Noor Ahmad Shaik
Khalid Rehman Hakeem
Babajan Banaganapalli · Ramu Elango
Editors

Essentials of Bioinformatics, Volume II

In Silico Life Sciences: Medicine

Springer

Essentials of Bioinformatics, Volume II

Noor Ahmad Shaik • Khalid Rehman Hakeem
Babajan Banaganapalli • Ramu Elango
Editors

Essentials of Bioinformatics, Volume II

In Silico Life Sciences: Medicine

Springer

Editors

Noor Ahmad Shaik
Princess Al-Jawhara Center of Excellence
in Research of Hereditary Disorders
Department of Genetic Medicine
Faculty of Medicine
King Abdulaziz University
Jeddah, Saudi Arabia

Babajan Banaganapalli
Princess Al-Jawhara Center of Excellence
in Research of Hereditary Disorders
Department of Genetic Medicine
Faculty of Medicine
King Abdulaziz University
Jeddah, Saudi Arabia

Khalid Rehman Hakeem
Department of Biological Sciences
King Abdulaziz University
Jeddah, Saudi Arabia

Princess Dr. Najla Bint Saud Al-Saud
Center for Excellence
Research in Biotechnology
King Abdulaziz University
Jeddah, Saudi Arabia

Ramu Elango
Princess Al-Jawhara Center of Excellence
in Research of Hereditary Disorders
Department of Genetic Medicine
Faculty of Medicine
King Abdulaziz University
Jeddah, Saudi Arabia

ISBN 978-3-030-18374-5 ISBN 978-3-030-18375-2 (eBook)
https://doi.org/10.1007/978-3-030-18375-2

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

Dr. Ramu Elango *dedicates the book to his wife, Karpagam; their children, Madhivanan and Ilankeeran; and his brothers, Govindan and Prof Krishnan Ramu and Anni-Vijaya Krishnan*

Dr. Babajan Banaganapalli *dedicates the book to his family, Khader Basha, Shanu, Shanawaz, Shavar, and Shaju, wife, Gowsia; and their children, Rayyan and Reenah*

Dr. Noor Ahmad Shaik *dedicates the book to his parents, Jan Ahmad and Abida Khatoon; siblings (Nazmul, Jeelani, Reshma and Nazia); wife, Dr. Fatima; and children, Maryam, Sakina, and Raheel*

Dr. Khalid Rehman Hakeem *dedicates the book to his beloved family: Papa, Mom, Asma, and Hibah*

for their sacrifices and continued support*.*

Foreword

The golden era of the biological understanding of the health and disease is now unfolding the nature secrets in an unprecedented level. This revolution is not possible without the contributions of the scientists across the world as well as in many subjects, ranging from biologist; physicists; engineers of the variety of fields like electrical, mechanical, and computer science; as well innovators of new ideas. This free flow of ideas from people with different skills resulted in bringing the new technologies and their implementation in a way which was never in doubt changed the way we look at the life around as well as inside us from the environment and the interaction with genomes resulting in adaption and better management plans of our lives with improved health.

Biomedical scientists were the major beneficiaries of such advances of the variety of fields mentioned above. Almost 50 years separated the discovery of DNA structure to the Human Genome Project achievement. Thousands of scientists developed new methods and technologies from sequencing and computers with power to deal with complex data generation to analysis. The proverb “Necessity is the mother of invention” explains the development of bioinformatics aptly. For example, with the generation of sequences came the first step of theoretical scientists, mathematicians and statisticians put the first seed for BLAST, to compare different sequences now to deal with high-throughput data from a spectrum of “-omic” technologies like genomics, proteomics, and metabolomics. Comparing fewer than 1000 bases in early years to millions and billions of bases and data points in biology in a short time with a variety of tools resulted in the rapid development in diagnosis of many genetic defects in rare diseases to identification of hundreds and thousands of risk markers for the complex diseases plaguing the human race at an alarming rate. Now, thanks to the bioinformatics tool, biologist with limited or no knowledge of computer programs can analyze the complex data from a variety of high-throughput “-omic” fields to search for the answer to their scientific queries.

This book series is trying to target the graduate students and young researchers who are keen in understanding and contemplating their future career in high-throughput biological fields of their choice. The chapters give the flavor of the various fields from genetic diagnosis, the dissection of complex diseases to application, and the collaborative efforts of bioinformatics scientists with geneticists, statisticians, biochemists, and engineers to deliver the new understanding of the human biology. The first volume showed a variety of tools available in bioinformatics field to address a variety of queries with different sets of biological data. The current second volume gives a glimpse of the success of such technologies and bioinformatics tools in many fields, changing the disease diagnostics and novel drug identification to better patient management with better drugs and exploring the revolutionary stem cell science to treat patients of devastating diseases. The editors were bold enough to take the task of assembling a group of senior scientists and young people, who understand the need and difficulties of young researchers and graduate students, in unravelling the myth that advanced biological research is unreachable to them in a simple format. I congratulate the senior and experienced authors of various chapters and editors for providing an excellent overview, highlighting the impact of bioinformatics and "-omic" technologies across many fields to improve the human welfare.

I strongly recommend this volume series to the young students and budding researchers wishing to enter this exciting era of biomedical revolutionary research. I am confident this series of volumes will provide the confidence to science students in different corners of the world, especially from the developing world with limited resources, to dream up the careers in this field to make an impact on the world.

Prof. Kaipa Prabhakar Rao
Department of Genetics
Osmania University
Hyderabad, India

Preface

Bioinformatics is growing along with the rapid advances in many different technological and scientific fields. The “big data” science is the result of combined work of ultrahigh-throughput technology development and high-performance computers. Genetics, genomics, proteomics, metabolomics, and metagenomics changed the biology more in the recent past. Next-generation sequencing technology is the result of Human Genome Project with whole-exome and whole-genome sequencing (NGS) possible within 24–56 hours. This revolutionized the genetic diagnosis of rare diseases around the world. Almost every country has the scientists equipped with the NGS data analysis skills for diagnostic purposes. Bioinformatics tools, especially in the public domain, make this technology for research and application in diagnosis a reality in every corner of the world. Many nations realized the potential of the national biobank and their potential contribution to the economy by reducing the healthcare burden enormously through the prevention of disease and/or better management of patients through novel drug discovery to personalized medicine.

This volume, like the first volume, is targeting the young researchers to make them aware of the recent developments in a variety of fields where bioinformatics along with the other multi-omics technologies changed the scientific world, making a large impact. It also focusses on the key development in key multi-omics technologies output and their impact in many aspects of biomedical fields. Human genome sequencing project witnessed a heightened activity of bioinformatics scientists and tools. Hundreds and thousands of the easy tools were developed for a variety of applications. It is not possible to discuss the examples for any single group of bioinformatics tools. Hand in hand with the first volume, this will help the young scientists and graduate students realize the role of bioinformatics play in the development of many applied biomedical advances toward better healthcare for all.

The chapters are organized in a way to highlight a particular “-omic” technology and its role in changing the biomedical scientific area. For example, the microarray and NGS technology, combined with the bioinformatics tools, made the genetic diagnosis rapid and accurate, even for rare diseases in any corner of the world with very little blood within days. Unknown diseases reveal novel hidden mutations, helping the scientists learn more about the disease biology to address the biological understanding in finer detail. Likewise, drug discovery and personalized medicine had the bioinformatics stamped its impact along with the technologies. We, as scientists, attempted to highlight the success of various biomedical fields in this volume to support the role of collaborative nature of modern science among the multidisciplinary scientists. It is the celebrations of the collaborative scientists ranging from physical to applied medical and clinical scientists with bioinformatics groups, directly or indirectly, through many software tools or specialized databases with hidden tools to provide the accurate answer to their queries. Hopefully, the young scientists will realize the importance of this type of multidisciplinary collaboration and gain success in their professional careers.

We sincerely thank the management, faculty members, staff, and students at Princess Al-Jawhara Center of Excellence in Research of Hereditary Disorders (PACER-HD), Department of Genetic Medicine, Faculty of Medicine, and Department of Biology, Faculty of Science, at King Abdulaziz University (KAU) for supporting our effort in bringing this book series a reality. Our special thanks go to Prof. Jumana Y. Al-Aama, director of PACER-HD, KAU, for letting us realize the importance of bioinformatics in clinical practice, for encouraging excellent scientific discussions and raising critical questions as clinicians, and for supporting our work throughout this long process. We would also like to thank the chairman of the Department of Biological Sciences, Prof. Khalid M. AlGhamdi, and the head of Plant Sciences Section, Dr. Hesham F. Alharby, for providing us the valuable suggestions and encouragement to complete this task. We also acknowledge the authors of all chapters who spared their precious time in bringing this book out with valuable contributions. Last but not the least, we would like to acknowledge Springer Nature publishers, especially Mr. Rahul Sharma, for their patience and regular communication with us to move the project forward.

Jeddah, Saudi Arabia Noor Ahmad Shaik
Jeddah, Saudi Arabia Khalid Rehman Hakeem
Jeddah, Saudi Arabia Babajan Banaganapalli
Jeddah, Saudi Arabia Ramu Elango

Contents

About the Editors

Noor Ahmad Shaik is an academician, researcher, and technologist working in the field human molecular genetics. Over the last 15 years, he has been working with different research groups whose fundamental focus is to offer genetic disease diagnostics, management, and therapy. Currently, he is working to improve the current mutation prediction methods through integrative computational algorithms, so that clinicians and scientists can better understand the functional relevance of genetic mutations to disease, and is rendering his editorial services to world-renowned journals like the *Frontiers in Pediatrics* and *Frontiers in Genetic Disorders*. He is interested in discovering the novel causal genes/biomarkers for rare hereditary disorders and also in understanding the effect of mutations on structure and function of causal proteins of human diseases. He has already published 42 research publications in reputed international journals of human genetics and bioinformatics and has been a recipient of several research grants from national and international funding agencies.

Khalid Rehman Hakeem is Professor at King Abdulaziz University, Jeddah, Saudi Arabia. He has completed his PhD (Botany) from Jamia Hamdard, New Delhi, India, in 2011 and has worked as postdoctorate fellow in 2012 and fellow researcher (associate prof.) from 2013 to 2016 at the Universiti Putra Malaysia, Selangor, Malaysia. His specialty includes Plant Ecophysiology, Biotechnology and Molecular Biology, Plant-Microbe-Soil Interactions, and Environmental Sciences. So far, he has edited and authored more than 25 books with Springer International, Academic Press (Elsevier), CRC Press, etc. and has also, to his credit, published more than 120 research publications in peer-reviewed international journals, including 42 book chapters in edited volumes with the international publishers.

Babajan Banaganapalli works as bioinformatics research faculty at King Abdulaziz University, where he initiated and successfully run the interdisciplinary Bioinformatics program from 2014 till to date. He has more than 12 years of research experience in bioinformatics; has published more than 40 journal articles, conference papers, and book chapters; and has also served in numerous conference program committees, organized several bioinformatics workshops and training programs, and acted as editor and reviewer for various international genetics/bioinformatics journals. His research interests spread across genomics, proteomics, and drug discovery for complex diseases. Recently, he was honored as young scientist for his outstanding research in bioinformatics by Venus International Research Foundation, India.

Ramu Elango is a well-experienced molecular geneticist and computational biologist with extensive experience at MIT, Cambridge, USA, and GlaxoSmithKline R&D, UK, where he contributed extensively in many disease areas of interest in identifying novel causal genes and tractable drug targets, after completing his PhD in Human Genetics at All India Institute of Medical Sciences, New Delhi, India. He presently heads the Research and Laboratories at the Princess Al-Jawhara Center of Excellence in Research of Hereditary Disorders, King Abdulaziz University. His research focus is on genetics and genomics of complex and polygenic diseases. His team exploits freely available large-scale genetic and genomic data with bioinformatics tools to identify the risk factors or candidate causal genes for many complex diseases.

Chapter 1
Driving Forces of Bioinformatics

Ramu Elango, Babajan Banaganapalli, and Noor Ahmad Shaik

Contents

1.1 Introduction

Bioinformatics focused on sequence analysis processes at inception, starting with Genbank and BLAST services. When more and more sequence data were submitted to the GenBank from various organisms by scientists from different parts of the world, the scope changed into more critical aspects of characterization of the gene, mapping, gene function and variant effect, etc. with dramatic increase in sequence data possible by various technological breakthroughs in many fields. There were many reviews which discussed these breakthrough technologies in detail over the

R. Elango (✉) · B. Banaganapalli · N. A. Shaik
Princess Al-Jawhara Center of Excellence in Research of Hereditary Disorders,
Department of Genetic Medicine, Faculty of Medicine,
King Abdulaziz University, Jeddah, Saudi Arabia
e-mail: relango@kau.edu.sa; bbabajan@kau.edu.sa; nshaik@kau.edu.sa

N. A. Shaik et al. (eds.), *Essentials of Bioinformatics, Volume II*,
https://doi.org/10.1007/978-3-030-18375-2_1

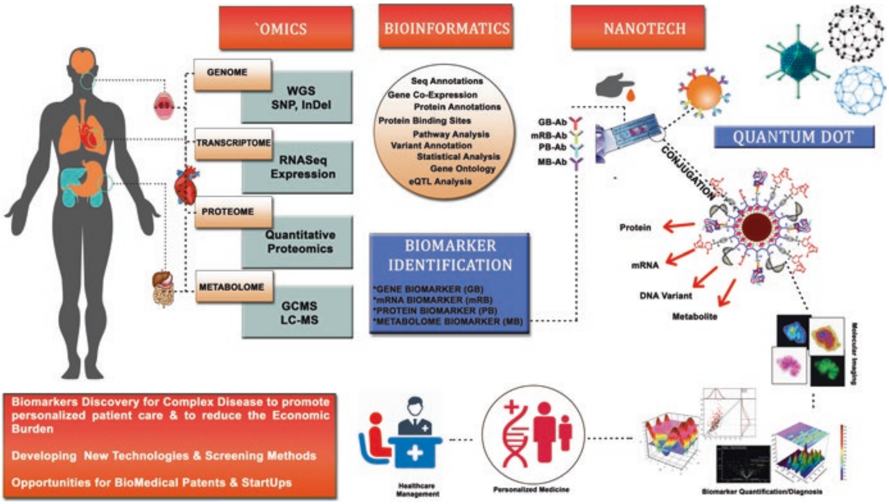


Fig. 1.1 Potential role of bioinformatics along with -Omic technologies in Health care

years. These will not be discussed here. For example, the polymerase chain reaction (PCR) using Taq polymerase enzyme is one key technology for amplifying very little amount of DNA into millions of copies. Different types of PCR, like real-time PCR, reverse transcriptase PCR (R-T PCR), SNaPshot, etc., helped drive the application of sequence variations and gene expression for the diagnosis or biology of the defective gene function. These techniques along with instruments like genome analyzer sequencing machines played the key role in early stages of genetic revolution in biomedical field. With more advanced instruments developed to generate high-throughput sequence data with minimal amount of source DNA as in next-generation sequencing machines turned bioinformatics into one of the main players in biomedical field (Fig. 1.1).

