functional conservation from yeast to humans in response to DNA replication arrest. Mol. Cell. Biol. **25**(23), 10665–10673 (2005)

Detection of Rare Mutations Using Beta-Binomial and Empirical Quantile Models in Next-Generation Sequencing Experiments

Sarunas Germanas, Audrone Jakaitiene, and Mario Guarracino

Abstract Next-generation sequencing is often used to identify genetic variants. The probability of variant detection also depends on the variant caller. Pooled data could be used to lower the sequencing cost. However identifying variants from pooled data is more challenging and demands more sophisticated mathematical methods. In this article we propose two novel SNP calling approaches: modification of Beta-binomial model as proposed in Flaherty et al. (2011) using posterior Beta distribution and empirical quantile method. Both offered methods and original Beta-binomial model were applied to pooled exome data of patients with neuromuscular diseases. The results showed that Beta-binomial model and modification of it were highly specific, however, with lower sensitivity compared to empirical quantile model. The positions could be identified as mutated using empirical quantile model much faster rather Beta-binomial models.

Keywords Variant calling • Pooling • NGS • Empirical quantile • Beta-binomial

1 Introduction

New generation sequencing (NGS) approach has revolutionized our possibilities to do a genetic and genomic research [7]. This approach provides cheap and fast way to sequence genetic data. NGS is used for de novo sequencing, for disease mapping,

S. Germanas
Institute of Mathematics and Informatics, Vilnius university, LT-08663 Vilnius, Lithuania
e-mail: sarunas.germanas@mf.vu.lt

A. Jakaitiene (✉)
Department of Human and Medical Genetics, Faculty of Medicine, Vilnius University,
LT-01513 Vilnius, Lithuania
e-mail: audrone.jakaitiene@mf.vu.lt

M. Guarracino
Laboratory for Genomics, Transcriptomics and Proteomics (Lab-GTP), High Performance
Computing and Networking Institute (ICAR), National Research Council (CNR),
Via Pietro Castellino 111, Naples, Italy
e-mail: mario.guarracino@cnr.it

A. Rogato et al. (eds.), *Dynamics of Mathematical Models in Biology*,
DOI 10.1007/978-3-319-45723-9_8

for variant calling, and for diagnostics [4, 8]. NGS suffers from high error rates which come from several error sources including base-calling and alignment.

During the single nucleotide polymorphism (SNP) calling process using NGS the variable sites of genome could be identified. Sophisticated SNP calling mathematical models could be used to reduce and quantify the uncertainty of variant calling process caused by high error rates. Another way could be target-sequencing of certain genetic region with higher sequencing rate (20x and more). However increasing demand of large samples suggests that high-depth sequencing is too expensive in time and cost. In such large sample cases alternative way of sequencing could be grouping of patients to pools and sequencing them together. This strategy gives a possibility to sequence more effective in terms of time and money. Although, this strategy also has a drawback—the allele frequency of certain individual from the pool cannot be estimated directly. The same applies to SNP and genotype calling. Therefore even more sophisticated mathematical methods and pooling strategies must be used in order to take advantage of pooled NGS data.

There are many SNP calling methods for pooled NGS data [1–3, 5, 6, 11, 13, 15]. One common property of these methods is that the non-referent allele frequency r_{iv} is modeled, where $i = 1, \dots, N$ is genetic position, $v = 1, \dots, J$ is the index of a pool. In [5, 6] hierarchical Beta-binomial model is offered and applied to synthetic pooled genetic data of virus. The quantity r_{iv} is assumed to have Beta-binomial distribution. In [3, 11, 15] random variable r_{iv} is assumed to have hierarchical binomial–binomial distribution. In models of [1, 2] statistically significant differences of r_{iv} between pools are searched to identify genetic variants. In [9, 14] sequencing error r_{iv} is modeled through genetic positions. This gives a benefit depending on the number of genetic positions, which often is very large. Although the sequencing error rate (θ) which is assumed constant in these models could be very different in different genetic loci. Applications of the methods mentioned above to real data with high minor allele frequency (>5 %) show that methods are quite sensitive and specific, but the rare event case is not clear or not sufficient [10, 12].

We offer two novel approaches for detecting SNPs in pooled NGS data. The first model is modification of Beta-binomial model [5], when the posterior distribution of r_{iv} is considered. In the second method we model the function of random variables $\frac{r_{iv}}{n_{iv}}$ with different v and use empirical quantile to detect the SNP calling threshold. Also, we use binomial approximation of r_{iv} to model the data in order to choose significant value of empirical quantile.

We use pooled NGS data from 128 patients with diagnosis of neuromuscular disease for SNP identification. Results show that the modification of Beta-binomial model detects variants with almost 100 % specificity. However the empirical quantile model has better sensitivity and is much faster compared to Beta-binomial models.

2 Materials and Methods

2.1 Modification of Beta-Binomial Model

We propose the first SNP calling model which is a modification of Beta-binomial model. The original Beta-binomial model was proposed in [5]:

$$r_{i,k_j(s)}|\theta_{i,k_j(s)} \sim Binomial(\theta_{i,k_j(s)}, n_{i,k_j(s)}), \tag{1}$$

$$\theta_{i,k_j(s)} \sim Beta(\mu_i, \Lambda), \tag{2}$$

where $r_{i,k_j(s)}$ is non-referent allele frequency (observed), $n_{i,k_j(s)}$ is read depth (observed), $\theta_{i,k_j(s)}$ is error rate parameter at position $i = 1, \dots, N$ in pool $k(s) = k_1(s), \dots, k_J(s)$, μ_i (expected value of Beta distribution) and Λ (precision of Beta distribution) are hyperparameters of the model; where $k_j = k_j(s)$ is a function which maps from the model set-up s to the k_j-th reference pool, $j = 1, \dots, J$. Parameter $\theta_{i,k_j(s)}$ and hyperparameters μ_i, Λ are estimated using Expectation-Maximization algorithm for the whole likelihood function:

$$L(r_{i,k_j(s)}, \theta_{i,k_j(s)}|\mu_i, \Lambda) = \prod_{i=1}^{N}\prod_{j=1}^{J} Pr(r_{i,k_j(s)}|\theta_{i,k_j(s)}, n_{i,k_j(s)})Pr(\theta_{i,k_j(s)}|\mu_i, \Lambda) \tag{3}$$

$$l(r_{i,k_j(s)}, \theta_{i,k_j(s)}|\mu_i, \Lambda) = \ln L(r_{i,k_j(s)}, \theta_{i,k_j(s)}|\mu_i, \Lambda), \tag{4}$$

Distribution of $r_{i,k_j(s)}$ which is assumed to be Beta-binomial is approximated with normal distribution with estimate of μ_i and standard deviation $\sigma_i = \frac{\mu_i(1-\mu_i)}{n_{i0}}(1 + \frac{n_{i0}-1}{\Lambda-1}), n_{i0} = \frac{1}{J}\sum_{j=1}^{J} n_{i,k_j(s)}$, and distribution of reference data is modeled. Z-test is applied for a main pool data.

We propose modification of Beta-binomial model. We use posterior expectation of $\theta_{i,k_j(s)}$ instead of prior:

$$E(\theta_{i,k_j(s)}|r_{i,k_j(s)}) = \mu_{post,i} = \frac{\mu_i\Lambda + \sum_{j=1}^{J} r_{i,k_j(s)}}{\Lambda + \sum_{j=1}^{J} n_{i,k_j(s)}}. \tag{5}$$

Therefore we use more information from the data and expect to get more accurate estimates of μ_i and $\theta_{i,k_j(s)}$. Estimate of standard deviation σ_i remains the same, but also the posterior standard deviation could be used. Z-test is applied for the case data.

We apply original Beta-binomial model and modification of it for set-ups of pooled data described in Sect. 2.3. Significance value for Z-test is chosen $\alpha = 10^{-6}$.

2.2 Empirical Quantile Model

In this section we present another SNP calling method as empirical quantile method. The idea of this method is do not use any theoretical distribution when predicting mutated positions and to use the data across all positions as it was applied in [9, 14].

We introduce the function $f : [0, 1]^{J+1} \to [-J, 1]$ (J is number of reference pools):

$$y_i^s = f(M_{i,l(s)}, R_{i,k_j(s)}) = M_{i,l(s)} - \sum_{j=1}^{J} R_{i,k_j(s)}, \tag{6}$$

where $M_{i,l(s)} = \frac{r_{i,l(s)}}{n_{i,l(s)}}$ is relative frequency of non-referent allele of main pool and $R_{i,k_j(s)} = \frac{r_{i,k_j(s)}}{n_{i,k_j(s)}}$ is relative frequency of non-referent allele of reference pool in i-th position, $k_j(s)$-th pool and s-th data set-up for a model; function $l = l(s)$ maps from model set-up s to the l-th main pool.

We expect that the value y_i^s is higher for mutated positions and lower for non-mutated positions. Therefore we use the empirical quantile q_α:

$$q_\alpha : Pr(y_i^s > q_\alpha) = \alpha. \tag{7}$$

Identification of mutated positions will depend on selection of α. For α estimation we model y_i^s as a sum of Bernoulli random variables. We assume that a pool is contributed by patients equally as Bernoulli random variables with the constant success probability which differs only between the main and reference pools.

$$M_{i,l(s)} = \sum_{c=1}^{Q} B_c^M = W_{p_M} \sim Binomial(Q, p_M), \tag{8}$$

$$\sum_{j=1}^{J} R_{i,k_j(s)} = \sum_{j=1}^{J} \sum_{c=1}^{Q} B_c^R = W_{p_R} \sim Binomial(QJ, p_R), \tag{9}$$

where $B_c^M \sim Bernoulli(p_M)$,$B_c^R \sim Bernoulli(p_R)$, $p_M = \frac{1}{SN} \sum_{s=1}^{S} \sum_{i=1}^{N} \frac{r_{i,l(s)}}{n_{i,l(s)}}$ and $p_R = \frac{1}{SJN} \sum_{s=1}^{S} \sum_{i=1}^{N} \sum_{j=1}^{J} \frac{r_{i,k_j(s)}}{n_{i,k_j(s)}}$ are, respectively, estimated error rate of the main and reference pools, $c = 1, \ldots, Q$ is the number of a patient in a pool, Q is assumed to be constant, and S is a number of data set-ups.

We do not use any prior information about mutation status of the positions and model distribution of y_i^s in three cases:

1. There is no information about position i (neither confirmed as mutated nor confirmed as non-mutated). y_i^s distribution using convolution formula is

$$P(y_i^s = t | i \textit{ is general}) = \sum_{h=1}^{Q} P_{W_R^{gen}}(h-t) P_{W_M^{gen}}(h), \tag{10}$$

where $W_R^{gen} \sim Binomial(QJ, p_R^{gen})$, $W_M^{gen} \sim Binomial(QJ, p_M^{gen})$, p_R^{gen} and p_R^{gen} are, respectively, error rates of main and reference pools estimated from all data.

2. Position i is mutated. For this case one pair of individuals has at least one mutation. We assume that this mutation is heterozygous and is present in both individuals. Therefore $M_{i,l(s)} = 1$ and $1 - R_{i,k_j(s)} \sim Binomial(QJ - Q + 2, p_R^{test})$.

$$P(y_i^s = t | i \textit{ is mutated}) = P_{W_R^{test}}(t-1), \tag{11}$$

where $W_R^{test} \sim Binomial(QJ - Q + 2, p_R^{test})$, p_R^{test} is error rate of main pools estimated from positions for which method is tested.

3. Position i is not mutated. For this case the pair of individuals has no mutation and $M_{i,l(s)} = 0$.

$$P(y_i^s = t | i \textit{ is not mutated}) = P_{W_R^{test}}(t). \tag{12}$$

In the second and the third case we use quantity $QJ-Q+2$ instead of QJ because we assume that part of main pool which is present in reference pools was canceled out (see Table 1). Modeled sensitivity and specificity are computed using (11) and (12) accordingly. Having calculated modeled sensitivity and specificity for different y_i^s, we determine the value of y_i^s and compute α from general distribution expressed in (10). Finally, for the assessment of the model, we calculate sensitivity and specificity using positions checked with Sanger.

2.3 Data Organization

We use pooled data from 128 patients with neuromuscular disease to identify mutated variants. Target exome regions were sequenced using Illumina sequencing platform. The target region consists of approximately 13,000 position with relative frequency of non-referent allele $M_{i,l(s)}$ varying from 0.01 to 0.06. Data consists of 8 original pools where each pool has 16 patients and 8 replicated pools which consist from the same 128 patients but with different pool composition (Table 1). For the models described above we used different organization of main and replicated pools as it is presented in Table 2.

Every pool from the original pool group was taken as main pool together with 7 pools from the replicated pool group as reference pools in such a way that every pair of patients from the main pool was not present in the reference pools. Every such combination of one main pool and seven reference pools we denote s, where $s = 1, \dots, 64$, and call data set-up for the model in the paper. Positions in every data set-up were filtered according to main pool—positions with $M_{i,l(s)} < 0.011$ where

Table 1 Organization of pools

Original pools (1–8)

Pool	P9	P9	P10	P10	P11	P11	P12	P12	P13	P13	P14	P14	P15	P15	P16	P16
P1	**1**	2	3	4	5	6	7	8	9	10	11	12	13	14	15	**16**
P2	17	18	**19**	20	21	22	23	24	25	26	27	28	29	**30**	31	32
P3	33	34	35	36	**37**	38	39	40	41	42	43	**44**	45	46	47	48
P4	49	50	51	52	53	54	**55**	56	57	**58**	59	60	61	62	63	64
P5	65	66	67	68	69	70	71	**72**	**73**	74	75	76	77	78	79	80
P6	81	82	83	84	85	**86**	87	88	89	90	**91**	92	93	94	95	96
P7	97	98	99	**100**	101	102	103	104	105	106	107	108	**109**	110	111	112
P8	113	**114**	115	116	117	118	119	120	121	122	123	124	125	126	**127**	128

Replicated pools (9–16)

Pool	P1	P1	P2	P2	P3	P3	P4	P4	P5	P5	P6	P6	P7	P7	P8	P8
P9	**1**	2	17	18	33	34	49	50	65	66	81	82	97	98	113	**114**
P10	3	4	**19**	20	35	36	51	52	67	68	83	84	99	**100**	115	116
P11	5	6	21	22	**37**	38	53	54	69	70	85	**86**	101	102	117	118
P12	7	8	23	24	39	40	**55**	56	71	**72**	87	88	103	104	119	120
P13	9	10	25	26	41	42	57	**58**	**73**	74	89	90	105	106	121	122
P14	11	12	27	28	43	**44**	59	60	75	76	**91**	92	107	108	123	124
P15	13	14	29	**30**	45	46	61	62	77	78	93	94	**109**	110	125	126
P16	15	**16**	31	32	47	48	63	64	79	80	95	96	111	112	**127**	128

not considered, because of the reasoning in [4]: when $M_{i,l(s)} < 0.011$ there cannot be any mutation because of finite number (16) of individuals in pool.

We have a list of mutated and non-mutated positions confirmed using Sanger sequencing which gives the possibility of golden standard for calculation of sensitivity and specificity.

3 Results

We apply Beta-binomial, modified Beta-binomial, and empirical quantile methods to pooled data from patients with neuromuscular diseases. For each model we use data set-up as in Table 2. For Beta-binomial model we model distribution of reference pools and apply Z-test for main pool for every model set-up.

An illustration of empirical quantile method is in Fig. 1 where values of y_i^s for all sequenced, mutated, and non-mutated positions are plotted. As we expected, the value y_i^s is positive for mutated positions and negative for non-mutated positions. This indicates that if we observe mutation in specific position, the error rate of the main pool is larger as error rate sum of the same position in reference pools. Therefore for the detection of mutated position, we need to find a threshold above which all positions would be detected as mutated only. For the determination

Table 2 Organization of pools: for every original pool eight combinations of replicated pools

Model set-up s	Main pools, $l(s)$	Reference pools, $k(s)$	Patient		Model set-up s	Main pools, $l(s)$	Reference pools, $k(s)$	Patient	
1	1	10, 11, 12, 13, 14, 15, 16	1	2	33	5	10, 11, 12, 13, 14, 15, 16	65	66
2		9, 11, 12, 13, 14, 15, 16	3	4	34		9, 11, 12, 13, 14, 15, 16	67	68
3		9, 10, 12, 13, 14, 15, 16	5	6	35		9, 10, 12, 13, 14, 15, 16	69	70
4		9, 10, 11, 13, 14, 15, 16	7	8	36		9, 10, 11, 13, 14, 15, 16	71	72
5		9, 10, 11, 12, 14, 15, 16	9	10	37		9, 10, 11, 12, 14, 15, 16	73	74
6		9, 10, 11, 12, 13, 15, 16	11	12	38		9, 10, 11, 12, 13, 15, 16	75	76
7		9, 10, 11, 12, 13, 14, 16	13	14	39		9, 10, 11, 12, 13, 14, 16	77	78
8		9, 10, 11, 12, 13, 14, 15	15	16	40		9, 10, 11, 12, 13, 14, 16	79	80
9	2	10, 11, 12, 13, 14, 15, 16	17	18	41	6	10, 11, 12, 13, 14, 15, 16	81	82
10		9, 11, 12, 13, 14, 15, 16	19	20	42		9, 11, 12, 13, 14, 15, 16	83	84
11		9, 10, 12, 13, 14, 15, 16	21	22	43		9, 10, 12, 13, 14, 15, 16	85	86
12		9, 10, 11, 13, 14, 15, 16	23	24	44		9, 10, 11, 13, 14, 15, 16	87	88
13		9, 10, 11, 12, 14, 15, 16	25	26	45		9, 10, 11, 12, 14, 15, 16	89	90
14		9, 10, 11, 12, 13, 15, 16	27	28	46		9, 10, 11, 12, 13, 15, 16	91	92
15		9, 10, 11, 12, 13, 14, 16	29	30	47		9, 10, 11, 12, 13, 14, 16	93	94
16		9, 10, 11, 12, 13, 14, 15	31	32	48		9, 10, 11, 12, 13, 14, 16	95	96
17	3	10, 11, 12, 13, 14, 15, 16	33	34	49	7	10, 11, 12, 13, 14, 15, 16	97	98
18		9, 11, 12, 13, 14, 15, 16	35	36	50		9, 11, 12, 13, 14, 15, 16	99	100
19		9, 10, 12, 13, 14, 15, 16	37	38	51		9, 10, 12, 13, 14, 15, 16	101	102
20		9, 10, 11, 13, 14, 15, 16	39	40	52		9, 10, 11, 13, 14, 15, 16	103	104
21		9, 10, 11, 12, 14, 15, 16	41	42	53		9, 10, 11, 12, 14, 15, 16	105	106
22		9, 10, 11, 12, 13, 15, 16	43	44	54		9, 10, 11, 12, 13, 15, 16	107	108
23		9, 10, 11, 12, 13, 14, 16	45	46	55		9, 10, 11, 12, 13, 14, 16	109	110
24		9, 10, 11, 12, 13, 14, 15	47	48	56		9, 10, 11, 12, 13, 14, 16	111	112
25	4	10, 11, 12, 13, 14, 15, 16	49	50	57	8	10, 11, 12, 13, 14, 15, 16	113	114
26		9, 11, 12, 13, 14, 15, 16	51	52	58		9, 11, 12, 13, 14, 15, 16	115	116
27		9, 10, 12, 13, 14, 15, 16	53	54	59		9, 10, 12, 13, 14, 15, 16	117	118
28		9, 10, 11, 13, 14, 15, 16	55	56	60		9, 10, 11, 13, 14, 15, 16	119	120
29		9, 10, 11, 12, 14, 15, 16	57	58	61		9, 10, 11, 12, 14, 15, 16	121	122
30		9, 10, 11, 12, 13, 15, 16	59	60	62		9, 10, 11, 12, 13, 15, 16	123	124
31		9, 10, 11, 12, 13, 14, 16	61	62	63		9, 10, 11, 12, 13, 14, 16	125	126
32		9, 10, 11, 12, 13, 14, 15	63	64	64		9, 10, 11, 12, 13, 14, 16	127	128

of the significance value for empirical quantile method, we approximate r_{iv} by binomial distribution and compute $M_{i,l(s)}, p_M^{gen}, p_M^{test}$ for main and $R_{i,k(s)}, p_R^{gen}, p_R^{test}$ for reference pools. We model distributions of y_i^s for general, mutated, and non-mutated positions. The latter distributions for all data set-ups are plotted in Fig. 2. We consider y_{i0}^s at integer values $\{-JQ, -JQ + 1, \ldots, 0, 1, 2, \ldots Q\}$ as possible threshold points. Specifically, we compute $\alpha = P(y_i^s \geq y_{i0}^s | i\ is\ general)$, modeled

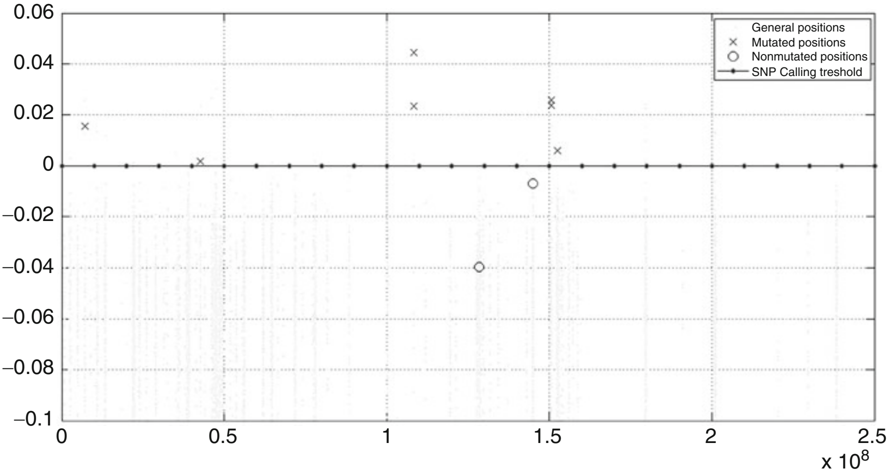


Fig. 1 In *X* axis all sequenced positions are represented, in *Y* axis values of y_i^s for model set-up $s = 1$ are represented

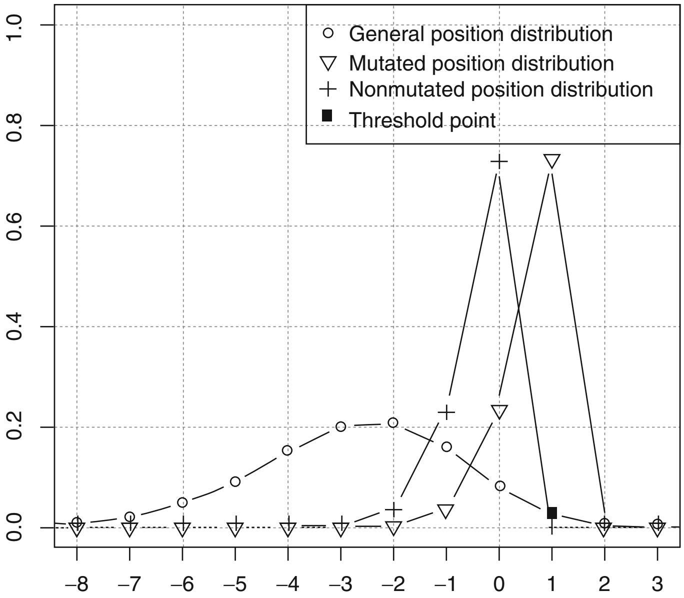


Fig. 2 In *X* axis values of y_i^s are represented, in *Y* axis modeled probabilities of y_i^s values for general, mutated, and non-mutated positions

Table 3 Sensitivity and specificity of the methods at different significance levels

	$\alpha = 10^{-6}$		$\alpha = 10^{-3}$		$\alpha = 0.03$	
Models	Sensitivity (%)	Specificity (%)	Sensitivity (%)	Specificity (%)	Sensitivity (%)	Specificity (%)
Beta-binomial	4.2	100	20.3	100	46.3	97.1
Modified Beta-binomial	5.9	100	35.6	100	69.5	97.1
Empirical quantile	0	100	48.3	100	95.7	82.3

sensitivity= $P(y_i^s \geq y_{i0}^s | i \textit{ is mutated})$, modeled specificity= $P(y_i^s < y_{i0}^s | i \textit{ is mutated})$ for every value y_{i0}^s. Our objective is to find y_{i0}^s for which both modeled specificity and sensitivity would be as close to 1. We obtained that only at $y_{i0}^s = 1$ both modeled sensitivity and specificity were larger than 0.7, i.e., ≈0.73, ≈1 and $\alpha \approx 0.03$. We use computed value $\alpha = 0.03$ in (7) to compute empirical quantile q_α. q_α is the threshold to distinguish between mutated and non-mutated positions.

For the model performance evaluation and comparison, we calculated sensitivity and specificity of Beta-binomial model, modification of Beta-binomial model, and empirical quantile method using positions checked with Sanger sequencing. As there is some methodological differences in significance value selection (α is selected for Beta-binomial models and estimated for empirical quantile method), we present sensitivity and specificity results at three different levels (see Table 3). $\alpha = 10^{-6}$ was used in [5] to account for of multiple testing and $\alpha = 0.03$ was estimated in empirical quantile method. Empirical quantile method gives better sensitivity for α equal 10^{-3} and 0.03. Additional advantage of empirical quantile method is speed. It takes approximately 4–5 s estimate mutated positions of all individuals. While the time for the implementation of Beta-binomial model is approximately 1 week. Therefore, we can conclude that empirical quantile method is applicable for detection of mutated positions in pooled NGS experiments.

However, empirical quantile method might be extended in several ways, as there are some strong assumptions made: (1) model parameters are not position dependent; (2) contributions of individuals into pools are equal; (3) sequencing error is not modeled; (4) read errors are independent between pools and positions; (5) selection of α must be done by researcher; (6) the method depends on the experiment structure; and (7) it was not considered in the model that observed frequencies in pools which have leastwise one common individual are statistically dependent.

Mentioned assumptions could be relaxed when, for example, Poisson-Binomial distribution instead of Binomial would be considered, dependence between pools and positions would be taken into account, weighted sums instead of sums would be calculated, selection of α would be automated.

There are several articles in which error rate across positions of target region is modeled. In [9] empirical quantile is computed with prescribed α and Poisson distribution is used to compute probability of SNP. In [14] also Poisson distribution is assumed as read error distribution, parameter of Poisson distribution is calculated from average error rate across positions, and threshold equal to 0.001 is used for SNP calling. These methods have limitations that distribution of read errors is assumed, parameter of Poisson distribution is assumed constant across positions, and SNP calling threshold is chosen without knowledge about analyzed data. Therefore the ability for selecting SNP calling threshold is advantage of proposed empirical quantile method.

4 Concluding Remarks

Pooled NGS experiments demand non-standard tools to accurately discover SNPs, therefore empirical quantile method could be appropriate, as it detects mutations with high sensitivity and specificity very fast. This method could be extended in many ways, for example, Poisson-Binomial distribution instead of Binomial could be used or dependence between pools and positions could be considered. These extensions could make the method even more efficient.

References

1. Altmann, A., Weber, P., Quast, C., Rex-Haffner, M., Binder, E.B., Müller-Myhsok, B.: vipR: variant identification in pooled DNA using R. Bioinformatics **27**, 77–84 (2011)
2. Bansal, V.: A statistical method for the detection of variants from next-generation resequencing of DNA pools. Bioinformatics **26**, 318–324 (2010)
3. Chen, Q., Sun, F.: A unified approach for allele frequency estimation, SNP detection and association studies based on pooled sequencing data using EM algorithms. BMC Genomics **14**, 1–14 (2013)
4. Ferraro, M.B., Savarese, M., di Fruscio, G., Nigro, V., Guarracino, M.R.: Prediction of rare single-nucleotide causative mutations for muscular diseases in pooled next-generation sequencing experiments. J. Comput. Biol. **21**, 665–675 (2014)
5. Flaherty, P., Natsoulis, G., Muralidharan, O., Winters, M., Buenrostro, J., Bell, J., Brown, S., Holodniy, M., Zhang, N., Ji, H.P.: Ultrasensitive detection of rare mutations using next-generation targeted resequencing. Nucleic Acids Res. **40**, 861–872 (2011)
6. He, Y., Zhang, F., Flaherty, P.: RVD2: an ultra-sensitive variant detection model for low-depth heterogeneous next-generation sequencing data. Bioinformatics **31–17**, 2785–2793 (2015)
7. Mardis, E.R.: A decade's perspective on DNA sequencing technology. Nature **470**, 198–203 (2011)
8. Nielsen, R., Paul, J.S., Albrechtsen, A., Song, Y.S.: Genotype and SNP calling from next-generation sequencing data. Nat. Rev. Genet. **12**, 443–451 (2011)
9. Out, A.A., van Minderhout, I.J.H.M., Goeman, J.J., Ariyurek, Y., Ossowski, S., Schneeberger, K., Weigel, D., van Galen, M., Taschner, P.E.M., Tops, C.M.J., Breuning, M.H., van Ommen, G.-J.B., den Dunnen, J.T., Devilee, P., Hes, F.J.: Deep sequencing to reveal new variants in pooled DNA samples. Hum. Mutat. **9**, 1703–1712 (2009)
10. Pabinger, S., Dander, A., Fischer, M., Snajder, R., Sperk, M., Efremova, M., Krabichler, B., Speicher, M.R., Zschocke, J., Trajanoski, Z.: A survey of tools for variant analysis of next-generation genome sequencing data. Brief. Bioinform. **15–2**, 256–278 (2012)

11. Raineri, E., Ferretti, L., Esteve-Codina, A., Nevado, B., Heath, S.: SNP calling by sequencing pooled samples. BMC Bioinf. **13**, 239–246 (2012)
12. Spencer, D.H., Tyagi, M.M., Vallania, F., Bredemeyer, A.J., Pfeifer, J.D., Mitra, R.D., Duncavage, E.J.: Performance of common analysis methods for detecting low-frequency single nucleotide variants in targeted next-generation sequence data. J. Mol. Diagn. **16**, 75–88 (2014)
13. Vallania, F.L.M., Druley, T.E., Ramos, E., Wang, J., Borecki, I., Province, M., Mitra, R.D.: Quantification of rare allelic variants from pooled genomic DNA. Nat. Methods **6**, 263–265 (2009)
14. Wang, C., Mitsuya, Y., Gharizadeh, B., Ronaghi, M., Shafer, R.W.: Characterization of mutation spectra with ultra-deep pyrosequencing: application to HIV-1 drug resistance. Genome Res. **17**, 1195–1201 (2007)
15. Wei, Z., Wang, W., Hu, P., Lyon, G.J., Hakonarson, H.: SNVer: a statistical tool for variant calling in analysis of pooled or individual next-generation sequencing data. Nucleic Acids Res. **39**, 132–144 (2011)

An Overview of Genotyping by Sequencing in Crop Species and Its Application in Pepper

Francesca Taranto, Nunzio D'Agostino, and Pasquale Tripodi

Abstract The exploitation of genetic variation in crops is essential to establish innovative breeding programs in the frame of global population increase and the sustainable intensification of agriculture. The advent of next generation sequencing technologies and the availability of complete or draft genome sequences of many crops allowed the development of several methods for SNP discovery. Genotyping by sequencing (GBS) has recently emerged as a promising approach to simultaneously allow SNP identification and genotyping. GBS provides a rapid, highly informative, high-throughput and cost-effective tool for exploring plant genetic diversity on a genome-wide scale and does not require any a priori knowledge on the genome of the species of interest. The features of GBS make it an attractive technology for (1) the assessment of population structure of germplasm collections; (2) the development of high density linkage maps and (3) genetic mapping studies. Herein, we present an overview of the GBS method and describe the main protocols in use, the principal methods for genetic diversity analysis and potential applications of the results in crop breeding programs. Finally, we illustrate the strategy we adopted to investigate the genetic diversity in cultivated pepper (*Capsicum annuum*).

Keywords Genotyping by sequencing • SNP markers • Genetic diversity • Population structure • Pepper

1 Introduction

The accessibility and use of natural genetic variation in plant breeding is currently restricted due to gaps in the genetic information that limit the comparison of germplasm accessions of different crops. The generation of novel varieties and the establishment of innovative breeding programs play a crucial role in food security and nutrition. In the last century, breeding programs have led to the selection of a small number of cultivars carrying genes for resistance to diseases

F. Taranto • N. D'Agostino • P. Tripodi (✉)
Consiglio per la ricerca in agricoltura e l'analisi dell'economia agraria (CREA) – Centro di ricerca per l'orticoltura, Via dei Cavalleggeri 25, Pontecagnano Faiano (SA), Italy
e-mail: pasquale.tripodi@crea.gov.it

A. Rogato et al. (eds.), *Dynamics of Mathematical Models in Biology*,
DOI 10.1007/978-3-319-45723-9_9

and pests with higher and more uniform yield [1], implicating a reduction of genetic diversity. In order to contrast this trend, international efforts are focusing on the recovery, protection and assessment of biodiversity and on the promotion of the sustainable use of plant genetic resources. Plant collections are constituted over time using both locally ecotypes, selected on the basis of a recognizable phenotype, and well-adapted crops selected for their fitness in climate change-affected production systems. The use of wild relatives and under-utilized varieties are instead challenging due to their unexplored genetic potentiality.

Crop improvement programs reaped the benefits from cutting-edge technologies in biological science, particularly in form of molecular markers, which in combination with conventional phenotype-based selection, define modern plant breeding practices. Molecular markers are extremely useful in plants to characterize germplasm collections and improve the conventional plant breeding schemes through marked-assisted selection (MAS). Different molecular markers have been successfully applied for genetic mapping [2], to infer phylogenetic relationships [3, 4], for the development of mapped genetic resources [5, 6] and comparative studies [7, 8].

Among various types of markers in use, single nucleotide polymorphisms (SNPs) are abundant in plant genomes; however, before the advent of next generation sequencing (NGS) technologies, they were considered costly for application in plant breeding [9, 10]. NGS is used for both whole genome sequencing and re-sequencing projects, leading to the discovery of a large number of SNPs useful to explore inter- and intra-species nucleotide diversity. As a consequence, SNPs have become the primary choice for many genetic studies thanks to their flexibility, speed and cost-effectiveness [11] inducing plant breeders to use them in their programs.

Genotyping by sequencing (GBS) has recently emerged as an innovative genomic approach for exploring plant genetic diversity on a genome-wide scale [12, 13]. GBS is based on genome reduction with restriction enzymes; it does not require a reference genome for SNP discovery and provides a rapid, high-throughput and cost-effective tool for the investigation of genetic variability in model and non-model species. Herewith it is provided an overview of the GBS method through the description of the main protocols in use and their applications in plants. In addition, research activity on the investigation of the genetic diversity in cultivated pepper (*Capsicum annuum*) is illustrated.

2 Why Genotyping by sequencing?

As mentioned above, a deep assessment of the available genetic variability within a crop is a necessary condition for a plant geneticist prior to plan a genetic improvement program. Moreover, the association between genetic variation and phenotypes of interest is the basis for MAS. The possibility to combine the processes of marker discovery and genotyping with a high-throughput and low-cost technology is the main achievement of GBS.

GBS was first introduced in plant science by Elshire et al. [13], and to date is one of the most powerful applications in the field of plant breeding. The information derived from GBS experiments have been widely used in genomic diversity studies and molecular marker discovery, genome-wide association studies (GWAS), genetic linkage analysis and genomic selection [9]. GBS can be performed through either a reduced-representation or a whole genome re-sequencing approach [12] generating a large number of genome-wide SNP data. It does not require any a priori knowledge on the genome of the species of interest, though several studies have been mainly carried out in species with reference genomes because SNP genotyping is much easier when a reference genome is available. Furthermore, GBS typically shows good results when it is applied to an inbred diploid species with a well-established reference genome as in the case of barley, maize, sorghum and brassica [9, 13]. Some studies have also made some progresses towards GBS of out-crossing species lacking reference genomes and of many agriculturally important polyploids crops such as wheat, cotton and potato [9, 14, 15]. Despite its benefits, GBS shows some limitations such as the presence of large amount of missing data, largely due to the use of low coverage sequencing and uneven genome coverage [16].

3 GBS Protocol and Data Analysis

The GBS protocol includes four major steps: (1) sample preparation, (2) NGS library construction, (3) SNP discovery and (4) genetic analysis (Fig. 1).

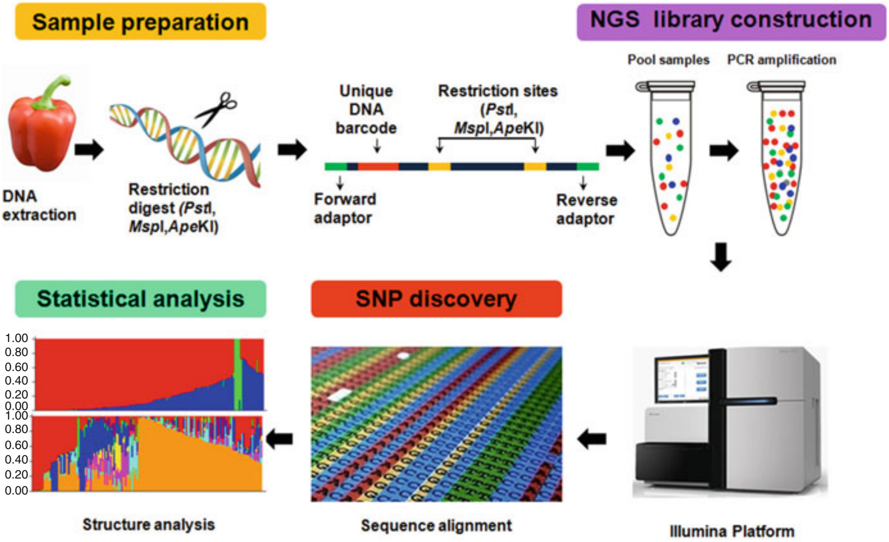


Fig. 1 Schematic steps of a genotyping by sequencing experiment

3.1 Sample Preparation

Sample preparation includes DNA extraction, assessment of quality parameters and library development. Total genomic DNA (gDNA) is extracted from plant tissue (young leaves or fruits) preferably using columns (i.e. DNeasy kits). The purity of DNA is confirmed by UV-Vis spectrophotometer and agarose gel electrophoresis. Using any spectrophotometer, the absorbance measurements consider the 260:280 nm to assess the purity of DNA. Usually, a ratio of ~1.8 is generally accepted as ‘pure’ for DNA. Agarose gel electrophoresis is used to check the quantity and quality of DNA and evaluate its possible degradation or contaminations from RNA, proteins, polysaccharides and other pigments that are difficult to separate from the DNA. However, DNA free from RNA can be obtained by treating the extracts with the RNAse A enzyme. Usually, thanks to the support of DNeasy kits, the purified DNA is free of proteins, nucleases and other contaminants and inhibitors, and therefore it is suitable for NGS. DNA requirements also include concentration above 50 ng/μl and trial digestions by methylation sensitive restriction enzyme (RE) (i.e. *Hind*III/*Eco*RI).

High-quality DNA is then digested with appropriate REs to reduce genomic complexity. The choice of the appropriate RE is a critical step in developing a GBS protocol for an organism. Since during library preparation no size selection step is performed, it is important to maximize the proportion of predicted restriction fragments that fall within the desired size range (100–400 bp) for sequencing. The original protocol provides *Ape*KI [14], which recognizes a degenerate 5-bp sequence (GCWGC, where W is A or T), creates a 5′ overhang (3-bp) and is partially methylation sensitive [13]. Other enzymes such as the rare cutter, *Pst*I (CTGCAG) and a frequent cutter, *Msp*I (CCGG) can be alternatively used. A larger pool of restriction fragments and consequently more unique sequences are generated by *Ape*KI. However *Pst*I and *Msp*I provide a greater degree of complexity reduction and uniform library for sequencing. Both REs have been widely used in genetic diversity studies on crops [14, 17]. Sonah et al. [18] described a modified library preparation protocol, in which selective amplification is used to increase both the number of SNPs called and their depth of coverage, resulting in a high efficiency and a reduction in per sample cost. After digestion, fragments are directly ligated to a pair of enzyme-specific adapters, which contain specific priming sites for the Illumina sequencing. Following ligation, the fragments are PCR amplified.

3.2 NGS Library Construction

Up to four amplicons with similar concentrations are generally pooled in order to assemble the library. A selection of Illumina-specific sequences is carried out followed by appropriate quantification and adjustment to preferred concentration. Afterwards, fragments are combined to form a sample library. As described in

Peterson et al. [16], the GBS protocol uses the Illumina 'Generate FASTQ' workflow, the 'FASTQ Only' application and 'TruSeq HT' assay to generate a de-multiplexed set of FASTQ files with the adapter sequences removed upon completion of the sequencing run. After sequencing run, raw data are downloaded. Each sample has two FASTQ files representing the forward and reverse sequenced reads. FASTQ files are text-based files for storing biological sequences (FASTA) with embedded quality scores.