1.2 Sequencing Technologies

Sequencing the DNA and RNA moved from Sanger sequencing few hundred bases at a time to hundreds of millions of bases in a day in the last decade. Introduction of genome analyzer to the research community led the scientists to embrace the new method quickly. Exciting research output with this method pushed many chemists and instrument engineers to develop newer and faster technologies to sequence the DNA and RNA. Availability of the cheaper technology and machines spurred more research on highly devastating diseases like cancer and complex diseases like myocardial infarction, dementia, and so on. Next-generation sequencing technologies introduced new machines especially from Illumina, Thermo Fisher, and Pacific Biosciences (PacBio), companies which made the high-throughput rapid sequencing to the forefront of the biomedical research. Large-scale generation of sequences of these NGS technologies provided the biggest challenge to the bioinformatics scientists to analyze and identify crucial markers and genes of interest for the bench scientists for functional validation.

Major NGS technologies and machines: Success of NGS depends on the chemistry and instrumentation behind the sample preparation. Three leading products from three different technologies contributed heavily for many major programs across the globe including the Human Genome Project, The Cancer Genome Atlas, etc.

Illumina/Solexa system uses the sequencing by synthesis method. They have successfully marketed different products targeting small laboratories to large genome centers. Major instruments in their portfolio include MiSeq, Nextseq, HiSeq, and NovaSeq. They are the market leaders with ~ 85–90% of NGS machines in labs around the world, producing large data every day for many groups.

1.2.1 MiSeq

This machine is the basic NGS machine that can be used for whole exome sequencing (WES) if required, but it is more suitable for targeted sequencing of the multigene panel for screening, especially in genetic diagnosis or screening for validation of novel sequence variants associated with complex diseases.

1.2.2 Nextseq

This is the midrange product from Illumina used for WES and whole genome sequencing (WGS) as well as RNA seq for gene expression profile of the specific tissues.

1.2.3 HiSeq

This group of instruments initially targeted the big genome centers and core laboratories in research organizations and universities across the world. This is one of the high-throughput sequencing platform machines which is suitable for large-scale WGS sequencing projects like Human Genome Project as well as running thousands of WES of populations for unique coding sequence variant identifications. Almost all genome centers across the world installed these machines along with commercial NGS providers like BGI, China; Macrogen, Korea; etc. Clinical diagnostics companies also use variety of these machines for the clinical diagnosis of rare diseases for clinicians from countries where the facility is not available.

1.2.4 NovaSeq

This is the latest new range of product released by Illumina for the large core facilities and commercial NGS providers for research and diagnosis. This uses refined and better technologies to provide better and faster results for the users.

1.2.5 Thermo Fisher Scientific Range

The *ion proton range* of products marketed by the Thermo-Fisher Life Biotechnologies is mainly used for multi-gene panel screening of cancer or different diseases. Though they targeted the WGS and WES customers initially, they are being used mainly for the mutation screening in many labs.

1.2.6 Nanopore Technology

Hand-held sequencing machine suitable for field laboratory and epidemic outbreak infection detection and work can be carried out in remote areas and data transferred to central laboratory for detailed analysis and to get rapid results. Newer machines are released with better options for long read sequences and high-throughput data generation by Oxford Nanopore technologies company (Oxford, UK). A nano-scale hole in proteins (biological nanopores) or in solid materials (solid-state nanopores) is called nanopore. One commercially scalable instrument set (MinION, GridION, and PromethION) from this company exploited this successfully with protein nanopores set in electrical-resistant polymer membrane. This instrument, passing the ionic current through nanopores, detects the differences in current variations of DNA/RNA when one nucleotide molecule at a time passes through it. Different nucleotides, A, T, G, and C, have unique current property – molecular signature – which can be used to identify that molecule, when the sample DNA goes through the nanopore.

1.2.6.1 Applications of Nanopore Sequence Technologies

Applications of this technology are extensively used for whole genome sequencing, exome sequencing, and RNA Seq and are reported in diverse fields of biology and biomedical sciences from genetics, genomics, and metagenomics. Main advantage of this technology is to reach remote areas where the sophisticated NGS technology is not available for rapid screening of biological samples. Its use can be noted by the more than 200 articles on epidemic infections in rural populations through the detection of causal organisms. They in turn help preventing the disease in living organisms. Excellent reviews were published in many leading journals which will give better understanding of the technology and its applications.

1.2.7 Single Cell Genomics

10xgenomics. This rapidly evolving technology is already making big impact on many complex disease biology, like in epigenomics, melanoma, inflammatory bowel disease, etc. (AlJanahi et al. 2017; Kinchen et al. 2018; Rambow et al. 2018). Publications and research is growing faster than one can catch up with advances in this exciting area, where collaborative interaction between scientists from different areas of science with bioinformatics is making a big impact.

1.2.8 Biostatistics

Genetic research in the population owes its growth and advancement to the statistical groups and population geneticists. Biostatistics has evolved through the interesting collaborations and questions raised in biology, especially in genetics. Face of genetic and genomic revolution of human diseases, especially the complex diseases seen across many countries, changed the biological understanding of the complex diseases to an unprecedented fine detail. One area of genetics which is dramatically changed by statistical groups is genetic association studies in complex diseases. From single genetic marker mapping to genetic association study to a trait with the help of human genome markers and sequences, to genome wide association studies (GWAS), the impact of statistical science is immense. The evolution of the GWAS studies through the eyes of one of the leading groups of statistical group is published in a recent review (Visscher et al. 2018). Gene expression analysis from microarray experiments was strengthened by the stringent statistical process, which continued its influence in all fields of high-throughput biological data generation areas like metabolomics, proteomics, metagenome, and gene enrichment analysis incorporated into the pathway analysis and other advanced applications. Statistics is one of the most important areas that all scientists, whether wet lab, biomedical or bioinformatic scientist, need to learn to apply to their study. Many biology-related degree programs made the introductory biostatistics course as prerequisite or compulsory courses to graduate. Bioinformatics scientists working with statisticians provided hundreds of tools which carry out many statistical tests in the background to provide the statistically stronger analysis output, like MetaboAnalyst for metabolomics or GEO-R2 in NCBI with gene expression data.

1.3 High-Throughput Technologies and Bioinformatics Changing the Biology

From simple sequence analysis to functional effect prediction of variants to miRNA localization, transcription factor analysis, etc. pushed the bioinformatics to the forefront of biomedical research. *Nucleic Acids Research* journal has been publishing special issues on different sequences, polymorphic variants, pathways, gene sets, etc. which have highlighted the contribution of bioinformatics groups around the world (Web Server Issue 2017). The bioinformatics groups not only built the databases but also created many powerful and simple-to-use analytical tools to query these databases. Such simple tools helped the bench scientists with limited bioinformatics exposure access to large experimental data at their fingertips for instant decision-making process, planning future experiments, comparing their own experimental results for independent validation, discovering the novel functions or mechanisms, etc.

The special database issues in *Nucleic Acids Research* provide update on newer tools, technologies, and applications. Many bioinformatics journals started publishing new tools developed by groups of scientists, driven by the requirement of specialized data analysis for many bench scientists. Such simple and better tools are being released by scientists every day for the benefit of bench scientists. Due to rapid growth in a variety of databases storing different sets of data, many bioinformatics groups design new powerful integrated genetic and genomic analysis tools to query databases. These novel tools help scientists to address the challenging questions in biomedical fields, from diagnosis to drug development, disease prevention, patient management, etc.

1.4 Gene Expression Profile

Gene expression profiling started with single gene expression in Northern blotting to high-throughput gene expression microarray chips and RNA Seq with genome wide expression profiles for the most commonly studied organisms. This high-throughput gene expression profiling technology with thousands of probes leads to “information overload” with limited suite of bioinformatics tools when these technologies were introduced. Illumina and ThermoFisher are the leading commercial companies which supply microarray-based gene expression chips for a variety of research activities. These array data are captured and analyzed by their own suite of bioinformatic software which incorporates the statistical functions to normalize the gene expression data across the chip, quantify the signal into gene expression level, and compare multiple samples across all genes with multiple testing correction options to suit the sample size and objectives. Realizing the limitations of the microarray and/or RNA Seq instrument-linked bioinformatic suite of programs, many groups started developing their own rigorous statistically robust software tools which can be flexible with added features to suit their interest to address the challenging queries. With these tools and easy access to the data generation technology, large number of studies explored complex questions of biological process of specific tissues, cell types, and organs.

The NCBI (Gene Expression Omnibus-GEO web link: https://www.ncbi.nlm.nih.gov/geo/) and EMBL (Expression Atlas Web link: https://www.ebi.ac.uk/gxa/home) are the major storehouses of gene expression profiles of different experimental condition for a variety of tissues and cell types to provide easy access to large-scale data to the scientific community for various organisms. These gene expression databases collected data from thousands of experiments carried out by scientists in many countries for other scientists. These databases along with bioinformatic analysis tools will be one of the powerful combinations in unravelling the biology of normal and affected tissues in patients with a disease of interest. New database of single cell transcriptomics in many organisms is established in European Bioinformatics Institute, UK, generated with advanced sequencing and other technologies (Single Cell Expression Atlas- Web link: https://www.ebi.ac.uk/gxa/sc/home). Rapid growth of this database will have much bigger impact on the understanding of the biology of complex diseases in specific tissues. Rapid acceleration of high-throughput data generation technologies along with the powerful bioinformatics tools drives many bench scientists to delve into these data to address scientifically challenging questions to propose new hypothesis, validate indirect evidence from other experimental data, and open up the new area of in silico biology with more questions. The Genotype Tissue Expression (GTEx: http://gtexportal.org/home/) project collected 53 tissues from 1000 individuals for high-throughput molecular studies using WES, WGS, and RNA Seq to understand the tissue-specific gene expression and regulation by genetic variants around the genes. This resource is valuable in validating many experimental results of groups of scientists who cannot afford such large-scale study on their own to support their scientific results.

The special issue of *Nucleic Acids Research* journal regularly publishes many special issues on databases and web server—bioinformatics tools available in the public domain. The recent issue of web server was published in July 2018 issue. This is the 16th annual issue published by the *Nucleic Acids Research* journal. Specialized bioinformatics journals also publish hundreds of new tools with their applications in detail, if one is interested in finding some good public domain tools.

The impact of bioinformatics along with the rapidly changing biological technologies from genetics, genomics to metabolomics and metagenomics is enormous. To cover all aspects, many volumes of updates are required in many fields. Integrated multi-omics data analysis, especially from DNA, RNA, protein, metabolite to Metagenomics, is now possible with the combinations of advanced technologies, some of which are covered, and many are not addressed in detail. Following chapters will give a glimpse of applications of this revolution in few selected fields. Many areas are not covered, not intentionally, as they are rapidly changing field with new development and applications like metagenomics and proteomics.

References

AlJanahi AA, Danielsen M, Dunbar CE (2017) An introduction to the analysis of single-cell RNA-sequencing data. Mol Ther Methods Clin Dev 10:P189–P196. https://doi.org/10.1016/j.omtm.2018.07.003

EBI Expression Atlas link: https://www.ebi.ac.uk/gxa/home

EBI Single Cell Expression Atlas link: https://www.ebi.ac.uk/gxa/sc/home

GEO database web link: https://www.ncbi.nlm.nih.gov/geo/

GTEx Database: http://gtexportal.org/home/ (Not works with IE interface)

Kinchen J, Chen HC, Parikh K, Antanaviciute A, Jagielowicz M, Fawkner-Corbett D, Ashley N, Cubitt L, Mellado-Gomez E, Attar M, Eshita Sharma E, Wills Q, Bowden R, Richter FC, Ahern D, Puri KD, Henault J, Gervais F, Koohy H, Simmons A (2018) Structural remodeling of the human colonic mesenchyme in inflammatory bowel disease. Cell 175:372–386. https://doi.org/10.1016/j.cell.2018.08.067

Rambow F, Rogiers A, Marin-Bejar O, Aibar S, Femel J, Dewaele M, Karras P, Brown D, Chang YH, Debiec-Rychter M, Adriaens C, Radaelli E, Wolter P, Bechter O'Dummer R, Levesque M, Piris A, Frederick DT, Boland G, Flaherty KT, van den Oord J, Voet T, Aerts S, Lund AW, Marine J-C (2018) Toward minimal residual disease-directed therapy in melanoma. Cell 174:1–13. https://doi.org/10.1016/j.cell.2018.06.025

Visscher PM, Wray NR, Zhang Q, Sklar P, McCarthy MI, Brown MA, Yang J (2018) 10 years of GWAS discovery: biology, function, and translation. Am J Hum Genet 101:5–22. https://doi.org/10.1016/j.ajhg.2017.06.005

Web Server issue (2017) Nucleic Acid Res J 45:W1

Chapter 2
Genetic Association from RFLPs to Millions of Variant Markers: Unravelling the Genetic Complexity of Diseases

Babajan Banaganapalli, Noor Ahmad Shaik, Jumana Y. Al-Aama, and Ramu Elango

Contents

2.1 Introduction

Rapid technological advancements in deciphering the DNA architecture and organization of the genomes at various stages revolutionized the role of genetics in health and disease conditions. The old proverb "Necessity is the mother of invention" is applicable for the development of bioinformatics field in general. With technology rapidly driving large-scale genetic and genomic data generation, bottleneck issue was the analysis of such data. Biologists and biostatisticians started collaborating to work on the statistical programs to simplify the analysis

B. Banaganapalli (✉) · N. A. Shaik · J. Y. Al-Aama · R. Elango
Princess Al-Jawhara Center of Excellence in Research of Hereditary Disorders,
Department of Genetic Medicine, Faculty of Medicine,
King Abdulaziz University, Jeddah, Saudi Arabia
e-mail: bbabajan@kau.edu.sa; nshaik@kau.edu.sa; jalama@kau.edu.sa; relango@kau.edu.sa

N. A. Shaik et al. (eds.), *Essentials of Bioinformatics, Volume II*,
https://doi.org/10.1007/978-3-030-18375-2_2

of such large-scale data. Flood of data resulted in highlighting the limitations of such statistical tool with large-scale data analysis. With more powerful computers, speed of analysis is better, but still many biologists started moving toward computer scientists to devise a simple bioinformatics tool that can be handled by the biologists with limited knowledge of statistics as well as computer programs and UNIX platforms. Computer scientists with interest in biological queries started the bioinformatics revolution.