3.3 SNP Discovery

Different computational pipelines have been specifically developed for SNP discovery and genotyping from FASTQ files. The TASSEL-GBS Discovery Pipeline is the most used in diploid plants with a reference genome [19]. It uses the first 64 nucleotides (nts) of the reads to minimize the effects of sequencing errors. As mentioned above, the sequencing produces million reads, split across multiple FASTQ files. All unique sequence tags from each sequence file are captured and then collapsed to generate a master tag file. The alignment of the unique 64-nts reads (tags) to reference genome is carried out using Bowtie2 [20] or BWA. A 'TagsOnPhysicalMap' (TOPM) file is returned as output and it can be used for SNP calling. SNP call is carried out for each set of tags originating from the same restriction enzyme cut site. Every set of tags aligns to the exact starting genomic position and strand, where the starting genomic position of a tag is identify by the cut site residue at the beginning of the tag. Raw SNP data output produced by the TASSEL-GBS pipeline are further filtered for studying purposes. Usually, the parameters considered are: inbreeding coefficient (F_{IT}) and minimum minor allele frequency (mnMAF). F_{IT} is largely used to filter SNPs from NGS data in inbred lines [21] and it is calculated based on the expectation–maximization (EM) algorithm [22]. In GBS analysis, spurious SNPs will appear to be excessively heterozygous, so it is necessary to calculate the F_{IT} and apply the minimum F_{IT} filter, generally 0.8 [19]. To detect and filter out error-prone SNPs, the TASSEL-GBS pipeline relies on population-genetic parameters such as MAF. The minimal filter used is in general set to MAF>0.01. Minimum minor allele count (mnMAC) and minimum locus coverage (mnLCov) are two additional parameters used in GBS analysis to count the number of minor alleles for each marker and to evaluate the proportion of taxa with a genotype, respectively [19].

The TASSEL-GBS pipeline provides SNP calls in both HapMap and VCF formats. The pipeline provides two sets of HapMap files: (1) a set without post SNP calling filtering; (2) a set with additional filtering on missingness and allele frequency. VCF format is an alternative format for holding SNP information that retains information on depth of coverage for each allele, and the genotype likelihood scores are calculated according to Etter et al. [23]. Specific software packages, such as VCFtools and VCFlib, have been developed for working with and manipulating VCFfiles [24]. For species with no reference genomes, a network-based algorithm

(UNEAK) and a computational pipeline, npGeno, were specifically developed for SNP discovery and genotyping [9, 16]. These two last bioinformatic pipelines have been developed particularly for polyploidy species, such as wheat, cotton and potato.

3.4 Genetic Analysis

The output data files generated from the bioinformatic pipelines are widely used in different genetic studies including conventional analysis to evaluate heterozygosity and genetic relationships among individuals, genetic diversity and population structure in large germplasm collections, high density linkage maps development, phylogenetic and association mapping studies. Each of these aspects requires complex analysis. In the next paragraph, a brief overview of the main methods used for genetic diversity studies is given.

4 Methods for Studying Genetic Diversity in Crops

The study of genetic variation is of great interest for trait association analysis and evolutionary researches. The first step is to investigate the population structure (the presence of genetic differences among groups of individuals and their assignment to different clusters based on allele frequency) given the large amount of SNP data. So far, several algorithms have been proposed which can be divided into two major computational paradigms: parametric and non-parametric. Parametric approaches assume a model in which there are K populations, each of which characterized by a set of allele frequencies at each locus. The assignment of individuals to a specific cluster is based on statistical likelihood method, using assumption such as Hardy–Weinberg equilibrium (HWE) for each marker and linkage equilibrium (LE) among markers [25]. The structure paradigm consists in a model-based clustering approach to infer the presence of distinct populations, assign each individual to a population and estimate ancestral population allele frequencies based on a statistical method known as the allele-frequency admixture model [26]. The most popular software to investigate the genetic structure in plants is STRUCTURE [26], although, in the last years, the ADMIXTURE [27, 28] program usage is growing. Both software used the same statistical model and input files (i.e. HapMap by the TASSEL-GBS pipeline) although ADMIXTURE performs much more rapidly since it employs a fast numerical optimization. STRUCTURE uses a Markov Chain Monte Carlo (MCMC) stochastic algorithm to produce sample-based estimates of a target distribution of choice and Bayesian approach based on the posterior distribution of defined population quantities. ADMIXTURE employs the same likelihood model but focuses on maximizing the likelihood rather than the posterior distribution. ADMIXTURE makes the further assumption of linkage equilibrium among the markers where dense marker sets should be pruned to mitigate background linkage disequilibrium (LD).

Both software estimates the best value of K. STRUCTURE HARVESTER (http://taylor0.biology.ucla.edu/structureHarvester/) is a web-based program developed to analyse the results generated by the program STRUCTURE. The algorithm implemented in STRUCTURE HARVESTER allows to assess and visualize likelihood values across multiple values of K and requires at least three values of sequential K with at least three replicates, and that the sample standard deviation of the log likelihood values across all K values is non-zero.

ADMIXTURE uses a cross-validation procedure to identify which K has the best predictive value, as determined by 'holding out' data points [27, 29].

The non-parametric approaches provide an alternative series of statistical methods that require few assumptions for data analysis [30]. There is a wide range of methods that can be used for different purposes. A viable tool to understand population diversity and structure is AWclust. AWclust has been firstly used to investigate genetic diversity in a human population [31] and it calculates the allele sharing distance (ASD) matrix, which represents the underlying genetic distance between every pair of individuals. The non-parametric analysis generated a multidimensional scaling (MDS) 2D/3D plots to recognize how the samples grouped, and the dendrogram tree to get a general relationships among individuals and to identify the number of population clusters. Furthermore, AWclust calculates the Gap statistics for estimating the optimal number of group (K) based on sample genetic relatedness.

Once the population structure is assessed, it is possible to select individuals in order to define a core collection and reduce the number of genotypes for downstream association mapping studies.

5 GBS Application in Crop Species

GBS technology is becoming pivotal as a cost-effective and unique tool for genomics-assisted breeding in numerous crop species. GBS has been successfully applied for a range of studies including genetic mapping [13, 17], assaying genetic diversity and population structure [32] and genomic selection [14]. Both monocots and dicots have been optimized by GBS for the efficient, low-cost and high-throughput SNP marker discovery (Table 1). The results achieved with GBS depend mainly on the type of population, the genome of the species and the protocol used.

Several examples of GBS studies were reported in diploid species and are focused on recombinant inbred lines (RILs) and germplasm collections. RILs are particularly feasible for GBS because of their high homozygosity, minimizing heterozygote genotyping errors caused by low read depth. Considering RILs, 2,815 maize inbred accessions were genotyped and 681,257 SNP markers distributed across the entire genome were detected, some of which linked to known candidate genes for quality traits and flowering time [33]. In sorghum, GBS analysis was conducted with an F_6 RIL population derived from an intra-specific cross between two *Sorghum bicolour* cultivars. The pilot study was performed by a single MiSeq

Table 1 Recent crop species analysed by GBS approach (non-exhaustive list)

Crops	Species	Genome size (Mb)	Sample size	N. SNPs	References
Maize	*Zea mays* L.	2,600	33,000	2,200 K	Romay et al. [33]
Rice	*Oryza sativa* L.	400	850	60 K	Spindel et al. [34]
Barley	*Hordeum vulgare* L.	5,427	160	1,949	Poland et al. [14]
Brassica	*Brassica oleracea*	628	89	683	Kim et al. [9]
Sorghum	*Sorghum bicolor* L.	700	90	576	Kim et al. [9]
Soybean	*Glycine max* L.	1,024	301	16,502	Jarquín et al. [35]
Potato	*Solanum tuberosum* L.	840	636	129,156	Uitdewilligen et al. [15]
Cotton	*Gossypium hirsutum* L.	197,632	39	956	Kim et al. [9]
Alfa-alfa	*Medicago sativa* spp.	800	48	11,694	Rocher et al. [36]
Oat	*Avena sativa* L.	11,300	2,664	45,117	Huang et al. [37]
Oil palm	*Elaeis guineensis*	1,800	108	21,471	Pootakham et al. [38]
Cassava	*Manihot esculenta* Crantz	530	917	56,489	Rabbi et al. [39]
Watermelon	*Citrullus lanatus var. lanatus*	425	183	11,483	Nimmakayala et al. [32]
Guinea yams	*Dioscorea* spp.	~1,200	95	6,371	Girma et al. [40]
Peach	*Prunus persica* (L.) Batsch	227	57	9,998	Bielenberg et al. [41]
Miscanthus	*Miscanthus sinensis*	2,592	230	49,007	Ma et al. [42]
Pine	*Pinus* spp.	~25,000	99	7,000–14,751	Pan et al. [43]

Genome size is in megabases (Mb), number of individuals genotyped (sample size) and number of SNPs used to assess genetic diversity

run (~25 million reads in total) with 90 mapping individuals plus parents (three redundant samples each), resulting in a total of 576 SNPs genetically mapped with the aid of the reference genome [9]. In rice, 30,894 SNPs were identified on 176 RILs and used to map the recombined hot and cold spots and QTLs for leaf width and aluminium tolerance [34].

Few studies have been performed on germplasm collections to characterize the genetic structure and to provide a tool for association mapping analysis for complex traits. As an example we report the work by Nimmakayala et al. [32], where the genetic structure of 183 domesticated watermelon accessions is investigated using a data set of 11,485 SNPs. Based on 5,254 filtered SNPs, linkage disequilibrium and population structure were estimated in order to identify agronomically important candidate genes. GBS has also been used for marker development in cassava (*Manihot esculenta* Crantz). Using a set of 917 accessions, 56,489 SNP loci were genotyped to assess population structure and perform varietal identification [39].

GBS was applied also in polyploid species such as potato, wheat and cotton. In potato, 12.4 gigabases of high-quality sequence data and 129,156 sequence variants have been identified [15]. In bread wheat, GBS was used to develop a high density map of 20,000 SNPs. To further evaluate GBS in wheat, a *de novo* genetic map was also constructed using only SNP markers from GBS experiment. The GBS approach presented here provides a powerful method of developing high density markers in species without a sequenced genome while providing valuable tools for anchoring and ordering physical maps and whole genome shotgun sequences [17].

Successful results in markers assisted breeding programs are reported. In cotton, GBS was used to genotype two BC_4F_1 populations and design strategies to obtain near isogenic lines (NILs) [9]. Two reciprocal sets of NILs by introgression between two tetraploid species were developed. In the first, 956 SNPs were used to genotype 39 individuals, which resulted in a total finding of 106 introgressions on average. The second set consisted of 39 individuals genotyped with 914 SNPs for a total of 114 introgressions. In pepper, GBS technology was used to develop a marker-assisted backcrossing (MABC) program for the constitution of new pepper varieties containing capsinoids, starting from BC_1F_1 and BC_2F_1 populations [44].

Despite the economical and nutritional importance of Solanaceae and the huge variability within, analytical studies on the genetic variability in germplasm collections using GBS are lacking. In the next paragraph we illustrate our research activity aiming to investigate genetic diversity in a population of cultivated pepper (*Capsicum annuum*) accessions.

6 Genetic Diversity Analysis in *Capsicum annuum* Using GBS

Pepper (*Capsicum* spp.) belongs to Solanaceae, which is an economically important family of flowering plants consisting of ~102 genera and ~2500 species. Plants belonging to the genus *Capsicum* had their origins in the South American regions and now are widely cultivated in tropical and temperate areas. Estimates report the existence of about 40 species (www.theplantlist.org) five of which (*C. annuum*, *C. baccatum*, *C. chinense*, *C. frutescens* and *C. pubescens*) were domesticated through distinct events at different primary diversification centres [45] and are widely consumed as sweet and hot peppers. Most of the species are diploids with 24 and 26 chromosomes and are distributed in three main gene pools based on morphological characteristics, chromosome banding or hybridization studies. Among the domesticated *Capsicum* spp., *C. annuum* ($2n = 2x = 24$) is the most widely grown species in the world, consumed as food or processed product and it is the most used in pepper breeding programs [46]. For most cultivated species the loss of genetic variability started as soon as the domestication process and subsequent steps of artificial selection. This led to the great variation in size, shape, colour and pungency of contemporary *C. annuum* fruits, depending on consumers' preference and product destination (fresh or powder). By contrast, the large number of landraces and ecotypes developed as a consequence of farmers' selection represents a wide source of diversity, particularly for alleles of agricultural interest and related to local adaptation. Therefore, the availability of large germplasm collections facilitates the evaluation of population diversity and genetic structure, providing vital information for genome-wide association mapping and allele mining studies to be exploited by plant breeders for the development of novel varieties and seed conservation programs [47, 48]. More recently, several approaches were developed in pepper to assess the genetic diversity and track allelic variants associated with phenotypic variations. The recent whole genome sequencing of *Capsicum* [49, 50] provides a more complete view to estimate chromosome wide molecular diversity and precisely infer pepper population structure.

Aiming to contribute in this scenario, we determined population structure and estimated genetic diversity in a collection of cultivated pepper (*Capsicum annuum*) using GBS.

Our approach consisted to collect and phenotype two hundred accessions of cultivated pepper for a wide range of agronomical and morphological traits. Genetic materials were retrieved from farmers and producers from over 20 world countries. Plants were previously stabilized through self-fertilization cycles and DNA was extracted using DNeasy plant column (QIAGEN). Quality parameters were measured by absorbance at 260 and 280 nm, respectively, using a UV-Vis spectrophotometer (ND-1000; Nanodrop, Thermo Scientific, USA). GBS was performed following the protocol described in Elshire et al. [13] using the *Ape*KI enzyme. About 8 million master tags were aligned to reference CM334 genome [49]. The TASSEL-GBS pipeline allowed to identify almost 100k filtered SNPs, which have been used to determine the population structure. Hereafter we show preliminary results on the population structure using a parametric approach of study. According to the Evanno's test [51] (Fig. 2), the population was divided into 3 clusters (Fig. 3a). A main group was identified which includes most of the

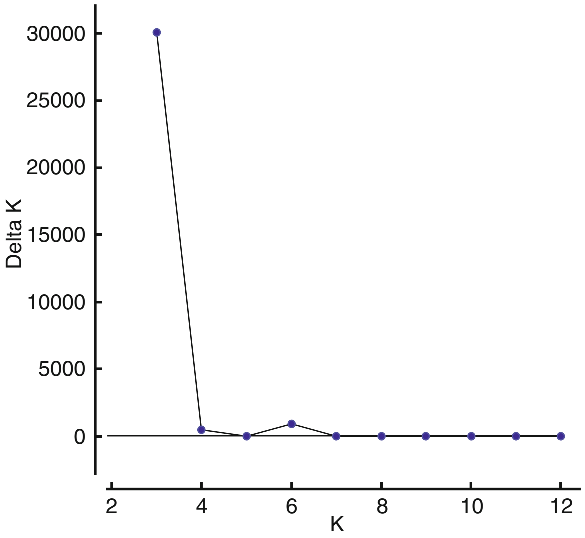


Fig. 2 Evanno's plot generated by STRUCTURE HARVESTER for the detection of the true number of clusters (the most likely value of K). The best value was at K = 3

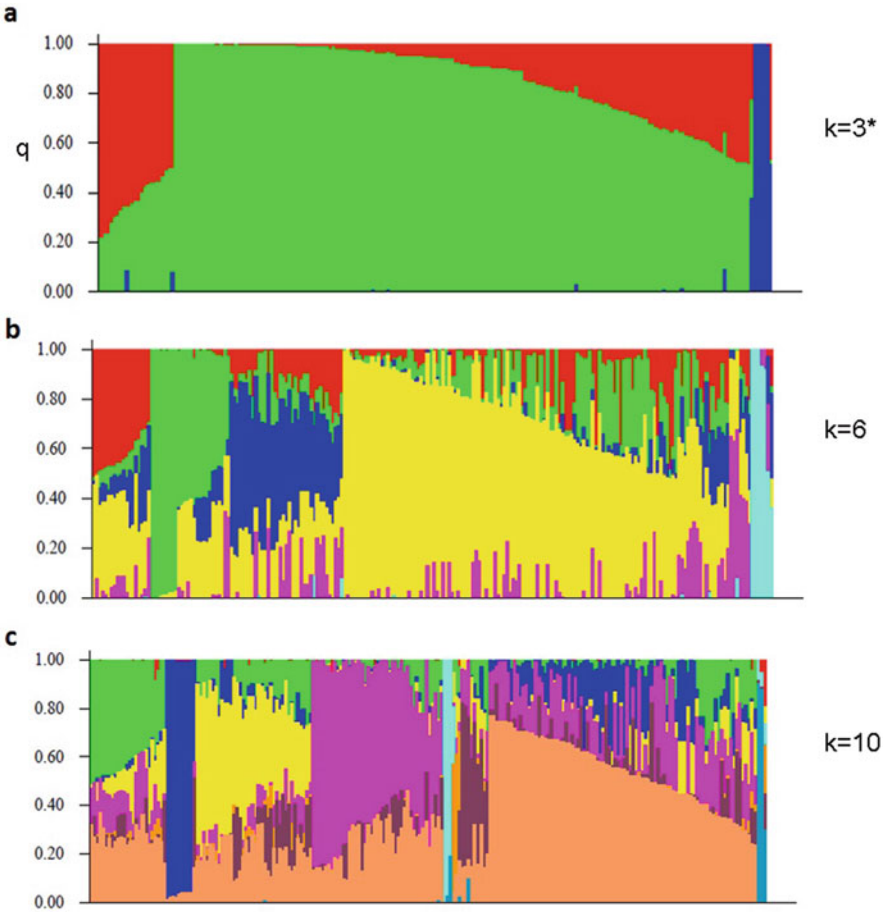


Fig. 3 Estimate of genetic diversity of *C. annuum* accessions using GBS-SNP markers. Barplot describing the population structure estimated by the Bayesian clustering. Each individual is represented by a thin vertical line, which is partitioned into *K* coloured segments whose length is proportional to the estimated membership coefficient (q). Population structure at (**a**) K = 3, (**b**) K = 6, (**c**) K = 10 is reported. Three, six and ten groups are identified, respectively. The *asterisk* shows the most informative K value (K = 3). Genotypes retrieved from the same geographical areas are represented by *yellow* and *brown lines* at K = 6 and K = 10, respectively

accessions having a common geographical origin. Considering the large variability within these sub-populations we performed the STRUCTURE analysis on *C. annuum* collection using other different K, particularly K = 6 and K = 10. At K = 6 (Fig. 3b) the clusters displayed some admixture, and the genetic structure of collection was not informative. Considering K = 10 (Fig. 3c), there was a better distinction considering other characteristics such as fruits morphology and pungency. As observed, increasing the number of sub-populations (K) it was possible to distinguish the accessions considering both geographical origin and fruit characteristics. Detailed assessment of morphological fruit-related characteristics was carried out using automated tools for the analysis (i.e. Chroma metre, 2D scanner). In total over 300 thousand data points for 38 fruit size and shape attributes were obtained. Main phenotypic variation was due to fruit size traits (i.e. perimeter, area, fruit height and fruit width) which could be considered the most relevant attributes for breeding new varieties. In order to identify genomic regions responsible for the phenotypic variation, high-quality SNP (mmMAF 0.01, coverage 90 %) were further selected. A first attempt to associate SNP alleles and morphological traits was carried out on the basis of General Linear Model. Several SNP highly correlated to the phenotypic variation were identified. For the main traits responsible for fruit size variation as well as for shape traits of high interest in breeding, highly correlated SNP were detected on chromosomes 2, 3, 6 and 9. Next step will involve the integration of a parametric (STRUCTURE) with a non-parametric approach (AWclust) in order to better refine the population structure with the aim to select a core-set of accessions. Moreover, Mixed Linear Model will be used for future association mapping analysis.

7 Conclusion

GBS is a high-throughput and low-cost technology used in several crop species in order to genotype breeding population, assess genomic diversity, discover and develop new molecular markers useful in plant breeding programs, and carry out GWAS. GBS, has proven useful and reliable for the identification of high-quality SNPs. It has several advantages, including the fact that no preliminary sequence information is required and that all newly discovered markers originate from the population under investigation.

Our study aims to unlock the genetic potentiality of cultivated pepper, which represents a major vegetable crop given its nutritional properties. GBS has been chosen to identify a large number of SNPs useful to precisely define the structure of a *C. annuum* population. Moreover large-scale phenomics has been carried out for fruit-related traits. Information concerning SNP markers and population structure developed in this study are the first step towards future genome-wide association mapping studies and marker-assisted selection programs in cultivated pepper.

Acknowledgements This project was supported by the 'GenHort' project funded by the Italian Ministry of University and Research (MIUR, PON02_00395_3215002) and the 'PEPIC' project funded by the Italian Ministry of Agricultural, Food and Forestry.

References

1. Hammer, K., Arrowsmith, N., Gladis, T.: Agrobiodiversity with emphasis on plant genetic resources. Naturwissenschaften **90**, 241–250 (2003)
2. Pei, C., Wang, H., Zhang, J., Wang, Y., Francis, D.M., Yang, W.: Fine mapping and analysis of a candidate gene in tomato accession PI128216 conferring hypersensitive resistance to bacterial spot race T3. Theor. Appl. Genet. **124**(3), 533–542 (2012)
3. Di Dato, F., Parisi, M., Cardi, T., Tripodi, P.: Genetic diversity and assessment of markers linked to resistance and pungency genes in Capsicum germplasm. Euphytica **1**, 103–119 (2015)
4. Xu, Y., Ma, R.C., Xie, H., Liu, J.T., Cao, M.Q.: Development of SSR markers for the phylogenetic analysis of almond trees from China and the Mediterranean region. Genome **47**(6), 1091–1104 (2004)
5. Alseekh, S., Ofner, I., Pleban, T., Tripodi, P., Di Dato, F., Cammareri, M., Mohammad, A., Grandillo, S., Fernie, A.R., Zamir, D.: Resolution by recombination: breaking up Solanum pennellii introgressions. Trends Plant Sci. **18**(10), 536–538 (2013)
6. Laidò, G., Mangini, G., Taranto, F., Gadaleta, A., Blanco, A., Cattivelli, L., Marone, D., Mastrangelo, A.M., Papa, R., De Vita, P.: Genetic diversity and population 387 structure of tetraploid wheats (Triticum turgidum L.) estimated by SSR, DArT and Pedigree Data. PLoS One **8**(6), e67280 (2013)
7. Wu, F., Eannetta, N.T., Durrett, Y.X.R., Mazourek, M., Jahn, M.M., Tanksley, S.D.: A COSII genetic map of the pepper genome provides a detailed picture of synteny with tomato and new insights into recent chromosome evolution in the genus Capsicum. Theor. Appl. Genet. **118**, 1279–1293 (2009)
8. Wu, F., Eannetta, N.T., Xu, Y., Plieske, J., Ganal, M., Pozzi, C., Bakaher, N., Tanksley, S.D.: COSII genetic maps of two diploid Nicotiana species provide a detailed picture of synteny with tomato and insights into chromosome evolution in tetraploid N. tabacum. Theor. Appl. Genet. **120**(4), 809–827 (2010)
9. Kim, C., Guo, H., Kong, W., Chandnani, R., Shuang, L.S., Paterson, A.H.: Application of genotyping-by-sequencing technology to a variety of crop breeding programs. Plant Sci. **242**, 12–14 (2016)
10. Rafalski, A.: Applications of single nucleotide polymorphisms in crop genetics. Curr. Opin. Plant Biol. **5**, 94–100 (2002)
11. He, J., Zhao, X., Laroche, A., Lu, Z.X., Liu, H., Li, Z.: Genotyping-by-sequencing (GBS), an ultimate marker-assisted selection (MAS) tool to accelerate plant breeding. Front. Plant Sci. **5**, 484 (2014)
12. Deschamps, S., Llaca, V., May, G.D.: Genotyping-by-sequencing in plants. Biology **1**, 460–483 (2012)
13. Elshire, R.J., Glaubitz, J.C., Sun, Q., Poland, J.A., Kawamoto, K., Buckler, E.S., Mitchell, S.E.: A robust, simple genotyping-by-sequencing (GBS) approach for high diversity species. PLoS One **6**(5), e19379 (2011)
14. Poland, J.A., Brown, P.J., Sorrells, M.E., Jannink, J.L.: Development of high-density genetic maps for barley and wheat using a novel two-enzyme genotyping-by-sequencing approach. PLoS One **7**, e32253 (2012)
15. Uitdewilligen, J.G., Wolters, A.A., D'hoop, B.B., Borm, T.J., Visser, R.G., van Eck, H.J.: A next-generation sequencing method for genotyping by-sequencing of highly heterozygous autotetraploid potato. PLoS One **8**(5), e62355 (2013)
16. Peterson, G.W., Dong, Y., Horbach, C., Fu, Y.B.: Genotyping-by-sequencing for plant genetic diversity analysis: a lab guide for SNP genotyping. Diversity **6**, 665–680 (2014)
17. Poland, J., Endelman, J., Dawson, J., Rutkoski, J., Wu, S., et al.: Genomic selection in wheat breeding using genotyping-by-sequencing. Plant Genome **5**, 103–113 (2012)

18. Sonah, H., Bastien, M., Iquira, E., Tardivel, A., Legare, G., Boyle, B., Normandeau, E., Laroche, J., Larose, S., Jean, M., Belzile, F.: An improved genotyping-by-sequencing (GBS) approach offering increased versatility and efficiency of SNP discovery and genotyping. PLoS One **8**(1), e54603 (2013)
19. Glaubitz, J.C., Casstevens, T.M., Lu, F., Harriman, J., Elshire, R.J., Sun, Q., Buckler, E.S.: TASSEL-GBS: a high capacity genotyping-by-sequencing analysis pipeline. PLoS One **9**(2), e90346 (2014)
20. Langmead, B., Salzberg, S.L.: Fast gapped-read alignment with Bowtie 2. Nat. Methods **9**, 357–359 (2012)
21. Vieira, F.G., Fumagalli, M., Albrechtsen, A., Nielsen, R.: Estimating inbreeding coefficients from NGS data: impact on genotype calling and allele frequency estimation. Genome Res. **23**, 1852–1861 (2013)
22. Smith, C.A.B., Thomson, R.: Estimation of inbreeding from population samples. J. Appl. Probab. **25**, 127–135 (1988)
23. Etter, P.D., Bassham, S., Hohenlohe, P.A., Johnson, E.A., Cresko, W.A.: SNP discovery and genotyping for evolutionary genetics using RAD sequencing. Methods Mol. Biol. **772**, 157–178 (2011)
24. Danecek, P., Auton, A., Abecasis, G., Albers, C.A., Banks, E., Depristo, M.A., Handsaker, R., Lunter, G., Marth, G., Sherry, S.T., McVean, G., Durbin, R., 1000 Genomes Project Analysis Group: The variant call format and VCFtools. Bioinformatics **27**, 2156–2158 (2011)
25. Deejai, P., Assawamakin, A., Wangkumhang, P., Poomputsa, K., Tongsima, S.: On assigning individuals from cryptic population structures to optimal predicted subpopulations: an empirical evaluation of non-parametric population structure analysis techniques. Comput. Syst. Biol. Bioinform. **115**, 58–70 (2010)
26. Pritchard, J.K., Stephens, M., Donnelly, P.: Inference of population structure using multilocus genotype data. Genetics **155**, 945–959 (2000)
27. Alexander, D.H., Novembre, J., Lange, K.: Fast model-based estimation of ancestry in unrelated individuals. Genome Res. **19**(9), 1655–1664 (2009)
28. Frichot, E., Mathieu, F., Trouillon, T., Bouchard, G., Francois, O.: Fast and efficient estimation of individual ancestry coefficients. Genetics **196**(4), 973–983 (2014)
29. Alexander, D.H., Lange, K.: Enhancements to the ADMIXTURE algorithm for individual ancestry estimation. BMC Bioinf. **12**, 246 (2011)
30. Whitley, E., Ball, J.: Statistics review 6: nonparametric methods. Crit. Care **6**, 509–513 (2002)
31. Gao, X., Martin, E.R.: Using allele sharing distance for detecting human population stratification. Hum. Hered. **68**, 182–191 (2009)
32. Nimmakayala, P., Levi, A., Abburi, L., Abburi, V.L., Tomason, Y.R., Saminathan, T., Vajja, V.G., Malkaram, G., Reddy, R., Wehner, T.C., Mitchell, S.E., Reddy, U.K.: Single nucleotide polymorphisms generated by genotyping-by-sequencing to characterize genome-wide diversity, linkage disequilibrium, and selective sweeps in cultivated watermelon. BMC Genomics **15**, 767 (2014)
33. Romay, M.C., Millard, M.J., Glaubitz, J.C., Peiffer, J.A., Swarts, K.L., Casstevens, T.M., Elshire, R.J., Acharya, C.B., Mitchell, S.E., Flint-Garcia, S.A., McMullen, M.D., Holland, J.B., Buckler, E.S., Gardner, C.A.: Comprehensive genotyping of the USA national maize inbred seed bank. Genome Biol. **14**, R55 (2013)
34. Spindel, J., Wright, M., Chen, C., Cobb, J., Gage, J., Harrington, S.: Bridging the genotyping gap: using genotyping-by-sequencing (GBS) to add high-density SNP markers and new value to traditional bi-parental mapping and breeding populations. Theor. Appl. Genet. **126**, 2699–2716 (2013)
35. Jarquín, D., Kocak, K., Posadas, L., Hyma, K., Jedlicka, J., Graef, G., Lorenz, A.: Genotyping by sequencing for genomic prediction in a soybean breeding population. BMC Genomics **15**, 740 (2014)
36. Rocher, S., Jean, M., Castongyay, Y., Belzile, F.: Validation of genotyping-by-sequencing analysis in populations of tetraploid alfalfa by 454 sequencing. PLoS One (2015). doi:10.1371/journal.pone.0131918

37. Huang, Y.F., Poland, J.A., Wight, C.P., Jackson, E.W., Tinker, N.A.: Using genotyping-by-sequencing (GBS) for genomic discovery in cultivated oat. PLoS One **9**(7), e102448 (2014)
38. Pootakham, W., Jomchai, N., Ruang-areerate, P., Shearman, J.R., Sonthirod, C., Sangsrakru, D., Tragoonrung, S., Tangphatsornruang, S.: Genome-wide SNP discovery and identification of QTL associated with agronomic traits in oil palm using genotyping-by-sequencing (GBS). Genomics **105**, 288–295 (2015)
39. Rabbi, Y.I., Kulakow, P.A., Manu-Aduening, J.A., Dankyi, A.A., Asibuo, J.Y., Parkes, E.Y., Abdoulaye, T., Girma, G., Gedil, M.A., Ramu, P., Reyes, B., Maredia, M.K.: Tracking crop varieties using genotyping by-sequencing markers: a case study using cassava (Manihot esculenta Crantz). BMC Genet. **16**, 115 (2015)
40. Girma, G., Hyma, K.E., Asiedu, R., Mitchell, S.E., Gedil, M., Spillane, C.: Next generation sequencing based genotyping, cytometry and phenotyping for understanding diversity and evolution of guinea yams. Theor. Appl. Genet. **127**, 1783–1794 (2014)
41. Bielenberg, D.G., Rauh, B., Fan, S., Gasic, K., Abbott, A.G., Reighard, G.L., Okie, W.R., Wells, C.E.: Genotyping by sequencing for SNP-based linkage map construction and QTL analysis of chilling requirement and bloom date in peach [Prunus persica (L.) Batsch]. PLoS One (2015). doi:10.1371/journal.pone.0139406
42. Ma, X.F., Jensen, E., Alexandrov, N., Troukhan, M., Zhang, L., Thomas-Jones, S., Farra, K., Clifton-Brown, J., Donnison, I., Swaller, T., Flavell, R.: High resolution genetic mapping by genome sequencing reveals genome duplication and tetraploid genetic structure of the diploid Miscanthus sinensis. PLoS One **7**(3), e33821 (2012)
43. Pan, J., Wang, B., Pei, Z.Y., Zhao, W., Gao, J., Mao, J.F., Wang, X.R.: Optimization of the genotyping-by-sequencing strategy for population genomic analysis in conifers. Mol. Ecol. Resour. **15**, 711–722 (2015)
44. Jeong, H.S., Jang, S., Han, K., Kwon, J.K., Kang, B.C.: Marker-assisted backcross breeding for development of pepper varieties (Capsicum annuum) containing capsinoids. Mol. Breed. **35**, 226 (2015)
45. Moscone, E.A., Scaldaferro, M.A., Grabiele, M., Cecchini, N.M., Sanchez Garcıa, Y., Jarret, R., Davina, J.R., Ducasse, D.A., Barboza, G.E., Ehrendorfer, F.: The evolution of chili peppers (Capsicum-Solanaceae): a cytogenetic perspective. Acta Hortic. **745**, 137–170 (2007)
46. Hernández-Verdugo, S., Luna-Reyes, R., Oyama, K.: Genetic structure and differentiation of wild and domesticated populations of Capsicum annuum (Solanaceae) from Mexico. Plant Syst. Evol. **226**(3–4), 129–142 (2001)
47. Cericola, F., Portis, E., Toppino, L., Barchi, L., Acciarri, N., Ciriaci, T., Sala, T., Rotino, G.L., Lanteri, S.: The population structure and diversity of eggplant from Asia and the Mediterranean Basin. PLoS One **8**(9), e73702 (2013)
48. Rodriguez, M., Rau, D., Bitocchi, E., Bellucci, E., Biagetti, E., Carboni, A., Biagetti, E., Carboni, A., Gepts, P., Nanni, L., Papa, R., Attene, G.: Landscape genetics, adaptive diversity and population structure in Phaseolus vulgaris. New Phytol. **209**(4), 1781–1794 (2015)
49. Kim, S., Park, M., Yeom, S.I., Kim, Y.M., Lee, J.M., Lee, H.A., Seo, E., Choi, J., Cheong, K., Kim, K.T., Jung, K., Lee, G.W., Oh, S.K., Bae, C., Kim, S.B., Lee, H.Y., Kim, S.Y., Kim, M.S., Kang, B.C., Jo, Y.D., Yang, H.B., Jeong, H.J., Kang, W.H., Kwon, J.K., Shin, C., Lim, J.Y., Park, J.H., Huh, J.H., Kim, J.S., Kim, B.D., Cohen, O., Paran, I., Suh, M.C., Lee, S.B., Kim, Y.K., Shin, Y., Noh, S.J., Park, J., Seo, Y.S., Kwon, S.Y., Kim, H.A., Park, J.M., Kim, H.J., Choi, S.B., Bosland, P.W., Reeves, G., Jo, S.H., Lee, B.W., Cho, H.T., Choi, H.S., Lee, M.S., Yu, Y., Do Choi, Y., Park, B.S., van Deynze, A., Ashrafi, H., Hill, T., Kim, W.T., Pai, H.S., Ahn, H.K., Yeam, I., Giovannoni, J.J., Rose, J.K., Sørensen, I., Lee, S.J., Kim, R.W., Choi, I.Y., Choi, B.S., Lim, J.S., Lee, Y.H., Choi, D.: Genome sequence of the hot pepper provides insights into the evolution of pungency in Capsicum species. Nat. Genet. **46**(3), 270–278 (2014)
50. Qin, C., Yu, C., Shen, Y., Fang, X., Chen, L., Min, J., Cheng, J., Zhao, S., Xu, M., Luo, Y., Yang, Y., Wu, Z., Mao, L., Wu, H., Ling-Hu, C., Zhou, H., Lin, H., González-Morales, S., Trejo-Saavedra, D.L., Tian, H., Tang, X., Zhao, M., Huang, Z., Zhou, A., Yao, X., Cui, J., Li, W., Chen, Z., Feng, Y., Niu, Y., Bi, S., Yang, X., Li, W., Cai, H., Luo, X., Montes-Hernández, S., Leyva-González, M.A., Xiong, Z., He, X., Bai, L., Tan, S., Tang, X., Liu, D., Liu, J., Zhang,

S., Chen, M., Zhang, L., Zhang, L., Zhang, Y., Liao, W., Zhang, Y., Wang, M., Lv, X., Wen, B., Liu, H., Luan, H., Zhang, Y., Yang, S., Wang, X., Xu, J., Li, X., Li, S., Wang, J., Palloix, A., Bosland, P.W., Li, Y., Krogh, A., Rivera-Bustamante, R.F., Herrera-Estrella, L., Yin, Y., Yu, J., Hu, K., Zhang, Z.: Whole-genome sequencing of cultivated and wild peppers provides insights into Capsicum domestication and specialization. Proc. Natl. Acad. Sci. U. S. A. **111**(14), 5135–5140 (2014)

51. Evanno, G., Regnaut, S., Goudet, J.: Detecting the number of clusters of individuals using the software STRUCTURE: a simulation study. Mol. Ecol. **14**, 2611–2620 (2005)

Hybridization-Based Enrichment and Next Generation Sequencing to Explore Genetic Diversity in Plants

Irma Terracciano, Concita Cantarella, and Nunzio D'Agostino

Abstract In plant research, targeted re-sequencing of enriched genomic DNA regions has become a scalable and cost-effective method for the discovery of genome-wide sequence variations to be exploited to address different biological questions.

In this manuscript, we describe the main strategies to reduce genome complexity in plants with a special focus on hybridization-based enrichment methods. Then, we provide an overview of applications of target enrichment-based next generation sequencing (NGS) protocols in plant genetics and illustrate major bioinformatic approaches and tools for the analysis of NGS data, the identification of sequence polymorphisms, and their annotation to predict possible biological effects. Finally, we introduce our research activity on the use of hybridization-based target enrichment system for the identification of interesting sequence variations at candidate genes controlling carotenoid biosynthesis in tomato.

Keywords Sequence capture • Target enrichment • Sequence polymorphisms • Variant calling • Allele mining

1 Introduction

Next generation sequencing (NGS) technologies have experienced an extraordinary increase in capacity and a significant advancement in data generation. In plant research, large datasets are now being generated across various model and non-model species by sequencing whole genomes. Recently, whole exome sequencing (WES) provided a cost-effective alternative aiming at the identification of nucleotide variability across the exome [1, 2], which is defined as the sequences encompassing all the exons of protein-coding genes in a genome. When there is interest on specific candidate *loci*, target enrichment (in which genomic regions are selectively captured from a DNA sample before sequencing) offers substantial reduction in sequencing

I. Terracciano • C. Cantarella • N. D'Agostino (✉)
Consiglio per la ricerca in agricoltura e l'analisi dell'economia agraria (CREA) - Centro di ricerca per l'orticoltura, via dei Cavalleggeri 25, 84089 Pontecagnano Faiano (SA), Italy
e-mail: nunzio.dagostino@crea.gov.it

A. Rogato et al. (eds.), *Dynamics of Mathematical Models in Biology*,
DOI 10.1007/978-3-319-45723-9_10

space and cost [3–6]. The hybridization-based method is one of the most efficient and widely adopted among the available target enrichment techniques [7, 8]. It has been demonstrated powerful, independently of the DNA capture protocol and the sequencing platform used and it is often replacing PCR as the main target enrichment method in plant sciences [3, 5].

Plant genomes can be extremely complex, repetitive, and are often polyploids; as a consequence, some species are not well suited for whole genome sequencing (WGS) approaches. By contrast, sequence capture and targeted re-sequencing have the advantage of providing higher read depth for individual *locus* and support the accurate identification of nucleotide polymorphisms also in plants with large genomes and higher ploidy levels [9, 10].

In this manuscript, we provide a brief overview of the available strategies to reduce genome complexity in plants with a special focus on hybridization-based enrichment methods currently used for the characterization of natural/induced genetic variation in plant species. Then, we highlight possible applications of these technologies to plant research and describe a typical bioinformatic workflow for the analysis of NGS data and the identification of sequence polymorphisms. Finally, we discuss our experience in a project aimed at the identification of naturally occurring sequence variation at candidate genes controlling carotenoid biosynthesis in tomato.

2 Strategies to Reduce Genome Complexity in Plants: Target Enrichment

For plants that possess large size or polyploid genomes, for which whole genomes cannot be readily assembled and the analysis of a large number of individuals results still very expensive, an alternative strategy to WGS is to generate a reduced representation of the genome. Genome reduction can be obtained using target enrichment strategies. Target enrichment consists in the isolation of specific genomic *loci* (e.g., genes, molecular markers, larger genomic regions, and organelle genomes) coupled with NGS. Compared to WGS, the reduction in sequencing space entails three main advantages: (1) sample multiplexing that implicates an overall reduction of the sequencing cost per sample; (2) significant reduction in the complexity of the analysis; and (3) the possibility of identifying the precise region of interest given the depth of sequencing provided by NGS.

At present, transcriptome-based, restriction enzyme-based, PCR-based, and hybridization-based methods, all compatible with the most popular NGS platforms, have been developed to enrich specific targets [3].

Transcriptome-Based Enrichment is one of the most widely used strategies to reduce genome complexity, since it focuses only on the transcribed portion of the genome. The key aim of transcriptome sequencing, also known as RNA-seq, is to determine gene expression profiles of each transcript during development and under different conditions [11]. SNP discovery and molecular marker development via

RNA-seq are often performed, especially in organisms with large genomes [12]. Noteworthy, since RNA-seq is independent from any a priori knowledge on the genome sequence of the species under investigation, it allows the analysis of poorly characterized species.

Restriction Enzyme-Based Enrichment makes use of the discriminatory power of the restriction endonucleases to produce restriction fragments among individuals in a population. Three main techniques have been developed so far: **RAD-seq** (restriction-site associated DNA sequencing) [13, 14], **GR-RSC** (genomic reduction based on restriction site conservation) [15], and **GBS** (genotyping-by-sequencing) [16]. All these methods, reviewed by Cronn et al. [3], are flexible and quite inexpensive and have been used to identify and score, in a group of individuals, thousands of genetic markers randomly distributed along the genome enabling SNP discovery, genotyping as well as quantitative genetic and phylo-geographic studies.

PCR-Based Target Enrichment includes the direct sequencing of small and long PCR products. NGS of PCR fragments has been preferentially applied to chloroplast genomes in systematic studies [17] and in some cases also to nuclear genomic regions despite their complexity [18]. The main disadvantages associated with this method are the high level of failed target amplifications and/or non-specific amplifications as well as the difficulty in obtaining an accurate pooling of samples for NGS multiplexing [5]. Anyway, PCR-based enrichment remains feasible for targeting small to medium-sized regions of the genome, but for high-throughput sequencing of tens of thousands of PCR amplicons its efficiency falls off, given the initial cost per sample and challenges in sample multiplexing. Microfluidic-based multiplexing PCR can reduce costs but continues to be more expensive than other enrichment methods [3].