Earliest Period (1948–1970)
Geneticists were keen to identify the genetic factors contributing to the common diseases as well as to the monogenic diseases. Such work started after the Second World War in the late 1940s and continued for many decades. The ABO blood groups were the first such genetic markers tested for association with many blood-related diseases (O'Hanlon and Stewert 1948; Prest et al. 1955). Such studies continued for a spectrum of diseases of all types for many years, till HLA and new serum biomarkers arrived (Cameron and Izatt 1962; McGinniss et al. 1964; Patel et al. 1969; Simon et al. 1971).

2.2 Early DNA Marker Development and Application (Late 1970s and 1980s)

In 1978, Kan and Dozy (1978) successfully used DNA markers for the prenatal diagnosis by testing the amniotic fluid cells for sickle cell anemia. Kan et al. (1980) used the DNA markers for beta-thalassemia screening in Italy. The early 1908s witnessed flurry of papers with the identification of many restriction fragment length polymorphism (RFLP) markers (Sarfarazi et al. 1983; Nussbaum et al. 1983) and their application in genetic association studies for beta-thalassemia (Wainscoat et al. 1983) and for Duchenne muscular dystrophy (Harper et al. 1983). The HLA genotyping by serological cytotoxicity methods was introduced by Terasaki group (Patel et al. 1969). Highly polymorphic nature of HLA loci in the world population triggered a new approach to study the genetic association between these markers and a variety of diseases, especially the autoimmune diseases in the 1980s and 1990s. Many research groups reported a strong association between many diseases like rheumatoid arthritis, spondylitis etc,.

Examples of changing world of genetic association studies through years: Genetic association studies in inflammatory bowel disease started with complement component C3 polymorphism in Norwegian IBD and healthy control population (Elmgreen et al. 1984) by high-voltage plasma electrophoresis. They have reported a positive association. Such studies were continued through early parts of the 1990s.

2.3 Microsatellite Markers in Genome-Wide Association Studies

Genome-level orderly mapping of the genes and markers was possible with tools and techniques in many fields. Radiation hybrid panels, chromosome-specific genomic libraries, gene-specific probes, and microsatellite markers contributed to the rapid development of the human genome mapping (Gyapay et al. 1996). Identification of CA repeat microsatellite markers across many of the mammalian genomes resulted in generating more accurate genetic maps of the genomes (Dietrich et al. 1996; Schuler et al. 1996). Microsatellite markers across the genome were found to be highly polymorphic in many ethnic and racial groups. This highly polymorphic nature of the markers helped the geneticists to explore the genetic association studies at the genome level. At that time, genotyping of the CA repeat markers for any disease studies is time-consuming and resource-demanding. The progress was limited to a few centers in the world. These studies are limited not only in generating the large-scale data, but also in analysis of such data. Statistical groups slowly realized their important role in getting to analyze such data to identify the meaningful statistically significant genes' contribution in complex diseases like schizophrenia, cardiovascular diseases, etc. (Pulver et al. 1994) as well as mapping many monogenic diseases such as myotonic dystrophy and Huntington disease (Brook et al. 1991; Doucette-Stamm et al. 1991).

Slowly and steadily the momentum built on the success of such studies, which accelerated the progress of bioinformatics integration into genetic studies. Such studies for many diseases started accumulating new markers and mapping information for the bioinformatics scientists to play major role in genetics and genomics providing simple tools to search through the data. This allowed bench scientists to make a meaningful conclusion and accurate mapping of many disease loci leading to the identification of causal genes.

2.4 Human Genome Project and GWAS

Human genome project generated multifaceted applications of the reference human genome in a variety of studies, including the genetic association studies (NCBI, dbSNP, etc.). Millions of SNPs were identified in different ethnic groups which can be used for large-scale association studies. Increasing numbers of the SNPs (from 1000s to million SNPs) were being analyzed with SNP microarray chip technology. These studies changed our understanding of the genetics of complex polygenic diseases with more refined details. International collaborations on complex disease genetics led to "information overload" and extensive data analysis, with hundreds of genetic loci being identified. For autoimmune diseases, a consortium of scientists led by Wellcome Trust Sanger Institute developed the immunochip (Illumina 2015—Infinium ImmunoArray-24 V2.0 Beadchip) with more than 250,000 SNPs from 186 most significantly associated loci (Cortes and Brown 2011; Liu et al. 2012), mainly for 17 major autoimmune diseases including type 1 diabetes, celiac disease, inflammatory bowel disease, multiple sclerosis, ankylosing spondylitis, rheumatoid arthritis, vitiligo, and systemic lupus erythematosus (Illumina 2015). One of the advantages of this chip is the refinement of all loci with dense marker set. Refinement of non-HLA loci associated with these autoimmune diseases has opened up many new avenues of research as well.

More than 120 research projects on the above-mentioned diseases in many countries were carried out with refinement of many known loci. Immunochip data for celiac disease yielded 13 new disease susceptibility loci, total of which now stands at 40. This dense genotyping of key autoimmune disease loci resulted in refining the known and new loci to one causal gene for almost all (Trynka et al. 2011). This data also identified "credible set" of variants, one of which most likely to be a causal variant for the risk locus. Rapid refinement of known loci and identification of potential causal gene for risk loci for many immune diseases came through such customized SNP analysis as well as general SNP microarray. Bioinformatics analysis of the candidate loci for IBD played a key role in providing supporting evidence for causal genes and their effect on crucial pathways (Jostins et al. 2012).

Large-scale collaborations like Wellcome Trust Case Control consortium was formed in 2005 to harness the power of such new technologies and bioinformatics, focusing on 14 complex diseases (Table 2.1) for GWAS data generation with large samples from the UK. For seven core diseases (*bold letters* in Table 2.1), 2000 cases and 3000 controls samples were genotyped with 500,000 SNPs by Affymetrix microarray chip technology, and association results were published in multiple publications from the consortium (Wellcome Trust Case Control Consortium 2007; Wellcome Trust Case Control Consortium et al. 2007; Barrett et al. 2008; Barton et al. 2008; Holmans et al. 2009; Imielinski et al. 2009; Perry et al. 2009; Wellcome Trust Case Control Consortium, et al. 2010). Host resistance to infectious diseases (TB and malaria) in Africa was funded by Wellcome Trust and Bill & Melinda Gates Foundation under MalariaGEN Initiative, providing the initiative for fatal diseases of developing countries (Jallow et al. 2009). This study also followed similar experimental design. Other diseases were genotyped with custom chip of about 15,000 known non-synonymous SNPs across the majority of the genes in the genome with Illumina Infinium custom SNP chip (Wellcome Trust Case Control Consortium, et al. 2010; Grozeva et al. 2010).

This is one of the first collaborative efforts of many groups to generate large-scale data from sample collections for many common diseases to identify the genetic risk factors. Bioinformatics scientists realized their key role in working with such large-scale data and provided much needed tools and databases for querying many aspects of biological information to interpret them from variant and gene annotations to gene expression and protein interaction for the whole genome of multiple organisms. This development led scientists in different fields to come together and realize the potential of the GWAS data in the following years with many more novel genetic risk loci identified than in the last 30 years put together (The GWAS catalogue 2018).

Table 2.1 WTCC disease areas and control cohorts

Disease	Manufacturer
Type 1 diabetes Type 2 diabetes Crohn's disease Coronary heart disease Hypertension Bipolar disorder Rheumatoid arthritis	Affymetrix 500 K SNPs
Breast cancer Multiple sclerosis Ankylosing spondylitis Autoimmune thyroid disease	Illumina Infinium custom chip
Malaria Tuberculosis	Affymetrix 500 K SNPs
[a]**Control cohorts** 1958 Birth Cohort UK Blood Service	Affymetrix 500 K SNPs

[a]Diseases in **bold letters** are the core diseases**. Illumina custom SNP chip** contains 15,000 non-synonymous SNPs across the genome (From https://www.wtccc.org.uk/)

The GWAS challenge to the bioinformatics is not only how to store and handle large-scale data but to analyze them to bring novel discoveries to contribute to the welfare of the populations across the globe. Many of these tools are in the public domain, and information about how it can be used by bench scientists and students is also available (Shaik et al. 2019). Many groups of scientists applied such bioinformatics tools to GWAS data to identify the novel target genes, pathways for the disease, as well as novel functional effect of variants on the complex diseases (Eyre et al. 2012; Banaganapalli et al. 2017; Uenaka et al. 2018).

2.5 GWAS and Genetically Isolated Populations

Genetically isolated or homogeneous populations, due to their physical isolation from admixture or migration, will be a good example to study complex genetic diseases. In 1996, Professor Kári Stefánsson, neurologist, of Iceland recognized that concept and found the *deCODE* genetics company. This company recognized the value of unlocking the potential genetic contribution to complex diseases in a uniquely homogeneous population of Iceland and the excellent personal genealogical data from about the 1700s to date and healthcare records of the total population—about 350,000 in total. This company and the national government built the Icelander database, which has records of more than 95% of the population born after the 1700s. The largest genealogy of the world and the new high-throughput technologies in genetic analysis—genotyping by microarray, whole exome and whole genome sequences of the population—combined to provide the strongest support to unlock the genetic architecture of many complex diseases observed in that population over the last two centuries. This company aimed to generate a 500,000 SNP genotype record for all Icelanders and then test for association of the diseases in the population. This group has changed the GWAS landscape of many diseases from cardiovascular diseases to epilepsy and cancers. Genetic homogeneity of the participants helped their research by providing many potential candidate loci and genes. These targets were used as potential biomarkers for disease progress or as a novel drug target to alleviate the health complications of the diseases (The deCode publications: https://www.decode.com/publications/). In the early stages of their startup, focus was on identifying the causal mutations in monogenic diseases of Icelanders. Identification of TEAD as a causal gene for Sveinsson's chorioretinal atrophy was possible with the 14 generation family records is one such example (Fossdal et al. 2004). Many more disease genes were identified over the years. Access to the individual national health records from birth to death analyzed along with the genome-wide genotype data resulted in the identification of causal mutation for rare Mendelian disease like Sveinsson's chorioretinal atrophy and significantly associated disease susceptibility loci for many complex diseases including myocardial infarction and cancers like prostate cancer (Fossdal et al. 2004; Helgadottir et al. 2007; Gudmundsson et al. 2007). Biosample collection from patients with cancers and other diseases across the country for many years resulted in the discovery of major genetic contributions to cancers and tumors. The strength of the company data and core bioinformatics team along with scientists across many fields and hospitals in Iceland led the pharmaceutical company Amgen to buy it recently. Amgen use their data to develop novel drug target, to stratify population, and to utilize in personalized medicine strategy for multiple diseases in their drug portfolio. The extensive clinical, family, and genetic data (in the form of whole genome genotyping, whole exome sequencing, etc.) is reused for multiple targets with powerful bioinformatics tools within their facility. Similar bioinformatics tools were in the public domain which is exploiting the large-scale genotyping project data in many countries and many novel discoveries followed.

Golden era of bioinformatics growth is linked to the technological developments of large-scale data generating capacity. National Center for Biotechnology Information (NCBI) of USA and European Bioinformatics Institute (EBI) of Europe started storing the spectrum of data sets from sequence to variants and gene expression profile and functional annotations of genes and proteins in multitude of databases. Access to these databases by scientists, with limited exposure to computer programming skills or Unix commands across the world, was made easy by the development of many bioinformatics tools integrated within these organizations, as well as many independent bioinformatics groups across the world developed many easy-to-use web interfaces to query the specific databases for the information for the bench scientists and clinical scientists.

From NCBI BLAST, dbSNP query to the GTEx analysis tools, many other useful bioinformatics tools helped the scientists to expand the GWAS outcome beyond markers to understanding the role of genetic marker to gene function and disease biology. Excellent reviews and meta-analyses of integrative genetic and genomic data reveal the extent of the success of thousands of GWAS on many diseases across the world. The *GWAS catalogue* of EBI in collaboration with NHGRI (National Human Genome Research Institute) made it easy for biologists and geneticists by capturing 3720 published GWAS data sets to query. The GWAS collection, as of December 2018, has 89,680 SNP-trait association and 70,459 SNPs associated with many traits and diseases in the database. Recent studies on the impact of GWAS on publications of biologists reveal that a new gene associated with a disease or trait by GWAS gets more attention and more publication citation immediately after the GWAS publications than other genes with no genetic association support (Struck et al. 2018). The combined efforts of microarray technology, bioinformatic tools, and access to hundreds and thousands of clinical samples and data led to rapid increase in GWAS studies, which was less than 200 before 2005 and reached more than 3200 between 2010 and 2018, with increasing number of samples and markers for spectrum of diseases (The GWAS catalogue – web link 2018).

2.6 National Biobank and Genome Projects

Many countries recognized the importance of the genetics in healthcare and its impact on reducing the economic burden of genetic diseases on the national budgets (Table 2.2). Many industrialized nations like the UK, the USA, and China followed the footpath of Iceland by creating national biobanks. Realization of direct benefit of genetic revolution will take time. The "big data" opportunities spur rapid scientific discoveries of the complexities of many common diseases, which was unimaginable two decades ago. This positive step was taken up by many governments with ambitious goals set for the scientific teams to scale in the form of discovery and development of tools and drugs to treat patients with many diseases. Many countries have initiated the national biobank with genetic data linked to long-term health records of the nations. Such national data accessibility is restricted by government policies or the national committee overseeing the effort in few countries for now. The UK Biobank gives access to the data it holds to scientists, whose proposed work will be published and the analyzed data and results return to their organization. Other countries have various levels of access to their genetic data. One of the first successes of the national biobank is discussed below as an example.