Hybridization-Based Enrichment or sequence capture methods exploit the high specificity of DNA or RNA probes (also called baits) which are designed to be complementary to target genomic regions. RNA baits have significant advantages over DNA probes because RNA–DNA hybrids have a higher affinity and melting temperature than DNA–DNA hybrids. Two main technologies have been developed for hybrid-capture applications: (1) **on-array-** or **solid-based hybridization** which implies sample hybridization on a solid support (i.e., glass slide, microarray) [8] and (2) **in-solution-** or **liquid-based hybridization** where pooled baits are used in reaction tubes [7]. Due to their moderate costs and high specificity, low amounts of required DNA per sample and power to simultaneously target large numbers of markers, several protocols and commercial kits have been developed. The most widespread ones and reliable in studies on plant species were provided by Agilent Technologies (SureSelect), Roche NimbleGen (SeqCap EZ), MYcroarray (MYbaits),and Ion Torrent (TargetSeq). Distinguishing features of these sequence capture platforms are reported in Table 1.

Table 1 List of the most important features of the commercially available target enrichment kits

	On-array hybridization-based capture		In-solution hybridization-based capture			
	NimbleGen Sequence Capture Array	Agilent Microarray	NimbleGen SeqCap EZ	Agilent SureSelect	MYcroarray MYbait	IonTorrent TargetSeq
Bait type	DNA	DNA	DNA	RNA	RNA	DNA
Bait length	60 bp	60 bp	55–105 bp	114–126 bp	80–120 bp	50–120 bp
Target size	Up to 30 Mb	N.D.	Up to 200 Mb	From 1 kb to 24 Mb	Up to 200,000 baits	From 100 kb up to 10 Mb

N.D. not determined

All these providers offer the opportunity to design custom kits for the species of interest and make available services and tools to support probe design. Evidently, it is necessary to have a reference sequence (complete or draft genome sequence, transcripts, Expressed Sequence Tag database, etc.) to accomplish this task.

3 Technical Considerations on Hybridization-Based Enrichment Methods

As previously mentioned, affordable costs combined with ease of use, multiplexing capacity, and scalability (from few genes or genomic regions to entire exomes) make sequence capture an attractive alternative to large-scale PCR and a method of choice for its wide potential use ranged from intra-specific population studies, typically for polymorphism identification, to deeper-level phylo-genomics.

Several authors compared the most popular sequence capture technologies (both liquid- and solid-phase) demonstrating that, although there are slight differences, results in terms of coverage efficiency, accuracy in genotype assignment, and variant discovery are basically very similar [19, 20].

However, liquid-based sequence capture systems are gradually replacing on-array-based hybridization methods because all the reaction steps of the protocol take place in a single tube making the process scalable to large numbers of samples and suitable for robotic automation. Furthermore, it requires less input DNA and simple laboratory equipment [7].

Regardless of the method used, several technical aspects and potential drawbacks must be taken into account in order to plan a successful target enrichment project and achieve predetermined objectives.

Ploidy level, genome size, and DNA compositional properties (e.g., high GC content) of the species under investigation together with several features of the baits (e.g., probe length, hybridization temperature) can affect enrichment efficiency. It has been demonstrated that enrichment efficiency level can be considerably reduced in promoter regions, 5′ UTR regions, and in the first exon of genes because of high GC content of these regions [21]. High or low GC content reduces the efficiency of PCR amplifications [22], bait synthesis, and hybridization. Since this latter aspect is related to nucleotide compositional properties of the probes, it can somehow be corrected by probe design. The GC bias effect on sequencing coverage has been studied by different authors, who plot GC content distribution against the normalized mean read depth [19, 23]. Enrichment efficiency depends also on the sequence capture protocol of choice as well as on the sequencing technology used.

The percentage of sequences that map to the selected targets (probe specificity) can be influenced by the presence of closely related sequences (orthologs/paralogs) of duplicated regions and/or interspersed repetitive elements in the genome [3]. Minimizing the number of off-target reads is desirable and it can be achieved by selecting probes with high specificity.

A crucial parameter of a sequence capture experiment is the sensitivity, which is the percentage of the target bases that are represented by one or more sequenced reads. In other words, the higher the sequencing depth, the higher the confidence that the base called at that position is correct, the better the estimation of SNP/InDel frequency for any particular SNP/InDel. Also the experimental design has a great impact on enrichment efficiency. Effectively, the right balance between the numbers of targets to be sequenced and the expected sequencing depth must be found.

4 Overview of Hybridization-Based Enrichment Applications in Plant Species

In the last few years, sequence capture and target enrichment followed by NGS have been used to identify a high number of mutations in whole exomes, selected gene families, and target genes or genomic regions of many plant species allowing (1) generation of useful polymorphism resources in a quick and rather inexpensive way; (2) biodiversity exploration and mining; (3) SNP marker development and generation of genetic maps; (4) population structure definition or evolutionary history in phylogenetics and phylogeography studies to be tracked; (5) QTL mapping and candidate gene identification; and (6) genomic selection.

All these applications are intended to accelerate plant breeder activity for crop improvement.

In this section we review recent literature on targeted re-sequencing of enriched genomic DNA regions in crops and other economically important plant species and briefly describe objectives and applications of each study (Table 2).

The first application of hybridization-based sequence capture in plant was published by Fu et al. [9] who demonstrated the effectiveness of the enrichment protocol in the identification of plant polymorphisms in divergent maize (*Zea mays*) lines.

Table 2 List of several examples from the recent literature on the applications of hybridization-based enrichment protocols to plant genetics

Species	N°	Plant material	Target	Target size	Assay type	Hybridization-based technology	NGS platform	References
Maize	1	2 inbred lines	Non-repetitive portion of 2.2 Mb and 43 genes	4.3 Mb	Roche NimbleGen array	SB	Roche 454	Fu et al. [9]
	2	21 inbred lines	Genomic regions including genes for biomass production	29 Mb	Roche NimbleGen array	SB	Roche 454	Muraya et al. [24]
Rapeseed canola	3	10 genotypes	47 QTL-associated genomic regions	51.2 Mb	Roche NimbleGen array and undefined liquid platform	SB and LB	Roche 454 and Illumina	Clarke et al. [25]
	4	4 accessions	29 regulatory flowering-time genes	614 kb	Agilent SureSelect	LB	Illumina	Schiessl et al. [26]
Cotton	5	2 accessions	1000 geni (500 pairs of homeologs)	550 kb	Roche NimbleGen array	SB	Roche 454	Salmon et al. [27]
Sugarcane	6	2 genotypes (one *Saccharum officinarum* and one *Saccharum* hybrid)	Gene-rich regions from the close relative *Sorghum bicolor*	5.8 Mb	Agilent SureSelect	LB	Illumina	Bundock et al. [28]
Swichgrass	7	4 tetraploid lowland and 4 octoploid upland cultivars	Whole exome	50 Mb	Roche NimbleGen Seq-EZ	LB	Illumina	Evans et al. [29]
Cassava	8	100 F1 progeny and 2 parental strains	27,469 biallelic SNPs from 10,105 regions	2.49 Mb	Ion TargetSeq Life Technologies	LB	Ion Torrent Proton System	Pootakham et al. [30]
Loblolly pine	9	24 haploid samples	Whole exome	6.57 Mb	Agilent SureSelect	LB	Illumina	Neves et al. [31]
	10	72 haploid samples from a mapping population	Whole exome	6.57 Mb	Agilent SureSelect	LB	Illumina	Neves et al. [32]

Black cottonwood	11	48 genotypes	Predicted exons and upstream regulatory regions and random genomic intervals	20.76 Mb	Agilent SureSelect	LB	Illumina	Zhou and Holliday [23]
Eucalyptus	12	3 genotypes	94 genes involved in xylogenesis	N.D.	Agilent custom array	SB	Illumina	Dasgupta et al. [33]
Wheat	13	The wild emmer accession (*Triticum dicoccoides*) and durum wheat cultivar Langdon (*Triticum turgidum var. durum*)	Exon regions of 3497 genes	3.5 Mb	Agilent SureSelect	LB	Illumina	Saintenac et al. [10]
	14	8 UK alloexaploid varieties	Significant proportion of the wheat exome	56.5 Mb	Roche NimbleGen array	SB	Illumina	Allen et al. [34]
	15	8 UK alloexaploid varieties	Significant proportion of the wheat exome	56.5 Mb	Roche NimbleGen array	SB	Illumina	Winfield et al. [35]
	16	2 RILs chosen as parent for a F2 mapping population	Gene-rich regions of the genome	110 Mb	Roche NimbleGen Seq-EZ	LB	Illumina	Gardiner et al. [36]
Rice and wheat	17	EMS-mutagenized rice (72) and wheat (6) individuals	Whole exome	42 Mb (rice) 107 Mb (wheat)	Roche NimbleGen Seq-EZ	LB	Illumina	Henry et al. [37]
Soybean	18	4 fast-neutron mutants	Whole exome	52.3 Mb	Roche NimbleGen array	SB	Illumina	Bolon et al. [38]
	19	2 individuals of the cultivar Williams 82	Whole exome	52.3 Mb	Roche NimbleGen array	SB	Illumina	Haun et al. [39]
Barley	20	36 samples from 13 barley cultivars and 7 samples from 3 wild relatives	Whole exome	61.6 Mb	Roche NimbleGen Seq-EZ	LB	Illumina	Mascher et al. [40]

(continued)

Table 2 (continued)

Species	N°	Plant material	Target	Target size	Assay type	Hybridization-based technology	NGS platform	References
	21	Parental lines and BC_1F_2 lines enriched for early flowering genotypes	Whole exome	61.6 Mb	Roche NimbleGen Seq-EZ	LB	Illumina	Pankin et al. [41]
	22	18 mutant plants and 30 randomly selected wild type plants	Whole exome	61.6 Mb	Roche NimbleGen Seq-EZ	LB	Illumina	Mascher et al. [40]
	23	3 Hv/Hb ILs and the respective donor lines	Whole exome	61.6 Mb	Roche NimbleGen Seq-EZ	LB	Illumina	Wendler et al. [42]
Tribe trifoliae	24	6 individuals representing major lineages within Medicago and Melilotus	62 low-copy nuclear genes and 257 short loci (exon sequences) distributed across all Medicago chromosomes	185 kb	MYbaits MYcroarray	LB	Illumina	de Sousa et al. [43]
Compositae	25	15 species	763 conserved ortholog set (COS) loci	N.D.	MYbaits MYcroarray	LB	Illumina	Mandel et al. [44]
Rosidae, asteridae, caryophyllales, asparagaceae, and poaceae	26	24 species (22 eudicots and 2 monocots)	300 plastidial genomes	N.D.	Agilent SureSelect	LB	Illumina	Stull et al. [45]

Strawberry	27	48 F1 individuals from MRD30xMDR60 and their parental lines	200 bp surrounding each of 6575 previously identified polymorphisms	149 Mb	MYbaits MYcroarray	LB	Illumina	Tennessen et al. [46]
Potato and tomato	28	Rpi-ber2 and Rpi-rzc1 F1 populations and *Solanum tuberosum* group Phureja clone DM1-3 516 R44	580 NB-LRR coding sequence from *Solanaceae*	N.D.	Agilent SureSelect	LB	Illumina	Jupe et al. [47]
Tomato	29	*Solanum pimpinellifolium* LA1589 and *Solanum lycopersicum* Heinz 1706	743 NB-LRR-like sequences	N.D.	Agilent SureSelect	LB	Illumina	Andolfo et al. [48]
Potato	30	83 tetraploid cultivars and 1 monoploid clone (DM 1-3 511)	807 genes	1.44 Mb	Agilent SureSelect	LB	Illumina	Uitdewilligen et al. [49]

LB liquid-based, *SB* solid-based, *N.D.* not determined

In the course of recent years, different authors proved the efficacy of exome capture for the investigation of nucleotide diversity in polyploid species with a large, repetitive, and heterozygous genomes [10, 29, 34, 35] and of intra-cultivar genomic heterogeneity in diploid species [39, 40].

Sequence capture assays have also been designed to target genomic regions associated with agronomically important traits and capture DNA sequence diversity in maize [24], rapeseed (*Brassica napus*) [25, 26], cotton (*Gossypium hirsutum*) [27], and cassava (*Manihot esculenta*) [30] to generate novel data for both research and breeding activities.

In addition, sequence capture and re-sequencing have been applied to several tree species, namely loblolly pine (*Pinus taeda*), black cottonwood (*Populus trichocarpa)*, and eucalyptus (*Eucalyptus globulus*), in order to identify sequence polymorphisms to be used for the generation of a dense reference gene-based genetic map [31, 32], perform genotyping [23], and develop xylogenesis associated-trait markers [33].

In order to reconstruct phylogenetic relationships across the Trifolie tribe [43] and the Compositae family [44] hybridization-based enrichment has been used to capture sequence variability within low-copy nuclear (LCN) and conserved ortholog set (COS) markers, respectively.

Mapping-by-sequencing, which combines genetic mapping with targeted-re-sequencing, has been exploited (1) to identify useful polymorphism to map candidate genes in barley (*Hordeum vulgare*) [41], wild strawberry (*Fragaria vesca ssp. bracteata*) [46], and einkorn wheat (*Triticum monococcum*) [36] and (2) to detect the precise allocation of *Hordeum bulbosum* introgression regions in the cultivated *H. vulgare* genetic background [42].

WES has been used to re-sequence ethyl methanesulfonate (EMS)- and fast neutron (FN)-mutagenized plant populations to discover induced mutations in rice (*Oryza sativa*), bread wheat (*Triticum aestivum*) [37], and soybean (*Glycine max*) [38].

The use of a closely related reference genome (i.e., *Sorghum bicolor*) for probe design has been applied to capture genomic regions of two sugarcane (*Saccharum officinarum*) genotypes [28] proving useful for polymorphism discovery in poorly described species.

Recently, even chloroplast genomes were subjected to target enrichment and massively parallel sequencing [45]. The strategy the authors adopted is based on the design of a custom RNA probe set based on the complete sequences of 22 previously sequenced eudicot chloroplast DNAs. Using this probe set an enrichment experiment was performed on 24 angiosperms (22 eudicots, 2 monocots), which were subsequently sequenced leading to the generation of complete plastid genomes with exceptionally high coverage (717× on average).

At present, very few studies are referred to solanaceous crops. In 2013, [49] described liquid-phase capture method to identify sequence variants within and across 84 potato (*Solanum tuberosum*) cultivars they afterwards used to genotype the same plant material by genotyping-by-sequencing.

In both tomato (*Solanum lycopersicum*) and potato, resistance gene enrichment and sequencing (RenSeq) has been used to discover and annotate novel NB-LRR genes [47, 48] that will undoubtedly provide breeders with a valuable tool to identify novel disease resistance traits.

5 Major Bioinformatic Strategies and Tools for Genome-Wide Sequence Variant Discovery

The demand of targeted re-sequencing is paralleled by the development of bioinformatic tools to analyze sequence data, with more than 500 tools published within a span of only 2 years [50]. Some of them were specifically developed to handle sequence data from targeted re-sequencing experiments and are constantly being improved and updated. A multi-step analysis performed with a combination of various tools is a general prerequisite to extract meaningful results from sequence capture and targeted re-sequencing experiments (Fig. 1). The first step of the analysis includes the evaluation of read quality. The command-line tools FastQC (http://www.bioinformatics.babraham.ac.uk/projects/fastqc/), FASTX-Toolkit (http://hannonlab.cshl.edu/fastx_toolkit/), and Trimmomatic [51], are often combined during the primary analysis to assessing the overall quality of a sequencing run, to trim off poor quality bases, and to filter on high quality scores. The next step is critical and involves the alignments of high quality reads to a reference genome or transcriptome when present. Read-to-reference alignments are released in a compressed, indexed, binary form called Binary sequence Alignment/Map format (BAM) [52].

The tool chosen for the alignment of the reads to the reference as well as the compositional properties of the reference sequence itself (e.g., presence of low complexity sequences, repetitive regions) affect the number of reads properly aligned and often influence final coverage and depth values [53]. Different algorithms dedicated to the mapping of short reads to a reference sequence have been developed so far [54], being the most popular BWA [55], Bowtie2 [56], and SOAP2 [57].

Alignments of the reads to the reference genome/transcriptome are generally used to estimate the coverage along target regions and to assess the level of on-target enrichment efficiency. In this regard, the *coverage* utility of the bedtools package [58] is normally used. Given a BAM alignment file and a BED file (a tab-delimited text file that defines the coordinates of a feature along the reference sequence) containing target regions, it computes the coverage over defined intervals. This allows evaluating if the coverage depth is uniform among the re-sequenced genotypes and at what extent the variation in coverage affects target regions. Indeed, some off-target reads are expected and this depends on the nature of the target sequences (e.g., genic, intergenic, etc.) as well as on the enrichment technology and sequencing platform used.

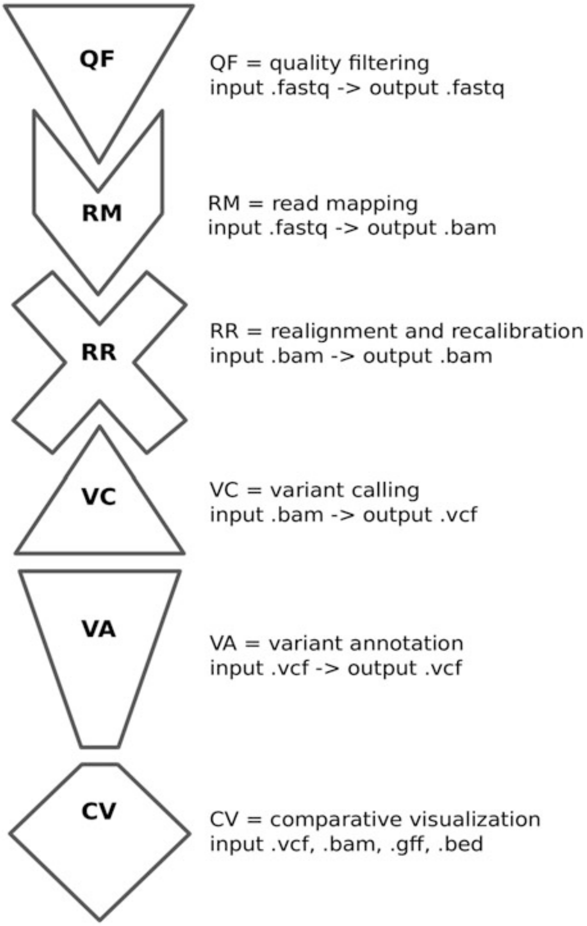

QF
QF = quality filtering
input .fastq -> output .fastq
RM
RM = read mapping
input .fastq -> output .bam
RR
RR = realignment and recalibration
input .bam -> output .bam
VC
VC = variant calling
input .bam -> output .vcf
VA
VA = variant annotation
input .vcf -> output .vcf
CV
CV = comparative visualization
input .vcf, .bam, .gff, .bed

Fig. 1 Typical variant calling workflow. Different analysis steps (each object in the figure) are concatenated to identify reliable sequence polymorphisms and derive meaningful biological interpretation of the results. The "RR" step is facultative in a variant calling NGS analysis

A BAM file can include reads with the same start and end coordinates. These might represent PCR duplicates, which should be removed/flagged in the BAM file since they are not informative and should not be counted as evidence of a putative variant. The Picard MarkDuplicates (http://picard.sourceforge.net) is the preferred tool for this task, although it only considers the starting position of the read as a way to indicate a putative duplicated read. As an alternative, the SAMtools "rmdup" command can be used [52].

Reads mapping to the edges of InDels often lead to mis-alignments and produce artifactual mis-matches. Therefore, the local re-alignment of the reads around InDels is necessary because it helps improve the accuracy of downstream processing steps. The strategy developed to accomplish this task combines short-read mapping with an assembly inspired approach to identify a local consensus sequence. Programs that implement this approach include SRMA [59] and IndelRealigner from the Genome Analysis Toolkit (GATK) [60].

A further improvement may be achieved running the base quality score recalibration (BQSR) on re-aligned BAM files. One of the most commonly used programs for BQSR is BaseRecalibrator from the GATK suite [60].

Variant calling, at first glance, may be pretty simple, as it involves the identification of sites where one or more samples display possible genomic variations. All the available tools allow the minimum coverage and minimum variant frequency threshold to be fixed in order to extract significant variants. Of course additional parameters can be configured to compute more stringent analyses. The GATK HaplotypeCaller or UnifiedGenotyper [61], SAMtools mpileup [52], and Freebayes (https://github.com/ekg/freebayes) are the most widely used programs for sequence variant calling. The Variant Call File (VCF) format allows the most prevalent types of sequence variations to be stored. In order to provide easily accessible methods for working with VCF files the VCFtools program package has been developed [62]. A binary representation of the variant call format (BCF), which is more compact and much faster to be processed than VCF, has also been implemented [63]. A limitation of VCF tools is not supporting filtering of polyploidy data, but this can be accomplished by VCFlib (https://github.com/vcflib/vcflib).