The *UK Biobank* project started as the epidemiological study to address the risk factor identification for many diseases based on the long-term population. One of the longest ongoing studies for the last 40+ years like the Framingham Heart study in the USA of more than 5000 people. In the UK, scientists wanted to expand the likes of Framingham Study on a much larger scale (100-fold increase) to 500,000 people in all walks of life. The Wellcome Trust and Medical Research Council in the UK funded the initial recruitment and data gathering effort. More funds from these agencies and extensive collaboration among multiple groups led to the large-scale data gathering on 500,000 people (Bycroft et al. 2018). With the genome-wide genotyping and whole exome sequencing of these participants along with clinical

Table 2.2 Major national genomics biobanks

Country	Projects
USA	One million Veterans project All of Us Research Program—1 million
United Kingdom	100,000 genomes project
China	100,000 genomes project
Saudi Human Genome Program	100,000 genomes project
Dubai, United Arab Emirates	Dubai Genomics whole population—3 million
Estonia	Personalized Medicine Program—100,000
French Plan for Genomic Medicine 2025	100,000 genomes project
The Australian Genomics Health Futures Mission	100,000 genomes project
Japan Initiatives on Rare and Undiagnosed Diseases	2000

data form National Health Service (NHS), a large amount of other test results and surveys opened up the unique opportunity to explore the novel connections between multiple genetic markers and thousands of traits of interest. The UK Biobank data is open to the scientists around the world, when the genotype data and clinical data are in secure databases. The scientists need to pay UK 2500 Sterling pounds to access the 8 terabytes of data on 500,000 individuals with genotypes of millions of markers across the genome. The first release of the data resulted in at least 600 articles in leading journals in a variety of fields (Jansen et al. 2019), and 1400 researchers registered their projects with the UK Biobank and analyzed the data for multiple diseases and traits association. Exome sequence data of 500,000 people will be released in March 2019. This will open again the floodgate for researchers to carry out extraordinary large-scale secondary data analysis with exome data and reveal the novel discoveries to benefit the world. Many ethnicity-specific major contributing coding variants will be identified for common diseases, which will trigger the application of such variants and genes in biomarker development and novel drug target molecule search with in silico screening of the compound library as well as many other exciting applications.

Secondary analysis of large-scale data is possible with the powerful bioinformatics analysis pipeline along with the strong statistical power. The above-mentioned two large data sets are unique. The deCode company is a commercial venture with restricted access to their benefit, and the UK Biobank is the largest resource with no restriction for access to the data. Such large-scale data are also available in the USA. The two major projects such as Million people project of Veterans Administrations group and a commercial company (23andMe) collections are much larger but with access limited by the participants. Recent publications of the secondary data from these can be accessed from their own websites (UK Biobank—https://www.ukbiobank.ac.uk/).

Spin-off of such large-scale data analysis led to many novel discoveries, which in turn resulted in novel drug targets for many complex diseases and novel biomarkers for disease development and many related fields of biomedical sciences.

For bioinformatics scientists, it is a boon that they can design innumerable pipelines to analyze the diverse data from this source to reveal the role of novel genetic associations, novel pathways for a variety of diseases and traits recorded, and novel biomarkers for disease development which will spur the bench scientists to validate the novel associations with elegant experiments like gene editing, single cell gene expression, etc.

The UK Biobank data released to global scientists resulted in the discovery of many novel disease risk loci for devastating diseases with extensive meta-analysis of combined GWAS studies of many ethnic and racial groups. Such an endeavor will benefit not only the nation which provided the data but the world. In the future, large meta-analysis studies from biobank data of many countries will provide confirmed target of intervention by the development of novel drug molecules or disease-specific biomarkers which will help the healthcare professionals in prevention strategies for high-risk individuals. Meta-analysis of large-scale multiple GWAS data are being generated with more powerful statistical and bioinformatics tools and methods for many diseases.

Genetic association studies with a variety of marker types like blood group, RFLP, and microsatellites to SNPs drive the bioinformatics groups to develop various tools to analyze large-scale data to identify the disease contributing genes, novel drug targets, or biomarkers. Many new databases resulting from the interdisciplinary collaboration with bioinformatics were useful to suit a variety of biomedical fields. The Open Target platform (Koscielny et al. 2017; https://www.targetvalidation.org/) from Wellcome Sanger Institute and EMBL-EBI with the collaboration of pharmaceutical industry partners like GSK, Biogen, Sanofi, Takeda, and Celgene is one such specialized database which captures and annotates data from many biomedical high-throughput platforms in one place. This database provides the integrated robust data from a variety of fields on genes which can be searched for their suitability as a novel drug target for the disease of interest.

The genetic association of diseases drives the identification of the role of many genes in their development. Advances in the genome-wide genotyping methods, tools to analyze large-scale data in the post-human genome sequencing project era made it possible to identify thousands of loci and markers associated with susceptibility to different diseases. Databases and tools created to store, annotate, and visualize the results changed the genomic research immensely in the past decade. Advances in GWAS and the birth of national biobanks are going to play an important role in better management of diseases in the population in the future.

Major Web Links

National Center of Biotechnology Information (NCBI): https://www.ncbi.nlm.nih.gov/

European Bioinformatics Institute (EBI), an outstation of European Molecular Biology organization (EMBO) https://www.ebi.ac.uk/

The GWAS Catalogue: https://www.ebi.ac.uk/gwas/

UK Biobank: https://www.ukbiobank.ac.uk/

Million Veterans Program: https://www.research.va.gov/mvp/

All of Us Research Program, NIH, USA: https://allofus.nih.gov/
Saudi Genome Project: https://www.saudigenomeprogram.org/en/about-us/
The Open Target Platform, EBI: https://www.targetvalidation.org/

References

Banaganapalli B, Rashidi O, Saadah O, Wang J, Muhammed IA, Al-Aama JY, Shaik NA, Elango R (2017) Comprehensive computational analysis of GWAS loci identifies CCR2 as a candidate gene for celiac disease pathogenesis. J Cell Biochem 118(8):2193–2207. https://doi.org/10.1002/jcb.25864. PubMed PMID: 28059456

Barrett JC, Hansoul S, Nicolae DL, Cho JH, Duerr RH, Rioux JD, Brant SR, Silverberg MS, Taylor KD, Barmada MM, Bitton A, Dassopoulos T, Datta LW, Green T, Griffiths AM, Kistner EO, Murtha MT, Regueiro MD, Rotter JI, Schumm LP, Steinhart AH, Targan SR, Xavier RJ, Genetics Consortium NIDDKIBD, Libioulle C, Sandor C, Lathrop M, Belaiche J, Dewit O, Gut I, Heath S, Laukens D, Mni M, Rutgeerts P, Van Gossum A, Zelenika D, Franchimont D, Hugot JP, de Vos M, Vermeire S, Louis E, Belgian-French IBD Consortium, Wellcome Trust Case Control Consortium, Cardon LR, Anderson CA, Drummond H, Nimmo E, Ahmad T, Prescott NJ, Onnie CM, Fisher SA, Marchini J, Ghori J, Bumpstead S, Gwilliam R, Tremelling M, Deloukas P, Mansfield J, Jewell D, Satsangi J, Mathew CG, Parkes M, Georges M, Daly MJ (2008) Genome-wide association defines more than 30 distinct susceptibility loci for Crohn's disease. Nat Genet 40(8):955–962. PUBMED: 18587394; PMC: 2574810. https://doi.org/10.1038/ng.175

Barton A, Thomson W, Ke X, Eyre S, Hinks A, Bowes J, Plant D, Gibbons LJ, Wellcome Trust Case Control Consortium, YEAR Consortium, BIRAC Consortium, Wilson AG, Bax DE, Morgan AW, Emery P, Steer S, Hocking L, Reid DM, Wordsworth P, Harrison P, Worthington J (2008) Rheumatoid arthritis susceptibility loci at chromosomes 10p15, 12q13 and 22q13. Nat Genet 40(10):1156–1159. PM ID: 18794857; PMC: 2662493. https://doi.org/10.1038/ng.218

Brook JD, Harley HG, Walsh KV, Rundle SA, Siciliano MJ, Harper PS, Shaw DJ (1991) Identification of new DNA markers close to the myotonic dystrophy locus. J Med Genet 28(2):84–88

Bycroft C, Freeman C, Petkova D, Band G, Elliott LT, Sharp K, Motyer A, Vukcevic D, Delaneau O, O'Connell J, Cortes A, Welsh S, Young A, Effingham M, McVean G, Leslie S, Allen N, Donnelly P, Marchini J (2018) The UK biobank resource withdeep phenotyping and genomic data. Nature 562(7726):203–209. https://doi.org/10.1038/s41586-018-0579-z. PubMed PMID: 30305743

Cameron JM, Izatt MM (1962) The ABO and rhesus blood groups in Perthes' disease. J Clin Pathol 15:163–168. PMID: 13875943

Cortes A, Brown MA (2011) Promise and pitfalls of the Immunochip. Arthritis Res Ther 13(1):101. https://doi.org/10.1186/ar3204. PMID: 21345260

Dietrich WF, Miller J, Steen R, Merchant MA, Damron-Boles D, Husain Z, Dredge R, Daly MJ, Ingalls KA, O'Connor TJ (1996) A comprehensive genetic map of the mouse genome. Nature 380(6570):149–152. https://doi.org/10.1038/380149a0. Erratum in: Nature 381(6578):172. PMID: 8600386

Doucette-Stamm LA, Riba L, Handelin B, Difilippantonio M, Ward DC, Wasmuth JJ, Gusella JF, Housman DE (1991) Generation and characterization of irradiation hybrids of human chromosome 4. Somat Cell Mol Genet 17(5):471–480. PMID: 1837181

Elmgreen J, Sorensen H, Berkowicz A (1984) Polymorphism of complement C3 in chronic inflammatory bowel disease. Acta Med Scand 215:375–378

Eyre S, Bowes J, Diogo D, Lee A, Barton A, Martin P, Zhernakova A, Stahl E, Viatte S, McAllister K, Amos CI, Padyukov L, Toes RE, Huizinga TW, Wijmenga C, Trynka G, Franke L, Westra

HJ, Alfredsson L, Hu X, Sandor C, de Bakker PI, Davila S, Khor CC, Heng KK, Andrews R, Edkins S, Hunt SE, Langford C, Symmons D, Biologics in Rheumatoid Arthritis Genetics and Genomics Study Syndicate; Wellcome Trust Case Control Consortium, Concannon P, Onengut-Gumuscu S, Rich SS, Deloukas P, Gonzalez-Gay MA, Rodriguez-Rodriguez L, Ärlsetig L, Martin J, Rantapää-Dahlqvist S, Plenge RM, Raychaudhuri S, Klareskog L, Gregersen PK, Worthington J (2012) High-density genetic mapping identifies new susceptibility loci for rheumatoid arthritis. Nat Genet 44(12):1336–1340. https://doi.org/10.1038/ng.2462

Fossdal R, Jonasson F, Kristjansdottir GT, Kong A, Stefansson H, Gosh S, Gulcher JR, Stefansson K (2004) A novel TEAD1 mutation is the causative allele in Sveinsson's chorioretinal atrophy (helicoid peripapillary chorioretinal degeneration). Hum Mol Genet 13(9):975–981. https://doi.org/10.1093/hmg/ddh106

Grozeva D, Kirov G, Ivanov D, Jones IR, Jones L, Green EK, St Clair DM, Young AH, Ferrier N, Farmer AE, McGuffin P, Holmans PA, Owen MJ, O'Donovan MC, Craddock N, Wellcome Trust Case Control Consortium (2010) Rare copy number variants: a point of rarity in genetic risk for bipolar disorder and schizophrenia. Arch General Psychiatr 67(4):318–327. PM ID: 20368508; PMC: 4476027. https://doi.org/10.1001/archgenpsychiatry.2010.25

Gudmundsson J, Sulem P, Manolescu A, Amundadottir LT, Gudbjartsson D, Helgason A, Rafnar T, Bergthorsson JT, Agnarsson BA, Baker A, Sigurdsson A, Benediktsdottir KR, Jakobsdottir M, Xu J, Blondal T, Kostic J, Sun J, Ghosh S, Stacey SN, Mouy M, Saemundsdottir J, Backman VM, Kristjansson K, Tres A, Partin AW, Albers-Akkers MT, Godino-Ivan Marcos J, Walsh PC, Swinkels DW, Navarrete S, Isaacs SD, Aben KK, Graif T, Cashy J, Ruiz-Echarri M, Wiley KE, Suarez BK, Witjes JA, Frigge M, Ober C, Jonsson E, Einarsson GV, Mayordomo JI, Kiemeney LA, Isaacs WB, Catalona WJ, Barkardottir RB, Gulcher JR, Thorsteinsdottir U, Kong A, Stefansson K (2007) Genome-wide association study identifies a second prostate cancer susceptibility variant at 8q24. Nat Genet 39(5):631–637. https://doi.org/10.1038/ng1999

Gyapay G, Schmitt K, Fizames C, Jones H, Vega-Czarny N, Spillett D, Muselet D, Prud'homme JF, Dib C, Auffray C, Morissette J, Weissenbach J, Goodfellow PN (1996) A radiation hybrid map of the human genome. Hum Mol Genet 5(3):339–346

Harper PS, O'Brien T, Murray JM, Davies KE, Pearson P, Williamson R (1983) The use of linked DNA polymorphisms for genotype prediction in families with Duchenne muscular dystrophy. J Med Genet 20(4):252–254

Helgadottir A, Thorleifsson G, Manolescu A, Gretarsdottir S, Blondal T, Jonasdottir A, Jonasdottir A, Sigurdsson A, Baker A, Palsson A, Masson G, Gudbjartsson DF, Magnusson KP, Andersen K, Levey AI, Backman VM, Matthiasdottir S, Jonsdottir T, Palsson S, Einarsdottir H, Gunnarsdottir S, Gylfason A, Vaccarino V, Hooper WC, Reilly MP, Granger CB, Austin H, Rader DJ, Shah SH, Quyyumi AA, Gulcher JR, Thorgeirsson G, Thorsteinsdottir U, Kong A, Stefansson K (2007) A common variant on chromosome 9p21 affects the risk of myocardial infarction. Science 316(5830):1491–1493. https://doi.org/10.1126/science.1142842. Epub 2007 May 3. PubMed PMID: 17478679

Holmans P, Green EK, Pahwa JS, Ferreira MA, Purcell SM, Sklar P, Wellcome Trust Case-Control Consortium, Owen MJ, O'Donovan MC, Craddock N (2009) Gene ontology analysis of GWA study data sets provides insights into the biology of bipolar disorder. Am J Human Genet 85(1):13–24. https://doi.org/10.1016/j.ajhg.2009.05.011. PMID: 19539887; PMC: 2706963

Illumina (2015) Infininium Immunochip. https://www.illumina.com/content/dam/illumina-marketing/documents/products/product_information_sheets/infinium-human-immunoarray-24-product-info-sheet-370-2015-002.pdf