Identifying functionally relevant polymorphisms in a *mare magnum* of genetic variations is the major challenge. Annotation of sequence variants includes the classification of the effects of single nucleotide polymorphisms and insertion–deletions (e.g., synonymous or non-synonymous SNPs, start-codon gain/loss, stop-codon gain/loss, frame-shift, etc.) on annotated genes. Annotations can also be based on the coordinate system used to describe the genomic position of each polymorphism (e.g., intronic, 5′ or 3′ un-translated region, upstream, downstream, inter-genic regions, etc.). In this regard, it is crucial to have an accurate, preferably gold standard, structural annotation of the reference genome. ANNOVAR [64] and SnpEff [65] are two of the most used tools in the variant annotation process.

Sequence polymorphisms in the coding regions are frequently associated with aberrant protein modifications. The interpretation of novel *missense* mutations (a type of non-synonymous substitutions) is challenging. Nevertheless, several computational tools have been developed in order to predict possible impact of an amino acid substitution on the structure and function of proteins [66, 67]. More recently, the six best performing tools were combined into a consensus classifier, called PredictSNP [68], which predictions on protein-related mutations represent a robust alternative to the predictions delivered by individual tools. More complicated is the study of splice-site polymorphisms as well as of sequence variations within intronic regions. It is known that nucleotide variants very close to splice junctions might alter the splicing pattern of a gene and/or affect splicing efficiency as well as that introns can harbor functional polymorphisms that can influence the expression of the genes that host them [69]. However, to the best of our knowledge, no tools are available for the automatic classification of the effects of such polymorphisms in plants. By contrast, strategies and tools to support investigations on promoters are well-defined. Indeed, regulatory regions are generally scanned to identify transcription factor binding sites (TFBSs). SNPs and InDels within these regions might modify the TFBS pattern and alter gene expression. A variety of databases have been established during the time to collect *cis-acting* regulatory DNA elements found in plant promoters. Most of them are now integrated into the most recent PlantPAN resource [70]. Of course, in silico predictions must be always interpreted with caution and additional experimental evidences are needed to confirm sequence variations within the identified alleles.

The last step of the workflow consists in the visual representation of NGS data. This can be amazingly useful when interpreting the obtained results. The integrative genomic viewer (IGV) supports users by displaying, along a reference genome, aligned reads (BAM files) and predicted genetic variants (VCF files) combined with annotations from the reference [71]. Aggregation of data on a single platform has significant consequences in the meaningful interpretation of sequencing data and it is essential to facilitate knowledge discovery.

6 Use of Liquid-Phase Sequence Capture and Target Enrichment System for Allele Mining in Tomato

Allele mining is a promising strategy to dissect allelic variation at candidate genes controlling key agronomic traits for potential crop breeding applications. Several factors determine a successful and efficient allele mining activity; mainly the availability of (1) information on genome and gene sequences for the species under investigation; (2) efficient and reliable phenotyping techniques; (3) high-throughput methods for easy generation of allelic data points; (4) cost-effective sequencing platforms, and (5) efficient bioinformatic tools for the identification of nucleotide variations and molecular marker development [72].

Allele mining approaches for traits associated to yield, quality, and important disease resistance have been successfully applied to many crop species, including tomato (*S. lycopersicum*) [73]. Tomato is the most widely consumed vegetable in the world and its fruits are an important source of bioactive compounds with known beneficial effects on human health [74]. Carotenoids are the major class of antioxidant compounds in ripe tomato fruits. They regulate pigmentation of many fruits and flowers and are involved in photo-reception and photo-protection mechanisms [75]. In plants, the carotenoid biosynthetic pathway is located into plastids. In tomato this pathway is highly active during fruit ripening leading to the accumulation of several metabolites, mainly α-/β-carotene and lycopene. Because a wide natural genetic variability associated with the accumulation of carotenoid pigments in the fruit exists across tomato species [76, 77], its exploration can be very useful to undertake breeding programs for tomato fruits bio-fortification [78]. The availability of the tomato genome sequence [79], combined with the existing genetic resources and genomic tools, has undoubtedly expedited the investigation on the genetic variability in large populations of individuals to identify sequence variations across candidate genes.

Aiming to contribute in this scenario, we are performing a research activity in order to capture interesting genetic variation affecting genes responsible for carotenoid accumulation in tomato fruits.

We decided to apply a liquid-phase sequence capture followed by Illumina re-sequencing to target a panel of genes responsible for β-carotene and lycopene accumulation in ripe tomato berries.

The version 2.50 of the *S. lycopersicum* reference genome, the 2.40 iTAG (international Tomato Annotation Group) annotation, and the Agilent's Sure design software have been used to generate 120mer RNA baits to cover all the target genes and their the regulatory regions. Then, the SureSelect target enrichment system (Agilent Technologies) for Illumina paired-end sequencing has been used to capture 230 kb target region in a panel of 48 genotypes differing for carotenoid content in the ripe tomato fruits as determined through RP-HPLC analysis. Illumina data were processed according to the workflow described in the previous paragraph. About 10,000 polymorphisms, including both SNPs and InDels, were identified and annotated to predict their biological effects (Terracciano et al.; manuscript in preparation). The association between the identified genotypic variation and the observed phenotypic variability is ongoing.

We are going to provide experimental validations of interesting mutations that we feel could be employed to generate improved tomato varieties for fruit quality.

7 Conclusions

Liquid- or solid-phase sequence capture and target enrichment coupled with NGS have been proven reliable in the identification of sequence polymorphisms in whole exomes, target genes, or genomic regions of many plant species. The demand of targeted re-sequencing is constantly growing and requires significant effort in data analysis and management. Bioinformatic strategies are essential to extract meaningful results from raw sequence data. Although all the steps of the complex workflow are well defined, the tools developed to accomplish basic tasks are constantly being improved and updated. Nevertheless, challenges associated with data analysis can be taken on with confidence. Indeed, several applications intended to accelerate plant-breeding activities for crop improvement can benefit from using this technology: these include genotyping, SNP marker development and biodiversity exploration, mapping-by-sequencing, etc. An additional application of sequence capture is the identification and characterization of novel alleles from non-reference genomes. By describing our research activity on the identification of sequence variation across a panel of tomato genotypes at candidate genes controlling carotenoid biosynthesis, we demonstrated that in solution-based hybridization method could be successfully applied to detect and study the effect of novel alleles in economically important crops.

Acknowledgements This work was carried out in the frame of the "GenoPom-pro - Integrating post-genomic platforms to enhance the tomato production chain" project (PON02_00395_3082360) and is supported by the PON R&C 2007–2013 grant funded by the Italian Ministry of Education, University and Research in cooperation with the European Funds for the Regional Development (FESR).

References

1. Hashmi, U., Shafqat, S., Khan, F., Majid, M., Hussain, H., Kazi, A.G., John, R., Ahmad, P.: Plant exomics: concepts, applications and methodologies in crop improvement. Plant Signal. Behav. **10**(1), e976152 (2015). doi:10.4161/15592324.2014.976152
2. Warr, A., Robert, C., Hume, D., Archibald, A., Deeb, N., Watson, M.: Exome sequencing: current and future perspectives. G3 **5**(8), 1543–1550 (2015). doi:10.1534/g3.115.018564
3. Cronn, R., Knaus, B.J., Liston, A., Maughan, P.J., Parks, M., Syring, J.V., Udall, J.: Targeted enrichment strategies for next-generation plant biology. Am. J. Bot. **99**(2), 291–311 (2012). doi:10.3732/ajb.1100356
4. Hodges, E., Xuan, Z., Balija, V., Kramer, M., Molla, M.N., Smith, S.W., Middle, C.M., Rodesch, M.J., Albert, T.J., Hannon, G.J., McCombie, W.R.: Genome-wide in situ exon capture for selective resequencing. Nat. Genet. **39**(12), 1522–1527 (2007). doi:10.1038/ng.2007.42
5. Mamanova, L., Coffey, A.J., Scott, C.E., Kozarewa, I., Turner, E.H., Kumar, A., Howard, E., Shendure, J., Turner, D.J.: Target-enrichment strategies for next-generation sequencing. Nat. Methods **7**(2), 111–118 (2010). doi:10.1038/nmeth.1419
6. Mertes, F., Elsharawy, A., Sauer, S., van Helvoort, J.M., van der Zaag, P.J., Franke, A., Nilsson, M., Lehrach, H., Brookes, A.J.: Targeted enrichment of genomic DNA regions for next-generation sequencing. Brief. Funct. Genomics **10**(6), 374–386 (2011). doi:10.1093/bfgp/elr033
7. Gnirke, A., Melnikov, A., Maguire, J., Rogov, P., LeProust, E.M., Brockman, W., Fennell, T., Giannoukos, G., Fisher, S., Russ, C., Gabriel, S., Jaffe, D.B., Lander, E.S., Nusbaum, C.: Solution hybrid selection with ultra-long oligonucleotides for massively parallel targeted sequencing. Nat. Biotechnol. **27**(2), 182–189 (2009). doi:10.1038/nbt.1523
8. Okou, D.T., Steinberg, K.M., Middle, C., Cutler, D.J., Albert, T.J., Zwick, M.E.: Microarray-based genomic selection for high-throughput resequencing. Nat. Methods **4**(11), 907–909 (2007). doi:10.1038/nmeth1109
9. Fu, Y., Springer, N.M., Gerhardt, D.J., Ying, K., Yeh, C.T., Wu, W., Swanson-Wagner, R., D'Ascenzo, M., Millard, T., Freeberg, L., Aoyama, N., Kitzman, J., Burgess, D., Richmond, T., Albert, T.J., Barbazuk, W.B., Jeddeloh, J.A., Schnable, P.S.: Repeat subtraction-mediated sequence capture from a complex genome. Plant J. **62**(5), 898–909 (2010). doi:10.1111/j.1365-313X.2010.04196.x
10. Saintenac, C., Jiang, D., Akhunov, E.D.: Targeted analysis of nucleotide and copy number variation by exon capture in allotetraploid wheat genome. Genome Biol. **12**(9), R88 (2011). doi:10.1186/gb-2011-12-9-r88
11. Martin, L.B., Fei, Z., Giovannoni, J.J., Rose, J.K.: Catalyzing plant science research with RNA-seq. Front. Plant Sci. **4**, 66 (2013). doi:10.3389/fpls.2013.00066
12. Egan, A.N., Schlueter, J., Spooner, D.M.: Applications of next-generation sequencing in plant biology. Am. J. Bot. **99**(2), 175–185 (2012). doi:10.3732/ajb.1200020
13. Baird, N.A., Etter, P.D., Atwood, T.S., Currey, M.C., Shiver, A.L., Lewis, Z.A., Selker, E.U., Cresko, W.A., Johnson, E.A.: Rapid SNP discovery and genetic mapping using sequenced RAD markers. PLoS One **3**(10), e3376 (2008). doi:10.1371/journal.pone.0003376
14. Rowe, H.C., Renaut, S., Guggisberg, A.: RAD in the realm of next-generation sequencing technologies. Mol. Ecol. **20**(17), 3499–3502 (2011)
15. Maughan, P.J., Yourstone, S.M., Jellen, E.N., Udall, J.A.: SNP discovery via genomic reduction, barcoding, and 454-pyrosequencing in Amaranth. Plant Genome J. **2**(3), 260 (2009). doi:10.3835/plantgenome2009.08.0022
16. He, J., Zhao, X., Laroche, A., Lu, Z.X., Liu, H., Li, Z.: Genotyping-by-sequencing (GBS), an ultimate marker-assisted selection (MAS) tool to accelerate plant breeding. Front. Plant Sci. **5**, 484 (2014). doi:10.3389/fpls.2014.00484

17. Uribe-Convers, S., Settles, M.L., Tank, D.C.: A phylogenomic approach based on PCR target enrichment and high throughput sequencing: resolving the diversity within the South American species of Bartsia L. (Orobanchaceae). PLoS one **11**(2), e0148203 (2016). doi:10.1371/journal.pone.0148203
18. Durstewitz, G., Polley, A., Plieske, J., Luerssen, H., Graner, E.M., Wieseke, R., Ganal, M.W.: SNP discovery by amplicon sequencing and multiplex SNP genotyping in the allopolyploid species Brassica napus. Genome **3**(11), 948–956 (2010). doi:10.1139/G10-079
19. Chilamakuri, C.S., Lorenz, S., Madoui, M.A., Vodak, D., Sun, J., Hovig, E., Myklebost, O., Meza-Zepeda, L.A.: Performance comparison of four exome capture systems for deep sequencing. BMC Genomics **15**, 449 (2014). doi:10.1186/1471-2164-15-449
20. Parla, J.S., Iossifov, I., Grabill, I., Spector, M.S., Kramer, M., McCombie, W.R.: A comparative analysis of exome capture. Genome Biol. **12**(9), R97 (2011). doi:10.1186/gb-2011-12-9-r97
21. Dohm, J.C., Lottaz, C., Borodina, T., Himmelbauer, H.: Substantial biases in ultra-short read data sets from high-throughput DNA sequencing. Nucleic Acids Res. **36**(16), e105 (2008). doi:10.1093/nar/gkn425
22. Aird, D., Ross, M.G., Chen, W.S., Danielsson, M., Fennell, T., Russ, C., Jaffe, D.B., Nusbaum, C., Gnirke, A.: Analyzing and minimizing PCR amplification bias in Illumina sequencing libraries. Genome Biol. **12**(2), R18 (2011). doi:10.1186/gb-2011-12-2-r18
23. Zhou, L., Holliday, J.A.: Targeted enrichment of the black cottonwood (Populus trichocarpa) gene space using sequence capture. BMC Genomics **13**, 703 (2012). doi:10.1186/1471-2164-13-703
24. Muraya, M.M., Schmutzer, T., Ulpinnis, C., Scholz, U., Altmann, T.: Targeted sequencing reveals large-scale sequence polymorphism in Maize candidate genes for biomass production and composition. PLoS One **10**(7), e0132120 (2015). doi:10.1371/journal.pone.0132120
25. Clarke, W.E., Parkin, I.A., Gajardo, H.A., Gerhardt, D.J., Higgins, E., Sidebottom, C., Sharpe, A.G., Snowdon, R.J., Federico, M.L., Iniguez-Luy, F.L.: Genomic DNA enrichment using sequence capture microarrays: a novel approach to discover sequence nucleotide polymorphisms (SNP) in Brassica napus L. PLoS One **8**(12), e81992 (2013). doi:10.1371/journal.pone.0081992
26. Schiessl, S., Samans, B., Huttel, B., Reinhard, R., Snowdon, R.J.: Capturing sequence variation among flowering-time regulatory gene homologs in the allopolyploid crop species Brassica napus. Front. Plant Sci. **5**, 404 (2014). doi:10.3389/fpls.2014.00404
27. Salmon, A., Udall, J.A., Jeddeloh, J.A., Wendel, J.: Targeted capture of homoeologous coding and noncoding sequence in polyploid cotton. G3 **2**(8), 921–930 (2012). doi:10.1534/g3.112.003392
28. Bundock, P.C., Casu, R.E., Henry, R.J.: Enrichment of genomic DNA for polymorphism detection in a non-model highly polyploid crop plant. Plant Biotechnol. J. **10**(6), 657–667 (2012). doi:10.1111/j.1467-7652.2012.00707.x
29. Evans, J., Kim, J., Childs, K.L., Vaillancourt, B., Crisovan, E., Nandety, A., Gerhardt, D.J., Richmond, T.A., Jeddeloh, J.A., Kaeppler, S.M., Casler, M.D., Buell, C.R.: Nucleotide polymorphism and copy number variant detection using exome capture and next-generation sequencing in the polyploid grass Panicum virgatum. Plant J. **79**(6), 993–1008 (2014). doi:10.1111/tpj.12601
30. Pootakham, W., Shearman, J.R., Ruang-Areerate, P., Sonthirod, C., Sangsrakru, D., Jomchai, N., Yoocha, T., Triwitayakorn, K., Tragoonrung, S., Tangphatsornruang, S.: Large-scale SNP discovery through RNA sequencing and SNP genotyping by targeted enrichment sequencing in cassava (Manihot esculenta Crantz). PLoS One **9**(12), e116028 (2014). doi:10.1371/journal.pone.0116028
31. Neves, L.G., Davis, J.M., Barbazuk, W.B., Kirst, M.: Whole-exome targeted sequencing of the uncharacterized pine genome. Plant J. **75**(1), 146–156 (2013). doi:10.1111/tpj.12193
32. Neves, L.G., Davis, J.M., Barbazuk, W.B., Kirst, M.: A high-density gene map of loblolly pine (Pinus taeda L.) based on exome sequence capture genotyping. G3 **4**(1), 29–37 (2014). doi:10.1534/g3.113.008714

33. Dasgupta, M.G., Dharanishanthi, V., Agarwal, I., Krutovsky, K.V.: Development of genetic markers in Eucalyptus species by target enrichment and exome sequencing. PLoS One **10**(1), e0116528 (2015). doi:10.1371/journal.pone.0116528
34. Allen, A.M., Barker, G.L., Wilkinson, P., Burridge, A., Winfield, M., Coghill, J., Uauy, C., Griffiths, S., Jack, P., Berry, S., Werner, P., Melichar, J.P., McDougall, J., Gwilliam, R., Robinson, P., Edwards, K.J.: Discovery and development of exome-based, co-dominant single nucleotide polymorphism markers in hexaploid wheat (Triticum aestivum L.). Plant Biotechnol. J. **11**(3), 279–295 (2013). doi:10.1111/pbi.12009
35. Winfield, M.O., Wilkinson, P.A., Allen, A.M., Barker, G.L., Coghill, J.A., Burridge, A., Hall, A., Brenchley, R.C., D'Amore, R., Hall, N., Bevan, M.W., Richmond, T., Gerhardt, D.J., Jeddeloh, J.A., Edwards, K.J.: Targeted re-sequencing of the allohexaploid wheat exome. Plant Biotechnol. J. **10**(6), 733–742 (2012). doi:10.1111/j.1467-7652.2012.00713.x
36. Gardiner, L.J., Gawronski, P., Olohan, L., Schnurbusch, T., Hall, N., Hall, A.: Using genic sequence capture in combination with a syntenic pseudo genome to map a deletion mutant in a wheat species. Plant J. **80**(5), 895–904 (2014). doi:10.1111/tpj.12660
37. Henry, I.M., Nagalakshmi, U., Lieberman, M.C., Ngo, K.J., Krasileva, K.V., Vasquez-Gross, H., Akhunova, A., Akhunov, E., Dubcovsky, J., Tai, T.H., Comai, L.: Efficient genome-wide detection and cataloging of EMS-induced mutations using exome capture and next-generation sequencing. Plant Cell **26**(4), 1382–1397 (2014). doi:10.1105/tpc.113.121590
38. Bolon, Y.T., Haun, W.J., Xu, W.W., Grant, D., Stacey, M.G., Nelson, R.T., Gerhardt, D.J., Jeddeloh, J.A., Stacey, G., Muehlbauer, G.J., Orf, J.H., Naeve, S.L., Stupar, R.M., Vance, C.P.: Phenotypic and genomic analyses of a fast neutron mutant population resource in soybean. Plant Physiol. **156**(1), 240–253 (2011). doi:10.1104/pp.110.170811
39. Haun, W.J., Hyten, D.L., Xu, W.W., Gerhardt, D.J., Albert, T.J., Richmond, T., Jeddeloh, J.A., Jia, G., Springer, N.M., Vance, C.P., Stupar, R.M.: The composition and origins of genomic variation among individuals of the soybean reference cultivar Williams 82. Plant Physiol. **155**(2), 645–655 (2011). doi:10.1104/pp.110.166736
40. Mascher, M., Richmond, T.A., Gerhardt, D.J., Himmelbach, A., Clissold, L., Sampath, D., Ayling, S., Steuernagel, B., Pfeifer, M., D'Ascenzo, M., Akhunov, E.D., Hedley, P.E., Gonzales, A.M., Morrell, P.L., Kilian, B., Blattner, F.R., Scholz, U., Mayer, K.F., Flavell, A.J., Muehlbauer, G.J., Waugh, R., Jeddeloh, J.A., Stein, N.: Barley whole exome capture: a tool for genomic research in the genus Hordeum and beyond. Plant J. **76**(3), 494–505 (2013). doi:10.1111/tpj.12294
41. Pankin, A., Campoli, C., Dong, X., Kilian, B., Sharma, R., Himmelbach, A., Saini, R., Davis, S.J., Stein, N., Schneeberger, K., von Korff, M.: Mapping-by-sequencing identifies HvPHYTOCHROME C as a candidate gene for the early maturity 5 locus modulating the circadian clock and photoperiodic flowering in barley. Genetics **198**(1), 383–396 (2014). doi:10.1534/genetics.114.165613
42. Wendler, N., Mascher, M., Noh, C., Himmelbach, A., Scholz, U., Ruge-Wehling, B., Stein, N.: Unlocking the secondary gene-pool of barley with next-generation sequencing. Plant Biotechnol. J. **12**(8), 1122–1131 (2014). doi:10.1111/pbi.12219
43. de Sousa, F., Bertrand, Y.J., Nylinder, S., Oxelman, B., Eriksson, J.S., Pfeil, B.E.: Phylogenetic properties of 50 nuclear loci in Medicago (Leguminosae) generated using multiplexed sequence capture and next-generation sequencing. PLoS One **9**(10), e109704 (2014). doi:10.1371/journal.pone.0109704
44. Mandel, J.R., Dikow, R.B., Funk, V.A., Masalia, R.R., Staton, S.E., Kozik, A., Michelmore, R.W., Rieseberg, L.H., Burke, J.M.: A target enrichment method for gathering phylogenetic information from hundreds of loci: an example from the Compositae. Appl. Plant Sci. **2**(2) (2014). doi:10.3732/apps.1300085
45. Stull, G.W., Moore, M.J., Mandala, V.S., Douglas, N.A., Kates, H.R., Qi, X., Brockington, S.F., Soltis, P.S., Soltis, D.E., Gitzendanner, M.A.: A targeted enrichment strategy for massively parallel sequencing of angiosperm plastid genomes. Appl. Plant Sci. **1**(2) (2013). doi:10.3732/apps.1200497