Imielinski M, Baldassano RN, Griffiths A, Russell RK, Annese V, Dubinsky M, Kugathasan S, Bradfield JP, Walters TD, Sleiman P, Kim CE, Muise A, Wang K, Glessner JT, Saeed S, Zhang H, Frackelton EC, Hou C, Flory JH, Otieno G, Chiavacci RM, Grundmeier R, Castro M, Latiano A, Dallapiccola B, Stempak J, Abrams DJ, Taylor K, McGovern D, Western Regional Alliance for Pediatric IBD, Silber G, Wrobel I, Quiros A, International IBD Genetics Consortium, Barrett JC, Hansoul S, Nicolae DL, Cho JH, Duerr RH, Rioux JD, Brant SR, Silverberg MS, Taylor KD, Barmuda MM, Bitton A, Dassopoulos T, Datta LW, Green T, Griffiths AM, Kistner EO, Murtha MT, Regueiro MD, Rotter JI, Schumm LP, Steinhart AH, Targan SR, Xavier RJ, Consortium NIDDKIBDG, Libioulle C, Sandor C, Lathrop M, Belaiche J, Dewit O, Gut I, Heath S, Laukens D, Mni M, Rutgeerts P, Van Gossum A, Zelenika D, Franchimont D, Hugot JP, de Vos M, Vermeire S, Louis E, Belgian-French IBD Consortium, Wellcome Trust Case Control Consortium, Cardon LR, Anderson CA, Drummond H, Nimmo E, Ahmad T, Prescott NJ, Onnie CM, Fisher SA, Marchini J, Ghori J, Bumpstead S, Gwillam R, Tremelling M, Delukas P, Mansfield J, Jewell D, Satsangi J, Mathew CG, Parkes M, Georges M, Daly MJ, Heyman MB, Ferry GD, Kirschner B, Lee J, Essers J, Grand R, Stephens M, Levine A, Piccoli D, Van Limbergen J, Cucchiara S, Monos DS, Guthery SL, Denson L, Wilson DC, Grant SF, Daly M, Silverberg MS, Satsangi J, Hakonarson H (2009) Common variants at five new loci associated with early-onset inflammatory bowel disease. Nat Genet 41(12):1335–1340. https://doi.org/10.1038/ng.489. PM ID: 19915574; PMC: 3267927

Jallow M, Teo YY, Small KS, Rockett KA, Deloukas P, Clark TG, Kivinen K, Bojang KA, Conway DJ, Pinder M, Sirugo G, Sisay-Joof F, Usen S, Auburn S, Bumpstead SJ, Campino S, Coffey A, Dunham A, Fry AE, Green A, Gwilliam R, Hunt SE, Inouye M, Jeffreys AE, Mendy A, Palotie A, Potter S, Ragoussis J, Rogers J, Rowlands K, Somaskantharajah E, Whittaker P, Widden C, Donnelly P, Howie B, Marchini J, Morris A, SanJoaquin M, Achidi EA, Agbenyega T, Allen A, Amodu O, Corran P, Djimde A, Dolo A, Doumbo OK, Drakeley C, Dunstan S, Evans J, Farrar J, Fernando D, Hien TT, Horstmann RD, Ibrahim M, Karunaweera N, Kokwaro G, Koram KA, Lemnge M, Makani J, Marsh K, Michon P, Modiano D, Molyneux ME, Mueller I, Parker M, Peshu N, Plowe CV, Puijalon O, Reeder J, Reyburn H, Riley EM, Sakuntabhai A, Singhasivanon P, Sirima S, Tall A, Taylor TE, Thera M, Troye-Blomberg M, Williams TN, Wilson M, Kwiatkowski DP, Wellcome Trust Case Control Consortium and Malaria Genomic Epidemiology Network (2009) Genome-wide and fine-resolution association analysis of malaria in West Africa. Nat Genet 41(6):657–665. PM ID: 19465909; PMC: 2889040. https://doi.org/10.1038/ng.388

Jansen IE et al (2019) Genome-wide meta-analysis identifies new loci and functional pathways influencing Alzheimer's disease risk. Nat Genet (Pre-publication). https://doi.org/10.1038/s41588-018-0311-9

Jostins L et al (2012) Host–microbe interactions have shaped the genetic architecture of inflammatory bowel disease. Nature 491(7422):119–124. https://doi.org/10.1038/nature11582

Kan YW, Dozy AM (1978) Antenatal diagnosis of sickle-cell anaemia by D.N.A. analysis of amniotic-fluid cells. Lancet 2(8096):910–912. PMID: 81926

Kan YW, Lee KY, Furbetta M, Angius A, Cao A (1980) Polymorphism of DNA sequence in the beta-globin gene region. Application to prenatal diagnosis of beta 0 thalassemia in Sardinia. N Engl J Med 302(4):185–188. PMID: 6927915

Koscielny G, An P, Carvalho-Silva D, Cham JA, Fumis L, Gasparyan R, Hasan S, Karamanis N, Maguire M, Papa E, Pierleoni A, Pignatelli M, Platt T, Rowland F, Wankar P, Bento AP, Burdett T, Fabregat A, Forbes S, Gaulton A, Gonzalez CY, Hermjakob H, Hersey A, Jupe S, Kafkas Ş, Keays M, Leroy C, Lopez FJ, Magarinos MP, Malone J, McEntyre J, Munoz-Pomer Fuentes A, O'Donovan C, Papatheodorou I, Parkinson H, Palka B, Paschall J, Petryszak R, Pratanwanich N, Sarntivijal S, Saunders G, Sidiropoulos K, Smith T, Sondka Z, Stegle O, Tang YA, Turner E, Vaughan B, Vrousgou O, Watkins X, Martin MJ, Sanseau P, Vamathevan J, Birney E, Barrett J, Dunham I (2017) Open targets: a platform for therapeutic target identification and validation. Nucleic Acids Res 45(D1):D985–D994. https://doi.org/10.1093/nar/gkw1055

Liu JZ, Almarri MA, Gaffney DJ, Mells GF, Jostins L, Cordell HJ, Ducker SJ, Day DB, Heneghan MA, Neuberger JM, Donaldson PT, Bathgate AJ, Burroughs A, Davies MH, Jones DE, Alexander GJ, Barrett JC, Sandford RN, Anderson CA, UK Primary Biliary Cirrhosis (PBC) Consortium; Wellcome Trust Case Control Consortium 3 (2012) Dense fine-mapping study identifies new susceptibility loci for primary biliary cirrhosis. Nat Genet 44(10):1137–1141. https://doi.org/10.1038/ng.2395. PMID: 22961000

Mcginniss MH, Schmidt PJ, Carbone PP (1964) Close association of I blood group and disease. Nature 202:606. PMID: 14195074

Nussbaum RL, Crowder WE, Nyhan WL, Caskey CT (1983) A three-allele restriction-fragment-length polymorphism at the hypoxanthine phosphoribosyltransferase locus in man. Proc Natl Acad Sci U S A 80(13):4035–4039

O'Hanlon RH, Stewert FS (1948) Maternal jaundice in association with haemolytic disease due to Rh sensitization. Ir J Med Sci (268):178. PMID: 18862316

Patel R, Mickey MR, Terasaki PI (1969) Leucocyte antigens and disease. I. Association of HL-A2 and chronic glomerulonephritis. Br Med J 2(5654):424–426. PMID: 5781489

Perry JR, McCarthy MI, Hattersley AT, Zeggini E, Wellcome Trust Case Control Consortium, Weedon MN, Frayling TM (2009) Interrogating type 2 diabetes genome-wide association data using a biological pathway-based approach. Diabetes 58(6):1463–1467. https://doi.org/10.2337/db08-1378. PM ID: 19252133; PMC: 2682674

Prest E, Bonnin JA, Simmons RT, Newland BT (1955) Haemolytic disease of the newborn due to ABO isosensitization in association with potent anti-M agglutinins. Med J Aust 42(5):153–156. PMID: 13253116

Pulver AE, Karayiorgou M, Wolyniec PS, Lasseter VK, Kasch L, Nestadt G, Antonarakis S, Housman D, Kazazian HH, Meyers D, Ott J, Lamacz M, Liang K-Y, Hanfelt J, Ullrich G, Demarchi N, Ramu E, Gordon CT, Kimberland M, Babb R, Puck J, Childs B (1994) A sequential strategy to identify a susceptibility gene for schizophrenia: report of potential linkage on chromosome 22q12-q13.1. Am J Med Genet 54:36–43

Sarfarazi M, Harper PS, Kingston HM, Murray JM, O'Brien T, Davies KE, Williamson R, Tippett P, Sanger R (1983) Genetic linkage relationship between the Xg blood group system and two X chromosome DNA polymorphisms in families with Duchenne and Becker muscular dystrophy. Hum Genet 65(2):169–171

Schuler GD, Boguski MS, Stewart EA, Stein LD, Gyapay G, Rice K, White RE, Rodriguez-Tomé P, Aggarwal A, Bajorek E, Bentolila S, Birren BB, Butler A, Castle AB, Chiannilkulchai N, Chu A, Clee C, Cowles S, Day PJ, Dibling T, Drouot N, Dunham I, Duprat S, East C, Edwards C, Fan JB, Fang N, Fizames C, Garrett C, Green L, Hadley D, Harris M, Harrison P, Brady S, Hicks A, Holloway E, Hui L, Hussain S, Louis-Dit-Sully C, Ma J, MacGilvery A, Mader C, Maratukulam A, Matise TC, McKusick KB, Morissette J, Mungall A, Muselet D, Nusbaum HC, Page DC, Peck A, Perkins S, Piercy M, Qin F, Quackenbush J, Ranby S, Reif T, Rozen S, Sanders C, She X, Silva J, Slonim DK, Soderlund C, Sun WL, Tabar P, Thangarajah T, Vega-Czarny N, Vollrath D, Voyticky S, Wilmer T, Wu X, Adams MD, Auffray C, Walter NA, Brandon R, Dehejia A, Goodfellow PN, Houlgatte R, Hudson JR Jr, Ide SE, Iorio KR, Lee WY, Seki N, Nagase T, Ishikawa K, Nomura N, Phillips C, Polymeropoulos MH, Sandusky M, Schmitt K, Berry R, Swanson K, Torres R, Venter JC, Sikela JM, Beckmann JS, Weissenbach J, Myers RM, Cox DR, James MR, Bentley D, Deloukas P, Lander ES, Hudson TJ (1996). A gene map of the human genome. Science 274(5287):540–546. Review PubMed PMID: 8849440.

Shaik NA, Banaganapalli B, Elango R, Hakkeem KR (2019). Essentials of bioinformatics volume 1: understanding bioinformatics: genes to proteins. Springer Publishers, 6330 Cham, Switzerland.

Simon J, Barcal R, Sova J, Kulich V (1971) Heredity of hypertensive disease from the view of the prevalence in sibs and parents and association with some blood and serum groups. Acta Univ Carol Med (Praha) 17(3):227–274. PMID: 5146158.

Struck TJ, Mannakee BK, Gutenkunst RN (2018) The impact of genome-wide association studies on biomedical research publications. Hum Genomics 12:38–45. https://doi.org/10.1186/s40246-018-0172-4

Trynka G, Hunt KA, Bockett NA, Romanos J, Mistry V, Szperl A, Bakker SF, Bardella MT, Bhaw-Rosun L, Castillejo G, de la Concha EG, de Almeida RC, Dias KR, van Diemen CC, Dubois PC, Duerr RH, Edkins S, Franke L, Fransen K, Gutierrez J, Heap GA, Hrdlickova B, Hunt S, Plaza Izurieta L, Izzo V, Joosten LA, Langford C, Mazzilli MC, Mein CA, Midah V, Mitrovic M, Mora B, Morelli M, Nutland S, Núñez C, Onengut-Gumuscu S, Pearce K, Platteel M, Polanco I, Potter S, Ribes-Koninckx C, Ricaño-Ponce I, Rich SS, Rybak A, Santiago JL, Senapati S, Sood A, Szajewska H, Troncone R, Varadé J, Wallace C, Wolters VM, Zhernakova A; Spanish Consortium on the Genetics of Coeliac Disease (CEGEC);

PreventCD Study Group; Wellcome Trust Case Control Consortium (WTCCC), Thelma BK, Cukrowska B, Urcelay E, Bilbao JR, Mearin ML, Barisani D, Barrett JC, Plagnol V, Deloukas P, Wijmenga C, van Heel DA (2011) Dense genotyping identifies and localizes multiple common and rare variant association signals in celiac disease. Nat Genet 43:1193–1201. https://doi.org/10.1038/ng.998

Uenaka T, Satake W, Cha P-C, Hayakawa H, Baba K, Jiang S, Kobayashi K, Kanagawa M, Okada Y, Mochizuki H and Toda T (2018). In silico drug screening by using genome-wide association study data repurposed dabrafenib, an anti-melanoma drug, for Parkinson's disease. Hum Mol Gen. Advance Access Publication. https://doi.org/10.1093/hmg/ddy279

Wainscoat JS, Bell JI, Old JM, Weatherall DJ, Furbetta M, Galanello R, Cao A (1983) Globin gene mapping studies in Sardinian patients homozygous for beta zero Thalassaemia. Mol Biol Med 1(1):1–10. PMID: 6092822