46. Tennessen, J.A., Govindarajulu, R., Liston, A., Ashman, T.L.: Targeted sequence capture provides insight into genome structure and genetics of male sterility in a gynodioecious diploid strawberry, Fragaria vesca ssp. bracteata (Rosaceae). G3 **3**(8), 1341–1351 (2013). doi:10.1534/g3.113.006288
47. Jupe, F., Witek, K., Verweij, W., Sliwka, J., Pritchard, L., Etherington, G.J., Maclean, D., Cock, P.J., Leggett, R.M., Bryan, G.J., Cardle, L., Hein, I., Jones, J.D.: Resistance gene enrichment sequencing (RenSeq) enables reannotation of the NB-LRR gene family from sequenced plant genomes and rapid mapping of resistance loci in segregating populations. Plant J. **76**(3), 530–544 (2013). doi:10.1111/tpj.12307
48. Andolfo, G., Jupe, F., Witek, K., Etherington, G.J., Ercolano, M.R., Jones, J.D.: Defining the full tomato NB-LRR resistance gene repertoire using genomic and cDNA RenSeq. BMC Plant Biol. **14**, 120 (2014). doi:10.1186/1471-2229-14-120
49. Uitdewilligen, J.G., Wolters, A.M., D'Hoop B.B., Borm, T.J., Visser, R.G., van Eck, H.J.: A next-generation sequencing method for genotyping-by-sequencing of highly heterozygous autotetraploid potato. PLoS One **8**(5), e62355 (2013). doi:10.1371/journal.pone.0062355
50. Li, J.W., Robison, K., Martin, M., Sjodin, A., Usadel, B., Young, M., Olivares, E.C., Bolser, D.M.: The SEQanswers wiki: a wiki database of tools for high-throughput sequencing analysis. Nucleic Acids Res. **40**(Database issue), D1313–D1317 (2012). doi:10.1093/nar/gkr1058
51. Bolger, A.M., Lohse, M., Usadel, B.: Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics **30**(15), 2114–2120 (2014). doi:10.1093/bioinformatics/btu170
52. Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., Durbin, R.: Genome project data processing, S.: the sequence alignment/map format and SAMtools. Bioinformatics **25**(16), 2078–2079 (2009). doi:10.1093/bioinformatics/btp352
53. Sims, D., Sudbery, I., Ilott, N.E., Heger, A., Ponting, C.P.: Sequencing depth and coverage: key considerations in genomic analyses. Nat. Rev. Genet. **15**(2), 121–132 (2014). doi:10.1038/nrg3642
54. Hatem, A., Bozdag, D., Toland, A.E., Catalyurek, U.V.: Benchmarking short sequence mapping tools. BMC Bioinf. **14**, 184 (2013). doi:10.1186/1471-2105-14-184
55. Li, H., Durbin, R.: Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics **25**(14), 1754–1760 (2009). doi:10.1093/bioinformatics/btp324
56. Langmead, B., Salzberg, S.L.: Fast gapped-read alignment with Bowtie 2. Nat. Methods **9**(4), 357–359 (2012). doi:10.1038/nmeth.1923
57. Li, R., Yu, C., Li, Y., Lam, T.W., Yiu, S.M., Kristiansen, K., Wang, J.: SOAP2: an improved ultrafast tool for short read alignment. Bioinformatics **25**(15), 1966–1967 (2009). doi:10.1093/bioinformatics/btp336
58. Quinlan, A.R., Hall, I.M.: BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics **26**(6), 841–842 (2010). doi:10.1093/bioinformatics/btq033
59. Homer, N., Nelson, S.F.: Improved variant discovery through local re-alignment of short-read next-generation sequencing data using SRMA. Genome Biol. **11**(10), R99 (2010). doi:10.1186/gb-2010-11-10-r99
60. McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., Garimella, K., Altshuler, D., Gabriel, S., Daly, M., DePristo, M.A.: The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res. **20**(9), 1297–1303 (2010). doi:10.1101/gr.107524.110
61. Van der Auwera, G.A., Carneiro, M.O., Hartl, C., Poplin, R., Del Angel, G., Levy-Moonshine, A., Jordan, T., Shakir, K., Roazen, D., Thibault, J., Banks, E., Garimella, K.V., Altshuler, D., Gabriel, S., DePristo, M.A.: From FastQ data to high confidence variant calls: the Genome Analysis Toolkit best practices pipeline. Curr. Protoc. Bioinformatics **11**(1110), 11.10.11–11.10.33 (2013). doi:10.1002/0471250953.bi1110s43
62. Danecek, P., Auton, A., Abecasis, G., Albers, C.A., Banks, E., DePristo, M.A., Handsaker, R.E., Lunter, G., Marth, G.T., Sherry, S.T., McVean, G., Durbin, R., Genomes Project Analysis, G.: The variant call format and VCFtools. Bioinformatics **27**(15), 2156–2158 (2011). doi:10.1093/bioinformatics/btr330

63. Li, H.: A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. Bioinformatics **27**(21), 2987–2993 (2011). doi:10.1093/bioinformatics/btr509
64. Yang, H., Wang, K.: Genomic variant annotation and prioritization with ANNOVAR and wANNOVAR. Nat. Protoc. **10**(10), 1556–1566 (2015). doi:10.1038/nprot.2015.105
65. Cingolani, P., Platts, A., le Wang, L., Coon, M., Nguyen, T., Wang, L., Land, S.J., Lu, X., Ruden, D.M.: 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 **6**(2), 80–92 (2012). doi:10.4161/fly.19695
66. Adzhubei, I.A., Schmidt, S., Peshkin, L., Ramensky, V.E., Gerasimova, A., Bork, P., Kondrashov, A.S., Sunyaev, S.R.: A method and server for predicting damaging missense mutations. Nat. Methods **7**(4), 248–249 (2010). doi:10.1038/nmeth0410-248
67. Ng, P.C., Henikoff, S.: Predicting the effects of amino acid substitutions on protein function. Annu. Rev. Genomics Hum. Genet. **7**, 61–80 (2006). doi:10.1146/annurev.genom.7.080505.115630
68. Bendl, J., Stourac, J., Salanda, O., Pavelka, A., Wieben, E.D., Zendulka, J., Brezovsky, J., Damborsky, J.: PredictSNP: robust and accurate consensus classifier for prediction of disease-related mutations. PLoS Comput. Biol. **10**(1), e1003440 (2014). doi:10.1371/journal.pcbi.1003440
69. Cooper, D.N.: Functional intronic polymorphisms: Buried treasure awaiting discovery within our genes. Hum. Genomics **4**(5), 284–288 (2010)
70. Chow, C.N., Zheng, H.Q., Wu, N.Y., Chien, C.H., Huang, H.D., Lee, T.Y., Chiang-Hsieh, Y.F., Hou, P.F., Yang, T.Y., Chang, W.C.: PlantPAN 2.0: an update of plant promoter analysis navigator for reconstructing transcriptional regulatory networks in plants. Nucleic Acids Res. **44**(D1), D1154–D1160 (2016). doi:10.1093/nar/gkv1035
71. Thorvaldsdottir, H., Robinson, J.T., Mesirov, J.P.: Integrative genomics viewer (IGV): high-performance genomics data visualization and exploration. Brief. Bioinform. **14**(2), 178–192 (2013). doi:10.1093/bib/bbs017
72. Kumar, G.R., Sakthivel, K., Sundaram, R.M., Neeraja, C.N., Balachandran, S.M., Rani, N.S., Viraktamath, B.C., Madhav, M.S.: Allele mining in crops: prospects and potentials. Biotechnol. Adv. **28**(4), 451–461 (2010). doi:10.1016/j.biotechadv.2010.02.007
73. Rose, L.E., Langley, C.H., Bernal, A.J., Michelmore, R.W.: Natural variation in the Pto pathogen resistance gene within species of wild tomato (Lycopersicon)I. Functional analysis of Pto alleles. Genetics **171**(1), 345–357 (2005). doi:10.1534/genetics.104.039339
74. Raiola, A., Rigano, M.M., Calafiore, R., Frusciante, L., Barone, A.: Enhancing the health-promoting effects of tomato fruit for biofortified food. Mediators Inflamm. **2014**, 139873 (2014). doi:10.1155/2014/139873
75. Giuliano, G.: Plant carotenoids: genomics meets multi-gene engineering. Curr. Opin. Plant Biol. **19**, 111–117 (2014). doi:10.1016/j.pbi.2014.05.006
76. Kavitha, P., Shivashankara, K.S., Rao, V.K., Sadashiva, A.T., Ravishankar, K.V., Sathish, G.J.: Genotypic variability for antioxidant and quality parameters among tomato cultivars, hybrids, cherry tomatoes and wild species. J. Sci. Food Agric. **94**(5), 993–999 (2014). doi:10.1002/jsfa.6359
77. Ruggieri, V., Francese, G., Sacco, A., D’Alessandro, A., Rigano, M.M., Parisi, M., Milone, M., Cardi, T., Mennella, G., Barone, A.: An association mapping approach to identify favourable alleles for tomato fruit quality breeding. BMC Plant Biol. **14**, 337 (2014). doi:10.1186/s12870-014-0337-9
78. Liu, L., Shao, Z., Zhang, M., Wang, Q.: Regulation of carotenoid metabolism in tomato. Mol. Plant **8**(1), 28–39 (2015). doi:10.1016/j.molp.2014.11.006
79. Tomato Genome Consortium: The tomato genome sequence provides insights into fleshy fruit evolution. Nature **485**(7400), 635–641 (2012). doi:10.1038/nature11119

DecontaMiner: A Pipeline for the Detection and Analysis of Contaminating Sequences in Human NGS Sequencing Data

Ilaria Granata*, Mara Sangiovanni*, and Mario Guarracino

Abstract Reads alignment is an essential step of next generation sequencing) data analyses. One challenging issue is represented by unmapped reads that are usually discarded and considered as not informative. Instead, it is important to fully understand the source of those reads, to assess the quality of the whole experiment. Moreover, it is of interest to get some insights on possible "contamination" from non-human sequences (e.g., viruses, bacteria, and fungi). Contamination may take place during the experimental procedures leading to sequencing, or be due to the presence of microorganisms infecting the sampled tissues. Here we propose a pipeline for the detection of viral, bacterial, and fungi contamination in human sequenced data. Similarities between input reads (query) and putative contaminating organism sequences (subject) are detected using a local alignment strategy (MegaBLAST). For each organism database DecontaMiner provides two main output files: one containing all the reads matching only a single organism; the second one containing the "ambiguous" matching reads. In both files, data is sorted by organism and classified by taxonomic group. Low quality, unaligned sequences, and those discarded by user criteria are also provided as output. Other information and summary statistics on the number of matched/filtered/discarded reads and organisms are generated. This pipeline has successfully detected foreign sequences in human Cancer RNA-seq data.

Keywords Contaminating sequences • Unmapped reads • NGS data

I. Granata • M. Sangiovanni (✉)
ICAR-CNR, Via Pietro Castellino 111, 80131 Napoli, Italy
e-mail: ilaria.granata@icar.cnr.it; mara.sangiovanni@icar.cnr.it

M. Guarracino
Laboratory for Genomics, Transcriptomics and Proteomics (Lab-GTP), High Performance Computing and Networking Institute (ICAR), National Research Council (CNR), Via Pietro Castellino 111, Naples, Italy
e-mail: mario.guarracino@cnr.it

A. Rogato et al. (eds.), *Dynamics of Mathematical Models in Biology*,
DOI 10.1007/978-3-319-45723-9_11

1 Introduction

The study of the human genome and its relationship with the environment is a crucial task in the context of modern biology.

The application of next generation sequencing technologies allows to characterize the genome-wide map of organisms. Genome investigation has been made possible by the construction of the reference genomes. Sequencing experiments produce a large amount of small sequences that have to be mapped to the reference. The alignment is probably the most challenging step of next generation sequencing (NGS) data analyses. It allows to obtain several information—such as read density, gene lists, and variant lists—crucial to the definition of the biological meaning underlying the data.

Typically the amount of reads that correctly map onto the human reference genome ranges between 70 and 90 % [1] leaving in some cases a consistent fraction of unmapped reads. Underestimating this portion may determine loss of precious information. Unmapped reads can be explained by errors during sequencing protocols, by the presence of repeat elements difficult to map, by novel transcripts that can be investigated by de novo assembly, and lastly, they can derive from non-human sequences. Indeed, microorganisms contamination can occur during samples processing or can be part of the normal or pathological tissues microbiome [2].

The interest in detecting microorganisms-derived sequences has grown up together with the spread of high-throughput approaches, allowing the extraction of information both about the quality of the experimental procedures and about the link between diseases and infections. The main appeal of these investigations is represented by the possibility to find new pathogen-disease associations. In literature there are many evidences which underline the importance of detecting contaminating organisms. Worth to note are the detection of polyomavirus in human Merkel cell carcinoma [3] and a novel Old World arenavirus in a cluster of patients with fatal transplant-associated disease [4]. Assembly of a novel bacterial draft genome starting from tissue specimens sequencing of cord colitis patients suggested an opportunistic pathogenic role for *Bradyrhizobium enterica* in humans [5].

Besides, environmental contaminations are routinely found in NGS datasets. Downstream contaminations or cross-contaminations can compromise the reliability of the whole experimental procedure. Strong et al. detected bacterial sequences, belonging to different taxa, in cell line data coming from different sequencing experiments and suggested the idea that a good portion of these bacterial reads did not derive from the specimens themselves but from downstream contamination. This suggestion has been supported by the detection of bacterial sequences in polyA RNA-seq [6]. Indeed, the polyA selection step should remove upstream contamination since bacteria are poorly polyadenylated. Moreover, to strengthen the hypothesis of downstream contamination occurrence, the authors analyzed

[*] Authors are contributed equally.

RNA-sequencing data of five Epstein-Barr virus (EBV)-positive lymphoblastoid cell lines obtained in six different Illumina laboratories. Across these labs the level of bacterial reads per million human mapped reads (RPMHs) differed by as much as 30-fold, while the transcript levels of the EB virus were similar.

Furthermore, another study also confirmed this laboratory-peculiar contamination, showing that different sequencing centers had specific signatures of contaminating genomes as "time stamps" [7]. Unmapped ChIP-Seq reads from *A. thaliana*, *Z. mays*, *H. sapiens*, and *D. melanogaster* datasets were investigated and found contaminated by foreign sequences. Taxonomic classification of these reads allowed authors to define the contaminants and to calculate the relative abundance for each dataset [8].

Several tools, based on different computational approaches, have been developed and used for the detection of pathogens in high-throughput sequencing data, especially in cancer samples. In particular, PathSeq [9] and CaPSID [10] are worth mentioning. Both are available as integrated open source softwares.

PathSeq applies a subtraction approach in which the reads are aligned on six different human genomes. After, it uses local aligners such as Mega BLAST and BLASTN [11] to re-align reads to microbial reference sequences and to two additional human sequence databases. PathSeq is implemented in a cloud-computing environment. However, the PathSeq pipeline can be computationally intensive, mostly due to the numerous subtraction steps. CaPSID overcame this limit using a single human reference genome with splice junctions. Although CaPSID might face the risk to fail the correct alignment, it provides a large reduction in elaboration time. Furthermore, PathSeq discards the ambiguous reads that map both to human and pathogen genomes, while CaPSID stores them in a database.

It should be noted that PathSeq also requires a commercial computing platform (i.e., Amazon Elastic Compute Cloud, EC2) to be used. CaPSID does not have this kind of restriction but it requires two files in *bam* format as input, obtained by the user with a separate alignment software. The user should take care of aligning the sequences both to human and to each pathogen (bacteria, viruses, and fungi) reference genome of interest, thus performing the most computationally intensive steps before CaPSID. Hence, the CaPSID pipeline is lighter and faster, and it can provide even gene annotations and a user-friendly web application that integrates a genome browser.

Another cloud-compatible bioinformatics pipeline aimed to pathogen discovery is SURPI ("Sequence-based Ultrarapid Pathogen Identification") [12], which provides a very useful and complete tool for the analysis of complex metagenomic NGS data. However, its purpose is the detection of microorganisms from complex clinical metagenomic samples open to the environment, using the entire NCBI nt and/or NCBI nr protein databases in comprehensive mode. The algorithm is particularly sensitive but, as consequence, the pipeline is likely not appropriate for a rapid analysis of the unmapped reads.

As far as we know, all the pipelines mentioned above are designed to analyze data primarily aimed to the detection of pathogens in human samples. Due to this, some of them, such as PathSeq and SURPI, provide intensive pipeline including alignment to host genome, while CaPSID, in order to reduce the required time and computational efforts, works on BAM files provided by the user, containing the resulted alignments to the human and to all the pathogen reference sequences.

Here we propose DecontaMiner, a pipeline designed and developed to detect contaminating sequences in NGS data. Our main purpose is to understand the nature of those reads that fail to map to the reference genome, as well as to provide an automatic pipeline that allows the quality filtering and the processing of these sequences.

From the detected output it is straightforward to extract information about the eventual samples contamination and/or tissue infection. As in the above-mentioned papers [6–8] the experimental setup and the study of the detected microorganism species might suggest the possible contamination sources. In general, it is not possible to automatically discriminate between upstream and downstream contamination.

Concluding, it can be said that DecontaMiner lies in the middle between the complex, intensive pipelines of PathSeq and SURPI, and the post-alignment approach of CaPSID.

2 DecontaMiner Pipeline

The DecontaMiner pipeline is a suite composed of several command-line tools wrapped together to identify, through digital subtraction, non-human nucleotide sequences generated by high-throughput sequencing of RNA or DNA samples. It is mainly written using Bash scripting and the Perl language. It requires in input the BAM files or the raw fastQ files containing the unmapped reads (i.e., all the reads discarded during the alignment on the human reference genome) if any. A schematic view of the pipeline is shown in Fig. 1.

All the files that have to be submitted to DecontaMiner can be collected in the same directory, and its path given as input. The entire pipeline can be subdivided into three main phases.

The first phase involves the filtering and file format conversion steps, needed to remove low quality reads and to obtain reads in fasta-format files, ready to be aligned to the genome databases. More in detail, DecontaMiner wraps in its pipeline two of the most used toolkits, Samtools [13] and Bedtools [14] used for the format conversions, and FastX [15] for the quality filtering. The filtering is mainly based on two parameters set by the user, namely the Phred quality threshold and the minimum percentage of bases within that threshold.

DecontaMiner works both on paired- and single-end experiments, a parameter that must be specified by the user. The conversion steps allow to sort the reads and switch from bam to fastq and then to fasta formats.