Wellcome Trust Case Control Consortium, Australo-Anglo-American Spondylitis Consortium (TASC), Burton PR, Clayton DG, Cardon LR, Craddock N, Deloukas P, Duncanson A, Kwiatkowski DP, McCarthy MI, Ouwehand WH, Samani NJ, Todd JA, Donnelly P, Barrett JC, Davison D, Easton D, Evans DM, Leung HT, Marchini JL, Morris AP, Spencer CC, Tobin MD, Attwood AP, Boorman JP, Cant B, Everson U, Hussey JM, Jolley JD, Knight AS, Koch K, Meech E, Nutland S, Prowse CV, Stevens HE, Taylor NC, Walters GR, Walker NM, Watkins NA, Winzer T, Jones RW, McArdle WL, Ring SM, Strachan DP, Pembrey M, Breen G, St Clair D, Caesar S, Gordon-Smith K, Jones L, Fraser C, Green EK, Grozeva D, Hamshere ML, Holmans PA, Jones IR, Kirov G, Moskivina V, Nikolov I, O'Donovan MC, Owen MJ, Collier DA, Elkin A, Farmer A, Williamson R, McGuffin P, Young AH, Ferrier IN, Ball SG, Balmforth AJ, Barrett JH, Bishop TD, Iles MM, Maqbool A, Yuldasheva N, Hall AS, Braund PS, Dixon RJ, Mangino M, Stevens S, Thompson JR, Bredin F, Tremelling M, Parkes M, Drummond H, Lees CW, Nimmo ER, Satsangi J, Fisher SA, Forbes A, Lewis CM, Onnie CM, Prescott NJ, Sanderson J, Matthew CG, Barbour J, Mohiuddin MK, Todhunter CE, Mansfield JC, Ahmad T, Cummings FR, Jewell DP, Webster J, Brown MJ, Lathrop MG, Connell J, Dominiczak A, Marcano CA, Burke B, Dobson R, Gungadoo J, Lee KL, Munroe PB, Newhouse SJ, Onipinla A, Wallace C, Xue M, Caulfield M, Farrall M, Barton A, Biologics in RA Genetics and Genomics Study Syndicate (BRAGGS) Steering Committee, Bruce IN, Donovan H, Eyre S, Gilbert PD, Hilder SL, Hinks AM, John SL, Potter C, Silman AJ, Symmons DP, Thomson W, Worthington J, Dunger DB, Widmer B, Frayling TM, Freathy RM, Lango H, Perry JR, Shields BM, Weedon MN, Hattersley AT, Hitman GA, Walker M, Elliott KS, Groves CJ, Lindgren CM, Rayner NW, Timpson NJ, Zeggini E, Newport M, Sirugo G, Lyons E, Vannberg F, Hill AV, Bradbury LA, Farrar C, Pointon JJ, Wordsworth P, Brown MA, Franklyn JA, Heward JM, Simmonds MJ, Gough SC, Seal S, Breast Cancer Susceptibility Collaboration (UK), Stratton MR, Rahman N, Ban M, Goris A, Sawcer SJ, Compston A, Conway D, Jallow M, Newport M, Sirugo G, Rockett KA, Bumpstead SJ, Chaney A, Downes K, Ghori MJ, Gwilliam R, Hunt SE, Inouye M, Keniry A, King E, McGinnis R, Potter S, Ravindrarajah R, Whittaker P, Widden C, Withers D, Cardin NJ, Davison D, Ferreira T, Pereira-Gale J, Hallgrimsdo'ttir IB, Howie BN, Su Z, Teo YY, Vukcevic D, Bentley D, Brown MA, Compston A, Farrall M, Hall AS, Hattersley AT, Hill AV, Parkes M, Pembrey M, Stratton MR, Mitchell SL, Newby PR, Brand OJ, Carr-Smith J, Pearce SH, McGinnis R, Keniry A, Deloukas P, Reveille JD, Zhou X, Sims AM, Dowling A, Taylor J, Doan T, Davis JC, Savage L, Ward MM, Learch TL, Weisman MH, Brown M (2007) Association scan of 14,500 nonsynonymous SNPs in four diseases identifies autoimmunity variants. Nat Genet 39(11):1329–1337. https://doi.org/10.1038/ng.2007.17. PM ID: 17952073; PMC: 2680141

Wellcome Trust Case Control Consortium (2007) Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. Nature 447(7145):661–678. PUBMED: 17554300; PMC: 2719288. https://doi.org/10.1038/nature05911

Wellcome Trust Case Control Consortium, Craddock N, Hurles ME, Cardin N, Pearson RD, Plagnol V, Robson S, Vukcevic D, Barnes C, Conrad DF, Giannoulatou E, Holmes C, Marchini JL, Stirrups K, Tobin MD, Wain LV, Yau C, Aerts J, Ahmad T, Andrews TD, Arbury H, Attwood

A, Auton A, Ball SG, Balmforth AJ, Barrett JC, Barroso I, Barton A, Bennett AJ, Bhaskar S, Blaszczyk K, Bowes J, Brand OJ, Braund PS, Bredin F, Breen G, Brown MJ, Bruce IN, Bull J, Burren OS, Burton J, Byrnes J, Caesar S, Clee CM, Coffey AJ, Connell JM, Cooper JD, Dominiczak AF, Downes K, Drummond HE, Dudakia D, Dunham A, Ebbs B, Eccles D, Edkins S, Edwards C, Elliot A, Emery P, Evans DM, Evans G, Eyre S, Farmer A, Ferrier IN, Feuk L, Fitzgerald T, Flynn E, Forbes A, Forty L, Franklyn JA, Freathy RM, Gibbs P, Gilbert P, Gokumen O, Gordon-Smith K, Gray E, Green E, Groves CJ, Grozeva D, Gwilliam R, Hall A, Hammond N, Hardy M, Harrison P, Hassanali N, Hebaishi H, Hines S, Hinks A, Hitman GA, Hocking L, Howard E, Howard P, Howson JM, Hughes D, Hunt S, Isaacs JD, Jain M, Jewell DP, Johnson T, Jolley JD, Jones IR, Jones LA, Kirov G, Langford CF, Lango-Allen H, Lathrop GM, Lee J, Lee KL, Lees C, Lewis K, Lindgren CM, Maisuria-Armer M, Maller J, Mansfield J, Martin P, Massey DC, WL MA, McGuffin P, McLay KE, Mentzer A, Mimmack ML, Morgan AE, Morris AP, Mowat C, Myers S, Newman W, Nimmo ER, O'Donovan MC, Onipinla A, Onyiah I, Ovington NR, Owen MJ, Palin K, Parnell K, Pernet D, Perry JR, Phillips A, Pinto D, Prescott NJ, Prokopenko I, Quail MA, Rafelt S, Rayner NW, Redon R, Reid DM, Renwick RSM, Robertson N, Russell E, St Clair D, Sambrook JG, Sanderson JD, Schuilenburg H, Scott CE, Scott R, Seal S, Shaw-Hawkins S, Shields BM, Simmonds MJ, Smyth DJ, Somaskantharajah E, Spanova K, Steer S, Stephens J, Stevens HE, Stone MA, Su Z, Symmons DP, Thompson JR, Thomson W, Travers ME, Turnbull C, Valsesia A, Walker M, Walker NM, Wallace C, Warren-Perry M, Watkins NA, Webster J, Weedon MN, Wilson AG, Woodburn M, Wordsworth BP, Young AH, Zeggini E, Carter NP, Frayling TM, Lee C, McVean G, Munroe PB, Palotie A, Sawcer SJ, Scherer SW, Strachan DP, Tyler-Smith C, Brown MA, Burton PR, Caulfield MJ, Compston A, Farrall M, Gough SC, Hall AS, Hattersley AT, Hill AV, Mathew CG, Pembrey M, Satsangi J, Stratton MR, Worthington J, Deloukas P, Duncanson A, Kwiatkowski DP, MI MC, Ouwehand W, Parkes M, Rahman N, Todd JA, Samani NJ, Donnelly P (2010) Genome-wide association study of CNVs in 16,000 cases of eight common diseases and 3,000 shared controls. Nature 464(7289):713–720. https://doi.org/10.1038/nature08979. PUBMED: 20360734; PMC: 2892339

Chapter 3
Diagnostic Revolution Post-Human Genome Sequence Project: High-Throughput Technologies and Bioinformatics

Noor Ahmad Shaik, Babajan Banaganapalli, Jumana Y. Al-Aama, and Ramu Elango

Contents

N. A. Shaik · B. Banaganapalli · J. Y. Al-Aama · R. Elango (✉)
Princess Al-Jawhara Center of Excellence in Research of Hereditary Disorders,
Department of Genetic Medicine, Faculty of Medicine,
King Abdulaziz University, Jeddah, Saudi Arabia
e-mail: nshaik@kau.edu.sa; bbabajan@kau.edu.sa; jalama@kau.edu.sa; relango@kau.edu.sa

N. A. Shaik et al. (eds.), *Essentials of Bioinformatics, Volume II*,
https://doi.org/10.1007/978-3-030-18375-2_3

3.1 Introduction

Development of novel high-throughput technologies has lasting impact on patient care at various stages. Along with such development come the bigger challenges of analysing the large-scale data. Necessity is the mother of invention is proven true here as well. Hundreds of scientists across multiple disciplines came across to address this big challenge. Such collaboration and overlapping interest resulted in many bioinformatics tools to address various aspects of the complex nature of the genome architecture and to predict functional effect of changes observed in patient samples. Before Human Genome Sequencing Project started in the 1990s, around 1000 genetic diseases were diagnosed. Midway through the Human Genome program, spurt of novel high-throughput laboratory techniques and methods were being developed to exploit the sequence data for a variety of purposes—from genetic diagnosis, novel drug target identification, prevention of adverse reaction to drugs, to personalized medicine and to unravel the secrets of the genome architecture, functional genomics and more.

3.2 Genetic Diagnostics Before Human Genome Sequence Project

Diagnosis of the inherited diseases started with chromosomal abnormalities from trisomy (Down syndrome, etc.) and monosomy (Turner syndrome, etc.) with the karyotyping of blood lymphocyte chromosomes. With new technologies like FISH (fluorescent in situ hybridization) and SKY (spectral karyotyping of chromosomes), more complex chromosomal microdeletion, insertion, and translocation abnormalities were detected, especially in congenital malformations, rare syndromes and cancer cells. Limitation of these technologies include failure to identify genetic defects at nucleotide level; Sanger sequencing of the DNA and RNA addressed this issue at an early stage.

Genetic diagnosis, before the human genome sequences started pouring into the GenBank and other sequence databases across many countries, was restricted to few genotyping methods like Sanger sequencing, SSCP (single-strand conformation polymorphism), RFLP (restriction fragment length polymorphisms), etc. These techniques helped scientists across the globe to identify causative genes for many genetic diseases from thalassemia, haemophilia to more devastating diseases like myotonic dystrophy and Huntington disease. Many such monogenic disease-causative genes were revealed in many laboratories. The process of identification of genetic defect in these diseases was hindered by lack of genomic sequences with limited technology available. With only sequencing of 350–600 bases at a time was possible, identification of specific genetic change in a gene requires large technical manpower and financial resources. The gene identification involves extensive use of lab resources to map the locus through linkage analysis in one or many families with multiple affected individuals. For linkage analysis, genotyping of highly polymorphic markers (microsatellite CA repeat markers) and restriction fragment length polymorphism (RFLP) markers spread across the genome, about 200–500 markers, was carried out by a large team of workers were generated. Segregation analysis of inheritance pattern of the disease with the marker through statistical analysis revealed markers which are tightly linked to the disease in independent families (Linkage program). These linked markers are the starting point for the causal gene search. Radiation hybrid mapping and subcloning the region of interest in bacteria and identification of unique RFLP or other markers to narrow down the candidate regions took place in the laboratories. Such "chromosome walking" is slow and labour-intensive exercise. This laborious process of narrowing down the candidate region from about 5–10 Mb to the defective gene forced many research groups to address the technological bottleneck issues of the disease gene cloning by increasing the collaboration across many fields.

For example, to identify the Huntington disease gene mutation, teams from across the globe worked many years. The first step in identifying the genetic defect involves identification of families with multiple affected individuals from remote villages in Venezuela. Next breakthrough came from Jim Gusella and his collaborators who identified a marker linked to the Huntington disease to chromosome 4 (Gusella et al. 1983). This major breakthrough brought the strong international collaborations for linkage studies for many single-gene disorders with many RFLP markers across the genome. After a decade of hard work of multiple groups of scientists across the world, genetic defect in Huntington disease was revealed (The Huntington's Disease Collaborative Research Group 1993). This example gives an overview of the hard work of large collaborating scientists at that time in identifying the genetic defects for a variety of diseases.

Many such hard-working groups of scientists discovered many mutations causing some of the common forms of genetic diseases, with rapidly changing technologies from different parts of the globe. These discoveries resulted in developing targeted screening methods for genetic defect in clinically diagnosed patients. Many diagnostic companies exploited the new information and started focusing on developing diagnostic tools and probes for diagnosis of the affected as well as for prenatal diagnosis in high risk pregnancies. This targeted mutation detection is a slow process when the genetic defects for the disease vary in different patients. This required sequencing of the coding regions of the full gene or, in many cases, different genes, like in Parkinson's disease. Genetic defects in many Mendelian diseases were identified through this approach till the birth of the Human Genome Project. McKusick and his colleagues collated the Mendelian diseases-related information from around the world and kept the catalogue—OMIM (Online Mendelian Inheritance in Man)—initially in the book format. With advancing technologies and gene discoveries, they have moved the content to digital version, which lists more than 4000 Mendelian diseases with causal gene defects. It is being updated at a regular interval with more information and used as reference material for geneticists for accurate diagnosis and treatment for the affected children under their care.

Many genetics groups generated large amount of sequences from the regions of interest to the diseases they were working. Repository of such sequences for the common use and to avoid the waste of resources is created by NIH funds as GenBank by NCBI (National Center for Biotechnology Information). Searching this growing database requires computational tool. Slowly the high-throughput genotyping and sequencing technologies entered the laboratories which resulted in the data analysis bottleneck issue. This issue was taken up by biologists with interest in computers and specialists across many fields including the mathematicians and engineers. Examples of the early bioinformatics tools which still are used extensively by biologists are FASTA (FAST All) and BLAST (Basic Local Alignment Search Tool) (Lipman and Pearson 1985; Altschul et al. 1990). FASTP (FAST-Protein) program searches local protein sequence alignment between query and reference database using the Smith-Waterman algorithm (Smith and Waterman 1981). Later, modification of the FASTP program to include searches for nucleotides as well resulted in FASTA (FAST All) program (Pearson and Lipman 1988). These two tools (FASTA and BLAST) provide the foundation for the birth of bioinformatics in the analysis of biological sequences. These tools contributed in the rapid mapping of query sequences, identification of gene mutations, etc. With more sequences deposited to the GenBank, many scientists were developing tools to address the challenging questions regarding the prediction of the role of the mutation in gene function and to understand the biology of the disease. This in turn contributed to biological function of many genes, which allowed scientists to explore larger challenges of developing drugs for common diseases based on the functional contribution of the gene in preventing or controlling the disease effect on the patient.

3.3 Human Genome Revolution

In the 1990s, the Human Genome Project was initiated with the support of many government funding agencies like NIH in the USA and non-governmental research funding agencies like Wellcome Trust in the UK. Different groups focused on sequencing different chromosomes. The yeast artificial chromosome (YAC) libraries and bacterial artificial chromosome (BAC) libraries of the human chromosomes played a crucial role in kick-starting the project in many countries. Random sequencing of YAC and BAC clones required better bioinformatics tools. Such necessity pushed computer scientists and data analytic teams to expand the spectrum of tools to handle such large-scale data. NCBI (sequence analysis suite and GenBank) and UCSD (genome browser) and Sanger Institute (Ensembl) developed independently sequence visualization tools, which allowed scientists from different corners of the world free access to the data when it was released daily. This contributed to the rapid mapping of sequences to different regions of the genome. Many groups working on various diseases which were mapped to chromosomes of their interest used the data to rapidly identify mutations in genes.

At the same time, rapid-sequencing technologies were the focus of many companies and academic groups. Genome Analyzer from Applied Biosystems started the revolution, which saw birth of many high-throughput technologies in the coming years, from Illumina's next-generation sequencing (NGS) range, Ion Torrent platform, PacBio platform and, recently, a portable Nanopore technology platform as well. With these NGS platforms, sequence data started pouring into many international databases. Analysis of such large-scale data pushed many scientific groups with varied interests to collaborate to develop many tools to analyse various aspects of the genome data, from orderly mapping of sequence, identification of polymorphic markers, genotyping these markers in families with monogenic disease to large number of sporadic cases of polygenic diseases along with statisticians and IT scientists.

3.4 Post-Human Genome Revolution

By 2005, more than 90–95% of the human genome is sequenced and mapped to correct locations with the exception of few regions. Freely available human genome sequence and analysis tools spurred the identification of mutations in many monogenic diseases by sequencing familial cases collected across the world. Mutations in many rare diseases are regularly identified even by small research groups, thanks to the NGS technologies and bioinformatics tools to analyse such data. Rare diseases programs in the USA, UK and Europe encouraged many clinical teams from other parts of the world to share their samples to unravel the novel mutations in novel genes for many diseases which are restricted to few families in a region or country. Presently, more than 5000 monogenic disease mutations were identified in familial cases and many from families with rare diseases. Middle Eastern countries provided large number of rare disease families due to the high rates of consanguineous marriages (first cousin marriages are common here). Collaborating with many international research groups, scientists from these countries revealed the complex nature of many different diseases, which will be the foundation stone for functional genomics of many unknown genes identified. More than 700 such novel gene mutations were identified from these regions alone. Now, the application of NGS method to identify the causal mutation for rare diseases is routinely used in many laboratories and hospitals across the world.

3.5 Next-Generation Sequencing Diagnosis

The SNP microarray chip and NGS technologies are being extensively used to diagnose rare familial diseases much more easily. Previously, the rare disease diagnosis is through probable candidate gene screen or genetic linkage in families with more than two affected patients. With the human genome sequence readily available for comparison, NGS diagnosis is accurate. International and national collaborations in the USA, Canada, European countries and Asia led to the large-scale screening for mutations in rare diseases at a unprecedented level (NIH rare disease projects-Texas, Yale group, DDD- Sanger, European, etc.).

With the NGS technology and the bioinformatics tools (freely available), the diagnosis is quicker than ever. The NGS technology is getting ultra-high throughput with faster sequencing with better genome coverage. One of the best examples of the dramatic changes in the diagnosis of one group of diseases includes primary ciliary dyskinesia (PCD). The PCD is a rare disease with variable clinical features in young children, who suffer from multiple organ functional defect due to the defective cilia, especially the lung, heart, kidney and other organs. First genetic defect was identified for PCD in 1999, when no NGS technology was available (Pennarun et al. 1999). With the availability of the human reference genome sequence through Human Genome Project and NGS technology, about 39 more genes causing PCD were identified, so far from many different racial/ethnic groups in many countries. The NGS technology, easy to use bioinformatic tools and large-scale exome and genome sequence in ExAC, 1000 Genome project, ESP6500 and other national genome projects of many countries resulted in identification of hundreds of novel gene mutations for many rare inherited diseases from many parts of the world. Rapid diagnostic screening for many genetic diseases can be carried out for the newborn, using the targeted gene panel NGS, where the targeted regions were sequenced 100s of times and the mutations are recorded.

Diagnostics of polygenic and complex diseases: The development of NGS gene panel for many diseases, which are caused by one single gene or multiple genes, is a boon to the clinical community. These gene panels detect accurately any type of mutations in the coding part of the gene, whether it is novel or known ones. Other methods will screen for only known mutations. For example, the cystic fibrosis (CF) is one of the most common diseases. Hundreds and thousands of mutations in the CFTR gene are found to be causing CF in affected patients in many parts of the world. The gene has 27 exons spanning 188,702 bases in chromosome 7. The CFTR protein transcript length is 6132 base pairs, coding for 1480 amino acids. So far, hundreds of mutations have been reported in this protein.

Rare Disease Diagnostics: With the advent of NGS technologies, especially the whole exome sequencing (WES), hundreds of rare diseases, seen in single family, revealed novel gene mutations causing a spectrum of defects. Many diagnostic companies like Centogene and Invitae as well as specialist NGS technology companies like BGI and Macrogen provided easier access to clinicians to diagnose genetic defects for rare diseases and possibly the carrier and prenatal screening for these mutations in those families. National and international rare disease consortiums pooled their limited resources and initiated diagnosis of many rare diseases, like the Deciphering Developmental disorders (DDD: https://decipher.sanger.ac.uk/ddd#overview) program based at Wellcome Sanger Institute with multitude of teams across the UK and other countries, resulting in DECIPHER (Firth et al. 2009; Deciphering Developmental Disorders Study 2015; https://decipher.sanger.ac.uk/)

platform, where clinicians can share and compare genotype and phenotype data with the 28,863 patients data in the DDD program. More than 1000 groups used this database to publish their work on identifying the genetic defects in rare families.

3.6 Microarray

The SNP microarray technology moved rapidly from thousands of SNPs in the chip to more than one million variants for genotyping samples rapidly. Before the Human Genome Project era, the GWAS (Genome wide Association Study) involved genotyping of 300–600 microsatellite CA repeat markers in few hundred samples. The microarray brought a dramatic change to the GWAS, increasingly using more SNP variants, though less informative than microsatellite markers. The unimaginable 100-fold increase in marker numbers for genotyping thousands of samples by this technology produced large number of risk loci for many complex diseases. For example, with CA repeat marker genotyping, there were about 20–30 disease susceptibility loci that were identified for any complex diseases like hypertension, coronary artery disease (CAD), etc. With the increasing number of microarray platforms and variants with thousands of samples, 100s of new loci for the same diseases are obtained. Such a dramatic turn provided an opportunity to bioinformatics and statistics groups to work together to reveal the role of many novel genes and pathways in disease development in complex diseases. This opened up the new area of functional genomics to test large number of candidate genes and their role in the biology of diseases with high-throughput technologies in that field of its own, like proteomics, metabolomics, etc.

The SNP microarray platform also plays a crucial role in diagnosis of submicroscopic changes, which were not detected by conventional karyotyping or high-resolution banding techniques in cytogenetics. Many undiagnosed patients with conventional karyotyping approach were found to carry small changes which were precisely detected by this technology. Application of this technology in the diagnosis extended to the cancers for prognosis and for the personalized medicine program or precision medicine.

The NGS technology and SNP microarray complement each other in the diagnosis of many rare diseases, especially in families where consanguinity is reported. Power of this approach can be seen by the identification of many novel rare disease mutations in Middle East Arab population, where the highest consanguinity is reported in the world (Alrayes et al. 2016; Scott et al. 2016; Reynolds et al. 2017; Monies et al. 2017; Mohamoud et al. 2018).

In the diagnosis of diseases with many known causative genes, screening such a long list of genes is not cost-effective with any conventional methods. Microarrays and targeted gene panel screening makes the process simple and rapid. Many diagnostic screening panels based on the known mutations are available for many diseases, but they have their limitations as well. These panels will detect only known mutations but not the unknown variants or mutations in the sample. Targeted NGS gene panels are available as well as one can easily make such for their work quickly. These gene panels are helping the clinical team to better manage the patients without waiting for a long time.

3.7 Diagnostic Companies

High-throughput technologies including NGS and microarray spurred the new type of diagnostics biotech companies around the world. Spin-off from the academic labs as well as support from venture capital groups changed the landscape of diagnostic companies around the globe. Some of the biggest companies like *Centogene* in Germany developed disease-specific mutation screen for 1000s of genetic diseases through direct sequencing the gene or specific mutation by Sanger sequencing or by real-time PCR. If there is more than one gene with many exons that need to be screened, these companies use the targeted sequencing using NGS and/or microarray gene panels. These developments generated large-scale genetic data, which are being used for the identification of novel mutations. These large databases are being searched by many large pharmaceutical companies for their drug development process from drug target discovery, validation and precision medicine.

Next generation of biotech companies, which exploit the large-scale NGS genetic data, powerful bioinformatics tools and robust statistical methods, are those which use WGS/WES to provide most common disease predictive risk scores based on worldwide population data as well as GWAS data for many such diseases to the public. Companies like *23andMe* are mainly focussed on selling this service to the general public directly. These NGS companies also carry out the ancestry search using the powerful bioinformatic and statistical platforms for the general public. They compare the world population frequencies of the highly polymorphic markers across the genome from the WGS studies and match with the customer DNA sequence and give ancestry roots for them. These companies also use the generated WGS data for drug development in collaboration with big pharmaceutical companies in the world. Large pharmaceutical companies realise the potential of such large-scale genetic and genomic databases from a variety of ethnic backgrounds and try to exploit the same for the novel drug development process. Major focus of such companies is to identify the potential drug target gene for common diseases, reduce the cost of clinical trials by selecting patients who will be responding better to the new drug molecule through predictive marker mapping, avoid the adverse drug reactions of the novel compounds and personalized medicine for patients. Many of these diagnostics companies market many gene panel tools based on their existing collection of mutations for a particular disease in certain ethnic communities or groups or for the worldwide population and offer this service for new patients. Cardiac arrhythmia panel in many companies feature known mutations or targeted sequencing of many genes which have been reported to be mutated in identified cases. Saudi Genome Project group recently published many such gene panels targeting the Saudi's Arab population for a variety of organ-specific diseases. They generated such panels based on exome sequencing of 100s of Saudi patients with such diseases, and unique mutations in these mutations were selected for the panel (Monies et al 2017).

3.8 Transition of Diagnostics to Drug Discovery and Precision Medicine

Diagnostics of rare diseases and thousands of monogenic diseases resulted in overload of information which is being exploited by many research groups and commercial companies, from spin-off start-up biotech companies of the universities to large pharmaceutical companies. Many large-scale diagnostics and genomic service companies like BGI, Centogene, 23andMe etc. were actively involved in one or all of these activities through commercial licensing of the data accumulated over the years to commercial organizations in drug development or in precision medicine activities of the existing drugs in stratification of patients to target those with the best response for future clinical trials or for marketing their product.

Leading pharmaceutical companies collect the large-scale data from these companies for comparison of their internally stored data from different clinical trials to increase the chances of success in clinical trials as well as conduct the trials with smaller number of patients with certain genetic profile, who might respond better to the drug molecule in question, or identify markers which stratify the clinical trial patients to be responders and non-responders for targeted marketing once the regulatory approval is obtained.

3.9 National Genome Projects

Impact of Human Genome Project over the last decade has been impressive in many areas of research. Many countries recognized the potential role of genomics in reducing the disease burden of the people, healthcare system burden and the economic burden to the family and the country through their suffering in all fronts. Major players of the Human Genome Project were the first ones to step up the effort in setting up multiple large-scale population sequencing studies to address potential biomarker panel development for screening and precision medicine and to develop novel drug-target identification by harnessing the power of genomic data of the population (Fig. 3.1). Objectives of most of the national genome projects are similar, but the scale and scope is limited by various other factors.

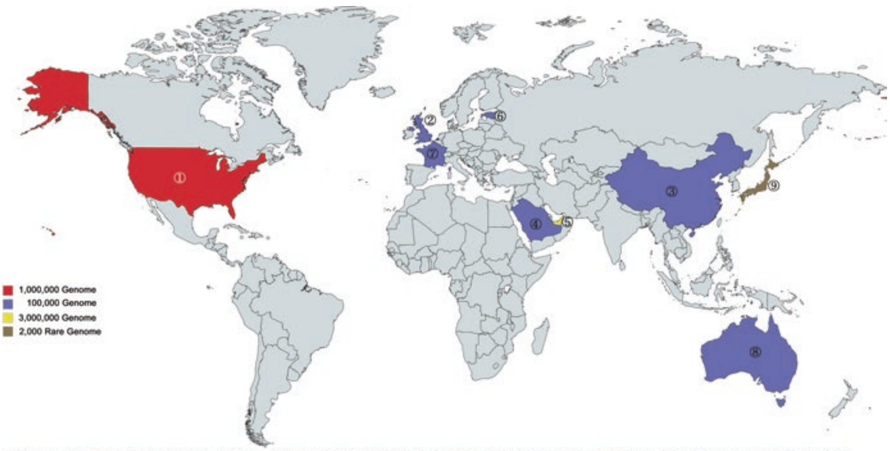
3.9.1 USA

It initiated few initiatives to address multiple objectives. In 2011, National Human Genome Research Institute (NHGRI) with co-funding from National Heart Lung and Blood Institute and National Eye Institute of NIH funded *Centres of Mendelian Genomics* to identify genetic defects in many Mendelian disorders. These four centres were sequencing the exomes of patients with rare diseases from around the USA.

1. Baylor College of Medicine-Johns Hopkins University CMG (http://bhcmg.org/)
2. Broad Institute Joint CMG (https://www.broadinstitute.org/news/7773)
3. University of Washington CMG (http://uwcmg.org/#/)
4. Yale University CMG (https://medicine.yale.edu/keck/ycga/)

The NIH initiated this program to identify the genetic defects in rare and unrecognized diseases seen in the population. These genetic discoveries will help boost the understanding of the biology of the disease development and explore the possibility of novel therapies for them and other associated diseases.

Recently, NIH (National Institute of Health), USA, released the new "*All of Us research program*" (https://allofus.nih.gov/). This program will target collecting health and wellness data from one million Americans over the age of 18 years. In the first phase of generating the genomic data, 100,000 participants each will be generated by 3 centres initially. From the second year, it will be scaled up to 200,000 samples per centre till they sequence all participants by the fifth year. The health data from electronic records and survey will be used to address the Precision Medicine Initiative (PMI) of the NIH which was launched in 2016. National Cancer



1.USA : 1,000,000 Genome @ Veterans Project & All of Us Reserach Program | 2.United Kingdom: 100,000 Genomes project | 3.China: 100,000 genomes project | 4.Saudi Arabia: 100,000 Genome Project (Saudi Genome) | 5.United Arab Emirates: 3,000,000 Genome Project | 6.Estonia: 100,000 Genome Project (Personalized Medicine Program) 7.France: 100,000 Genome Project (French Plan for Genomic Medicine 2025) | 8.Australia: 100,000 Genome Project (The Australian Genomics Health Futures Mission) 9.Japan: 2,000 Genome Project (Initiatives on Rare and Undiagnosed Diseases)

Fig. 3.1 Caption

Institute (NCI) leads the cancer genomics efforts of the PMI that will have a big impact on this program by reducing the cost of cancer patient management.