Once terminated the conversion phase, the mapping module can start. In the case of RNA-seq data, it is crucial to remove the ribosomal RNA (rRNA). Indeed, rRNA represents up to 90 % of the total RNA. Although the wet lab procedures provide an rRNA removal step, often this procedure is not totally satisfactory, due to high

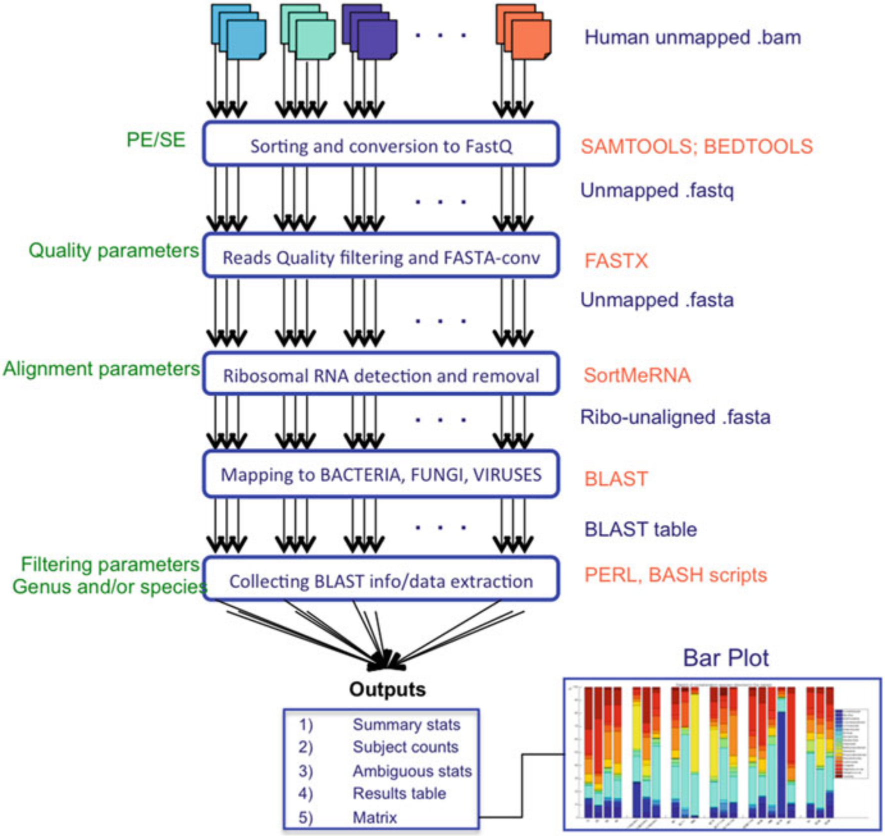


Fig. 1 The pipeline
A scheme of the DecontaMiner pipeline. On the right, in blue are the input files, and in red the tools used to process the data. In the central part, as a flux, the processing steps are described. On the left, the parameters that can be set for each step are indicated in green. Several tab-delimited files and one matrix are the pipeline outputs. All the discarded reads are also provided, as well as all the different file formats generated (fastQ, FASTA, etc.). The matrix, containing all the samples, can be easily used to create a bar plot

number of rRNA copies. We downloaded the fasta sequences of human ribosomal RNA (28S, 18S, 5S, 5.8S and mitochondrial 12S, 16S) from NCBI website. The rRNA alignment is performed using the SortmeRNA tool [16], which is a software designed to this aim. All the reads that do not map to the human rRNA will undergo mapping to bacteria, viruses, and fungi genome databases (NCBI nt) using the MegaBLAST [17] algorithm.

The rRNA alignment reliability is evaluated using the *E*-value score. This threshold can be either set by the user or left at the default SortMeRna value. The user can specify also the alignment length and number of allowed mismatches/gaps when aligning to contaminating genomes.

The BLAST outputs, in table format, are then submitted to the third and last phase, that involves the collection and extraction of information from the local alignments.

This module, mainly composed of Perl scripts, is executed accordingly to some user-specified parameters specifying the filtering and collecting options. In particular, the filtering is based on the threshold number of total reads successfully mapped and on the minimum threshold of reads mapped to a single organism. Instead, the collecting options involve the choice of organizing the results according either to genus or to species names.

DecontaMiner stores the output reads into three main files: unaligned, ambiguous, and aligned. The "unaligned" file contains the reads that do not satisfy the filtering parameters (i.e., length of alignment, number of allowed gaps, and mismatches). The ambiguous reads are those that map to different Genera or, in case of paired-end reads, those having mates mapping to different genera. Ambiguous reads mapping to more than one Genus might derive from ortholog sequences. Since Reads matching all the filtering criteria are stored into the "aligned" file.

The results are available in a tabular format, one for each sample, containing the names of the detected organisms and the relative reads count. Furthermore, DecontaMiner generates a matrix that can be easily used to create a barplot or other types of diagrams in which all the data are collected together.

Lastly, the summary statistics about the number of matched/filtered/discarded reads and organisms are generated and stored into tabular textual files.

3 Case Studies

3.1 Cancer Datasets

In order to assess the usefulness of the DecontaMiner pipeline and its efficiency in detecting non-human sequences in NGS data, we used two publicly available datasets downloaded from the GEO portal (GSE68086 and GSE69240).

The first study, from which the dataset GSE68086 was generated, concerns the total RNA-sequencing experiments of blood platelet samples from patients with six different malignant tumors (non-small cell lung cancer, colorectal cancer, pancreatic cancer, glioblastoma, breast cancer, and hepato-biliary carcinomas) and from healthy donors [18]. The experiment was performed with single-end 100 bp reads.

The second one, GSE69240, derives from the expression profiling by high-throughput sequencing of High-Grade Ductal Carcinoma In Situ (DCIS) [19]. The dataset contains 25 pure HG-DCIS and 10 normal breast organoids samples. The reads are paired-end 76 nucleotides long. This second dataset was used for testing our pipeline on polyA RNA-seq data.

Table 1 Decontaminer parameter settings

Parameter name	Value
Phred quality threshold	20
Minimum % of bases with the Phred set quality	100
E-value rRNA alignment	$\leq$10–20
Match length	= Query length
Mismatch number	1
Gap number	0
Minimum threshold of reads mapped to a single organism	100

3.2 Pre-processing

The Sequence Read Archive (SRA) file of each sample was downloaded and converted to fastq format using the SRAToolkit [20]. The sequencing reads were cleaned by eventual poor quality ends by Trimmomatic [21]. The quality assessment of the trimmed reads was performed with FastQC [22]. The fast splice junction mapper TopHat [23] was chosen to align the fastq files to the reference genome (assembly hg19) guided by UCSC gene annotation. The sequence features in mapped data were checked by SamStat [24]. The unmapped bam files provided by TopHat were the input to our pipeline.

The parameter setting used for analyzing the two datasets is listed in Table 1.

3.3 Results

The analysis of the overall read mapping rate showed a high variability among the samples of the GSE68086 dataset, with a range of 5–40 % of unmapped reads.

In the case of the GSE69240 dataset, instead, we observed a good mapping rate in all the samples, with a percentage of unmapped reads below 10 %. The mapping statistics of the two datasets immediately suggested a different probability to detect non-human sequences.

In order to test the reliability of our pipeline we submitted to the analysis also the samples with a small amount of unmapped sequences.

As we expected, we did not find any significant match to contaminating genomes for the samples of the GSE69240 dataset. We also re-analyzed the data, lowering the stringency of the parameters in terms of allowed mismatches and gaps (2 for each), with the same negative outcome.

This result completely agrees with the type of experimental procedure used. As mentioned before, an efficient polyA RNA-seq process and a set of samples not contaminated by the environment should guarantee reads free of contamination. Hence, this result supports the reliability of the pipeline in terms of false positives detection.

Table 2 Number of reads in the Decontaminer pipeline for two tumor sample

	Number of obtained reads (% of raw reads)	
Pipeline step	Sample A	Sample B
Human unmapped (input)	4,698,672 (31.0 %)	4,961,067 (36.6 %)
Quality filtering	1,355,915 (22.5 %)	2,020,118 (14.9 %)
Ribosomal alignment	1,043,952 (22.2 %)	1,795,032 (13.3 %)
BLAST alignment	1,670,204 (11.0 %)	4478 (0.03 %)
Bacteria alignment filtering	1,434,098 (9.5 %)	49 (0.0004 %)

Instead, in the GSE68086 dataset DecontaMiner detected several matches to bacterial reference sequences. In particular, we focused on those samples having more than 10 % of human-unaligned reads. A modest amount of reads matched to fungal genomes, whereas many reads aligned to bacteriophages specific for the identified bacteria (namely *Enterobacteria phage* and *Propionibacterium phage*). This last finding further confirmed the accuracy of the bacteria identification. As an example, the number of reads in two samples before and after the filtering and rRNA alignment processes are shown in Table 2. Sample A and Sample B had low mapping rates on the reference genome 69 and 63.4 %, respectively. However, the reason for such a high number of unmapped reads is completely different. Most of the alignment failure of the Sample B is due to the presence of low quality reads, that are approximately 22 % of the total raw reads, and only 0.0004 % reads matched correctly to bacteria, according to our setting. Instead, only 7.5 % of the sample A are low quality reads and almost 10 % significantly matched to bacteria. As shown by the barplots generated by a Matlab in-house script, Figs. 2 and 3, both healthy and tumor samples contain non-human sequences. For each sample, we plotted only the organisms having number of matched reads greater than 20 % of the total. All the species that do not fit this criterion are reported as "Others."

Propionibacterium acnes and *Escherichia coli* species were detected in almost all tumor samples and healthy donors, suggesting the possibility of a downstream contamination of the samples or some kind of machine artifacts. *P. acnes* is a gram-positive bacterium that forms part of the normal flora of the skin [25] and it is usually considered a contaminant of blood cultures [26]. *E. coli* is a gram-negative bacterium, host of the normal intestinal flora, but also one of the most common responsible of a wide variety of hospital and community-onset infections, affecting patients with normal immune systems as well as those immunodepressed [27].

One of the healthy samples, as well as one of the hepato-biliary carcinoma group, did not have a significant number of reads matching to any bacterial species, according to our thresholds.

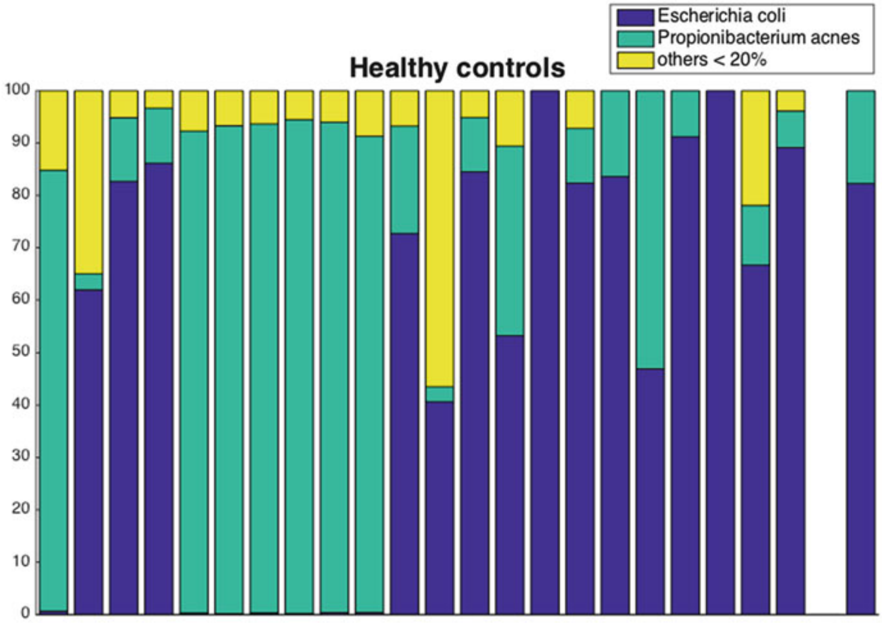


Fig. 2 Healthy controls barplot. For each healthy sample a bar reports the detected contaminating organism (*colors*) and percentage of unmapped reads assigned to each of them

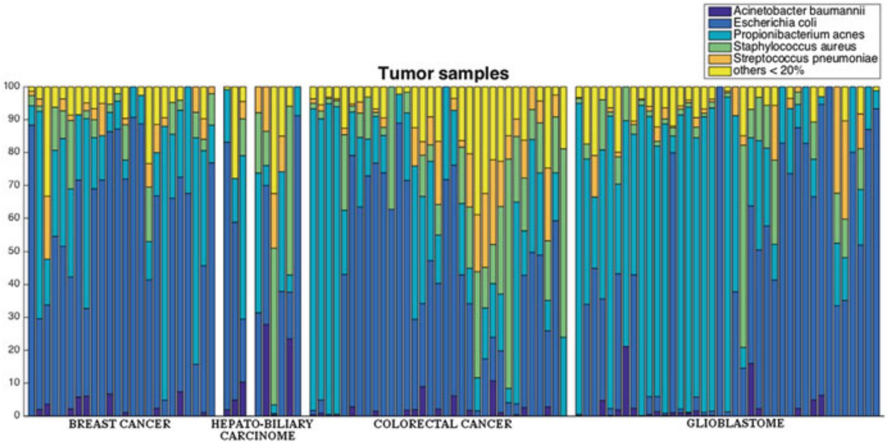


Fig. 3 Tumor samples barplot

The tumor samples barplot shows the presence of some bacterial species that are absent in control samples, or present with a very low reads number. Among them is worth to note the bacterium *Acinetobacter baumannii*. The percentage of reads aligned to *A. baumannii* is particularly evident in hepato-biliary carcinoma, although its presence seems to be independent of cancer type.

The genus Acinetobacter, as currently defined, comprises gram-negative, strictly aerobic, nonfermenting, nonfastidious, nonmotile, catalase-positive, and oxidase-negative bacteria [28]. *A. baumannii* normally inhabits human skin, mucous membranes, and soil [29]. *Acinetobacter baumannii*, in particular, has become one of the major causes of nosocomial infections during the past two decades [28, 30–32] and its correlation with outcomes of cancer patients is a clinical issue under study [33, 34].

4 Conclusions

The DecontaMiner pipeline was designed and developed to investigate the presence of contaminating sequences in NGS data. It has a dual utility, both as a filtering tool to remove foreign reads from the raw sequencing file, usually in fastq format, and as a detection tool to identify contaminating sequences among the unmapped reads, provided as a bam file. In order to test our pipeline we used two different RNA-seq datasets. The lack of matches to microorganisms in case of the polyA-RNA (GSE69240) demonstrates that the risk of incurring into false positive results is very low. The reliability of our pipeline is further proved on the total RNA (GSE68086) dataset analysis. Indeed, we found some kind of background contamination in almost all the samples. The most present organisms are *P. acnes* and *E. coli* and, in addition, some tumor samples significatively matched to *A. baumannii*, that it is a well-known nosocomial pathogen, even probably associated with outcomes of cancer diseases. It is important to underline that DecontaMiner can suggest the presence of contaminating sequences, but this results must be confirmed by an experimental validation. As an added value, the output fasta files and BLAST tables can be easily uploaded to MEGAN5 [35], a metagenome analyzer, which allows to obtain more detailed information about the taxonomy profile of the samples in several graphical modes. We are currently working to provide DecontaMiner as a Bash shell command-line tool, usable on a common laptop as well as in a distributed computing environment. We are also planning to put together the pipeline here developed and the Transcriptator tool [36] developed in our lab to provide an integrated environment for the analysis of omics data.

Acknowledgements Ilaria Granata was supported by MIUR PON02-00619. Mara Sangiovanni was supported by Interomics Italian Flagship Project. Mario R. Guarracino work has been conducted at National Research University Higher School of Economics (HSE) and has been supported by the RSF grant n. 14-41-00039.

References

1. Conesa, A., et al.: A survey of best practices for RNA-seq data analysis. Genome Biol. **17**(1), 1–19 (2016)
2. Laurence, M., Hatzis, C., Brash, D.E.: Common contaminants in next-generation sequencing that hinder discovery of low-abundance microbes. PLoS One **9**(5), e97876 (2014)
3. Feng, H., et al.: Clonal integration of a polyomavirus in human Merkel cell carcinoma. Science 319(5866), 1096–1100 (2008)
4. Palacios, G., et al.: A new arenavirus in a cluster of fatal transplant-associated diseases. N. Engl. J. Med. **358**(10), 991–998 (2008)
5. Bhatt, A.S., et al.: Sequence-based discovery of Bradyrhizobium enterica in cord colitis syndrome. N. Engl. J. Med. **369**(6), 517–528 (2013)
6. Strong, M.J., et al.: Microbial contamination in next generation sequencing: implications for sequence-based analysis of clinical samples. PLoS Pathog. **10**(11), e1004437 (2014)
7. Tae, H., et al.: Large scale comparison of non-human sequences in human sequencing data. Genomics **104**(6), 453–458 (2014)
8. Ouma, W.Z., et al.: Important biological information uncovered in previously unaligned reads from chromatin immunoprecipitation experiments (ChIP-Seq). Sci. Rep. **5**, 8635–8635 (2015)
9. Kostic, A.D., et al.: PathSeq: software to identify or discover microbes by deep sequencing of human tissue. Nat. Biotechnol. **29**(5), 393–396 (2011)
10. Borozan, I., et al.: CaPSID: a bioinformatics platform for computational pathogen sequence identification in human genomes and transcriptomes. BMC Bioinf. **13**(1), 206 (2012)
11. Altschul, S.F., et al.: Basic local alignment search tool. J. Mol. Biol. **215**(3), 403–410 (1990)
12. Naccache, S.N., et al.: A cloud-compatible bioinformatics pipeline for ultrarapid pathogen identification from next-generation sequencing of clinical samples. Genome Res. **24**(7), 1180–1192 (2014)
13. Li, H., et al.: The sequence alignment/map format and SAMtools. Bioinformatics **25**(16), 2078–2079 (2009)
14. Quinlan, A.R., Hall, I.M.: BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics **26**(6), 841–842 (2010)
15. Gordon, A., Hannon, G.J.: FastX Toolkit (2010) http://hannonlab.cshl.edu/fastx_toolkit/index
16. Kopylova, E., Noé, L., Touzet, H.: SortMeRNA: fast and accurate filtering of ribosomal RNAs in metatranscriptomic data. Bioinformatics **28**(24), 3211–3217 (2012)
17. Zhang, Z., Schwartz, S., Wagner, L., Miller, W.: A greedy algorithm for aligning DNA sequences. J. Comput. Biol. **7**, 203–214 (2000)
18. Best, M.G., et al.: RNA-Seq of tumor-educated platelets enables blood-based pan-cancer, multiclass, and molecular pathway cancer diagnostics. Cancer Cell **28**(5), 666–676 (2015)
19. Abba, M.C., et al.: A molecular portrait of high-grade ductal carcinoma in situ. Cancer Res. **75**(18), 3980–3990 (2015)
20. Leinonen, R., Sugawara, H., Shumway, M.: The sequence read archive. Nucleic Acids Res. **39**, D19–D21 (2010).
21. Bolger, A.M., Lohse, M., Usadel, B.: Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics **30**, 2114–2120 (2014).
22. Andrews, S.: FastQC: a quality control tool for high throughput sequence data. Reference Source (2010)
23. Trapnell, C., Pachter, L., Salzberg, S.L.: TopHat: discovering splice junctions with RNA-Seq. Bioinformatics **25**(9), 1105–1111 (2009)
24. Lassmann, T., Hayashizaki, Y., Daub, C.O.: SAMStat: monitoring biases in next generation sequencing data. Bioinformatics **27**(1), 130–131 (2011)
25. Perry, A., Lambert, P.: Propionibacterium acnes: infection beyond the skin. Expert Rev. Anti-Infect. Ther. **9**(12), 1149–1156 (2011)
26. Park, H.J., et al.: Clinical significance of Propionibacterium acnes recovered from blood cultures: analysis of 524 episodes. J. Clin. Microbiol. **49**(4), 1598–1601 (2011)

27. Pitout, J.D.D.: Extraintestinal pathogenic Escherichia coli: an update on antimicrobial resistance, laboratory diagnosis and treatment. Expert Rev. Anti-Infect. Ther. **10**(10), 1165–1176 (2012)
28. Peleg, A.Y., Seifert, H., Paterson, D.L.: Acinetobacter baumannii: emergence of a successful pathogen. Clin. Microbiol. Rev. **21**(3), 538–582 (2008)
29. Manchanda, V., Sanchaita, S., Singh, N.P.: Multidrug resistant acinetobacter. J. Global Infect. Dis. **2**(3), 291 (2010)
30. Fukuta, Y., et al.: Risk factors for acquisition of multidrug-resistant Acinetobacter baumannii among cancer patients. Am. J. Infect. Control **41**(12), 1249–1252 (2013)
31. Al-Hassan, L., El Mehallawy, H., Amyes, S.G.B.: Diversity in Acinetobacter baumannii isolates from paediatric cancer patients in Egypt. Clin. Microbiol. Infect. **19**(11), 1082–1088 (2013)
32. Dijkshoorn, L., Nemec, A., Seifert, H.: An increasing threat in hospitals: multidrug-resistant Acinetobacter baumannii. Nat. Rev. Microbiol. **5**(12), 939–951 (2007)
33. Ñamendys-Silva, S.A., et al.: Outcomes of critically ill cancer patients with Acinetobacter baumannii infection. World J. Crit. Care Med. **4**(3), 258 (2015)
34. Nazer, L.H., et al.: Characteristics and Outcomes of Acinetobacter baumannii Infections in Critically Ill Patients with cancer: a matched case-control study. Microb. Drug Resist. 21(5), 556–561 (2015)
35. Huson, D.H., Weber, N.: Microbial community analysis using MEGAN. Methods Enzymol. **531**, 465–485 (2012)
36. Tripathi, K.P., Evangelista, D., Zuccaro, A., Guarracino, M.R.: Transcriptator: an automated computational pipeline to annotate assembled reads and identify non coding RNA. PLoS One **10**(11), e0140268 (2015)