Million Veterans Program (MVP): (https://www.research.va.gov/MVP/default.cfm) also is making big impact on research. Long-term medical history of large group of veterans will be a boon to the science in the future.

It is one of the largest voluntary programs funded by Department of Veterans Affairs Research and Development. One million volunteers' baseline health survey collects all health-related data and blood samples to understand the role of genes to a variety of health conditions including cancer, cardiovascular diseases, diabetes, kidney diseases, etc. (Klarin et al. 2018). This program is part of the Precision Medicine Initiative of the USA in 2015.

3.9.2 UK

Genomics England is wholly owned by Department of Health and Social Care to carry out the 100,000 genome project. This project is to collect blood samples from NHS (National Health Services) patients with rare diseases, families and cancers for whole genome sequencing. For the recruitment of the patients and families for the study, 11 Genomic Medicine Centres were created, who will collect all the necessary clinical and blood samples for processing and analysis. This project already sequenced 70,000 genomes, and researchers from around the world are exploring the data to identify various diseases risk loci. This will pave the way for the NHS to transform how the patients are cared for with the advanced technology (Gräf et al. 2018; Turnbull et al. 2018; Klintman et al. 2018; Grant and Maytum 2018; Barwell et al. 2018). Based on the success of this project, the NHS already initiated nationwide genomic medicine service through genomic medical centres for the routine use for clinicians for the accurate diagnosis, precision medicine and better patient care.

3.9.3 UK Biobank

The UK Biobank has recruited 500,000 volunteers to provide blood, urine and saliva samples and provide health data from the NHS as well as provide data through an extended survey in the UK. Access to the genetic and health information of these participants was given to many research groups across the world for analysis to unravel the role of genetic factors to many diseases of interest to the research groups. This resulted in some of the highly impactful research to open up the window of opportunities to address the disease management through early diagnosis, novel drug target identification, identification of biomarkers, etc. (Bycroft et al. 2018; Elliott et al. 2018; Haas et al. 2018; Inouye et al. 2018). This is being funded by many charities, including Wellcome Trust, British Heart Foundation, Cancer

Research UK, Diabetes UK and government arms of research and service including Medical Research Council (MRC, UK), Scottish and Welsh governments and National Health Services (NHS, UK).

Many other countries are following these trends and setting up their own national genomics programs, listed below with the web links.

China (http://encs.hit.edu.cn/2018/0611/c5396a210190/page.htm)
Japan (https://www.amed.go.jp/en/program/IRUD/)
Estonia (https://www.sm.ee/en/personalised-medicine)
Australia (https://www.australiangenomics.org.au/)
France (https://www.france-genomique.org/spip/?lang=fr)
Saudi Arabia http://shgp.kacst.edu.sa/site/)
Dubai (https://www.dha.gov.ae/en/pages/dubaigneomicsabout.aspx)

Next few years will witness the outcome of these population combined genomic and clinical data analyses will drive applied research towards functional genomics, personalized medicine, pharmacogenomics and drug discovery of novel targets for many diseases of the mankind. Gene-editing technology is being explored for functional validation of the genetic mutations and correction of the genetic defect. This will bring the new era of personalized genetic surgery to a reality in the near future.

3.10 Conclusion

Rapidly changing technology and methods made it possible to diagnose many rare diseases. These technologies in combination with bioinformatics are helping the patient and family with the accurate detection of the mutations in their samples. This helps the family and clinical teams in better planning and management of the patient care. This chapter highlights few of the technological revolutions in the diagnosis.

References

Alrayes N, Mohamoud HS, Ahmed S, Almramhi MM, Shuaib TM, Wang J, Al-Aama JY, Everett K, Nasir J, Jelani M (2016) The alkylglycerol monooxygenase (AGMO) gene previously involved in autism also causes a novel syndromic form of primary microcephaly in a consanguineous Saudi family. J Neurol Sci 15/363:240–244

Altschul S, Gish W, Miller W, Myers E, Lipman D (1990) Basic local alignment search tool. J Mol Biol 215(3):403–410. https://doi.org/10.1016/S0022-2836(05)80360-2. PMID 2231712

Barwell JG, O'Sullivan RBG, Mansbridge LK, Lowry JM, Dorkins HR (2018) Challenges in implementing genomic medicine: the 100,000 genomes project. J Transl Genet Genom 2:13. https://doi.org/10.20517/jtgg.2018.17

Bycroft C, Freeman C, Petkova D, Band G, Elliott LT, Sharp K, Motyer A, Vukcevic D, Delaneau O, O'Connell J, Cortes A, Welsh S, Young A, Effingham M, McVean G, Leslie S, Allen N, Donnelly P, Marchini J (2018) The UK Biobank resource with deep phenotyping and genomic data. Nature 562(7726):203–209. https://doi.org/10.1038/s41586-018-0579-z

Deciphering Developmental Disorders Study -265 collaborators (2015) Large-scale discovery of novel genetic causes of developmental disorders. Nature 519(7542):223–228. https://doi.org/10.1038/nature14135

Elliott LT, Sharp K, Alfaro-Almagro F, Shi S, Miller KL, Douaud G, Marchini J, Smith SM (2018) Genome-wide association studies of brain imaging phenotypes in UK Biobank. Nature 562(7726):210–216. https://doi.org/10.1038/s41586-018-0571-7. Epub 2018 Oct 10

Firth HV, Richards SM, Paul Bevan AP, Clayton S, CorpasM RD, Van Vooren S, Moreau Y, Pettett RM, Carter NP (2009) DECIPHER: database of chromosomal imbalance and phenotype in humans using ensembl resources. Am J Hum Genet 84:524–533

Gräf S, Haimel M, Bleda B, Hadinnapola C, Southgate L, Li W, Hodgson J, Liu B, Salmon RM, Southwood M, Machado RD, Martin JM, Treacy CM, Yates K, Daugherty LC, Shamardina O, Whitehorn D, Holden S, Aldred M, Bogaard HJ, Church C, Coghlan G, Condliffe R, Corris PA, Danesino C, Eyries M, Gall H, Ghio S, Ghofrani HA, JSR G, Girerd B, Houweling AC, Howard L, Humbert M, Kiely DG, Kovacs G, RV MKR, Moledina S, Montani D, Newnham M, Olschewski A, Olschewski H, Peacock AJ, Pepke-Zaba J, Prokopenko I, Rhodes CJ, Scelsi L, Seeger W, Soubrier F, Stein DF, Suntharalingam J, Swietlik EM, Toshner MR, van Heel DA, Vonk Noordegraaf A, Waisfisz Q, Wharton J, Wort SJ, Ouwehand WH, Soranzo N, Lawrie A, Upton PD, Wilkins MR, Trembeth RC, Morrell NW (2018) Identification of rare sequence variation underlying heritable pulmonary arterial hypertension. Nat Commun 9:1416. https://doi.org/10.1038/s41467-018-03672-4

Grant M, Maytum JP (2018) What will follow the first hundred thousand genomes in the NHS? Per Med 15(4):239–241. https://doi.org/10.2217/pme-2018-0025

Gusella JF, Wexler NS, Conneally PM, Naylor SL, Anderson MA, Tanzi RE, Watkins PC, Ottina K, Wallace MR, Sakaguchi AY, Young AB, Shoulson I, Bonilla E, Martin JB (1983) A polymorphic DNA marker genetically linked to Huntington's disease. Nature 306(5940):234–238

Haas ME, Aragam KG, Emdin CA, Bick AG, International Consortium for Blood Pressure, Hemani G, Davey Smith G, Kathiresan S (2018) Genetic Association of Albuminuria with cardiometabolic disease and blood pressure. Am J Hum Genet 103(4):461–473. https://doi.org/10.1016/j.ajhg.2018.08.004. Epub 2018 Sep 13

Inouye M, Abraham G, Nelson CP, Wood AM, Sweeting MJ, Dudbridge F, Lai FY, Kaptoge S, Brozynska M, Wang T, Ye S, Webb TR, Rutter MK, Tzoulaki I, Patel RS, Loos RJF, Keavney B, Hemingway H, Thompson J, Watkins H, Deloukas P, Di Angelantonio E, Butterworth AS, Danesh J, Samani NJ, UK Biobank CardioMetabolic Consortium CHD Working Group (2018) Genomic risk prediction of coronary artery disease in 480,000 adults: implications for primary prevention. J Am Coll Cardiol 72(16):1883–1893. https://doi.org/10.1016/j.jacc.2018.07.079

Klarin D, Damrauer SM, Cho K, Sun YV, Teslovich TM, Honerlaw J, Gagnon DR, DuVall SL, Li J, Peloso GM, Chaffin M, Small AM, Huang J, Tang H, Lynch JA, Ho YL, Liu DJ, Emdin CA, Li AH, Huffman JE, Lee JS, Natarajan P, Chowdhury R, Saleheen D, Vujkovic M, Baras A, Pyarajan S, Di Angelantonio E, Neale BM, Naheed A, Khera AV, Danesh J, Chang KM, Abecasis G, Willer C, Dewey FE, Carey DJ, Global Lipids Genetics Consortium, Myocardial Infarction Genetics (MIGen) Consortium, Geisinger-Regeneron DiscovEHR Collaboration, VA Million Veteran Program, Concato J, Gaziano JM, O'Donnell CJ, Tsao PS, Kathiresan S, Rader DJ, Wilson PWF, Assimes TL (2018) Genetics of blood lipids among ~300,000 multi-ethnic participants of the Million Veteran Program. Nat Genet 50(11):1514–1523. https://doi.org/10.1038/s41588-018-0222-9

Klintman J, Barmpouti K, Knight SJL, Robbe P, Dreau H, Clifford R, Ridout K, Burns A, Timbs A, Bruce D, Antoniou P, Sosinsky A, Becq J, Bentley D, Hillmen P, Taylor JC, Caulfield M, Schuh AH (2018) Clinical-grade validation of whole genome sequencing reveals robust detection of low-frequency variants and copy number alterations in CLL. Br J Haematol 182:412–417. https://doi.org/10.1111/bjh.15406

Lipman DJ, Pearson WR (1985) Rapid and sensitive protein similarity searches. Science 227(4693):1435–1441. https://doi.org/10.1126/science.2983426. PMID 2983426

Mohamoud HS, Ahmed S, Jelani M, Alrayes N, Childs K, Vadgama N, Almramhi MM, Al-Aama JY, Goodbourn S, Nasir J (2018) A missense mutation in TRAPPC6A leads to build-up of the protein, in patients with a neurodevelopmental syndrome and dysmorphic features. Sci Rep 8(1):2053. https://doi.org/10.1038/s41598-018-20658-w

Monies D, Abouelhoda M, AlSayed M, Alhassnan Z, Alotaibi M, Kayyali H, Al-Owain M, Shah A, Rahbeeni Z, Al-Muhaizea MA, Alzaidan HI, Cupler E, Bohlega S, Faqeih E, Faden M, Alyounes B, Jaroudi D, Goljan E, Elbardisy H, Akilan A, Albar R, Aldhalaan H, Gulab S, Chedrawi A, Al Saud BK, Kurdi W, Makhseed N, Alqasim T, El Khashab HY, Al-Mousa H, Alhashem A, Kanaan I, Algoufi T, Alsaleem K, Basha TA, Al-Murshedi F, Khan S, Al-Kindy A, Alnemer M, Al-Hajjar S, Alyamani S, Aldhekri H, Al-Mehaidib A, Arnaout R, Dabbagh O, Shagrani M, Broering D, Tulbah M, Alqassmi A, Almugbel M, AlQuaiz M, Alsaman A, Al-Thihli K, Sulaiman RA, Al-Dekhail W, Alsaegh A, Bashiri FA, Qari A, Alhomadi S, Alkuraya H, Alsebayel M, Hamad MH, Szonyi L, Abaalkhail F, Al-Mayouf SM, Almojalli H, Alqadi KS, Elsiesy H, Shuaib TM, Seidahmed MZ, Abosoudah I, Akleh H, AlGhonaium A, Alkharfy TM, Al Mutairi F, Eyaid W, Alshanbary A, Sheikh FR, Alsohaibani FI, Alsonbul A, Al Tala S, Balkhy S, Bassiouni R, Alenizi AS, Hussein MH, Hassan S, Khalil M, Tabarki B, Alshahwan S, Oshi A, Sabr Y, Alsaadoun S, Salih MA, Mohamed S, Sultana H, Tamim A, El-Haj M, Alshahrani S, Bubshait DK, Alfadhel M, Faquih T, El-Kalioby M, Subhani S, Shah Z, Moghrabi N, Meyer BF, Alkuraya FS (2017) The landscape of genetic diseases in Saudi Arabia based on the first 1000 diagnostic panels and exomes. Hum Genet 136(8):921–939. https://doi.org/10.1007/s00439-017-1821-8

Pearson WR, Lipman DJ (1988) Improved tools for biological sequence comparison. Proc Natl Acad Sci U S A 85(8):2444–2448. https://doi.org/10.1073/pnas.85.8.2444

Pennarun G, Escudier E, Chapelin C, Bridoux A-M, Cacheux V, Roger G, Clément A, Goossens M, Amselem S, Duriez B (1999) Loss-of-function mutations in a human gene related to *Chlamydomonas reinhardtii* dynein IC78 result in primary ciliary dyskinesia. Am J Hum Genet 65:1508–1519. https://doi.org/10.1086/302683

Reynolds JJ, Bicknell LS, Carroll P, Higgs MR, Shaheen R, Murray JE, Papadopoulos DK, Leitch A, Murina O, Tarnauskaitė Ž, Wessel SR, Zlatanou A, Vernet A, von Kriegsheim A, Mottram RM, Logan CV, Bye H, Li Y, Brean A, Maddirevula S, Challis RC, Skouloudaki K, Almoisheer A, Alsaif HS, Amar A, Prescott NJ, Bober MB, Duker A, Faqeih E, Seidahmed MZ, Al Tala S, Alswaid A, Ahmed S, Al-Aama JY, Altmüller J, Al Balwi M, Brady AF, Chessa L, Cox H, Fischetto R, Heller R, Henderson BD, Hobson E, Nürnberg P, Percin EF, Peron A, Spaccini L, Quigley AJ, Thakur S, Wise CA, Yoon G, Alnemer M, Tomancak P, Yigit G, Taylor AM, Reijns MA, Simpson MA, Cortez D, Alkuraya FS, Mathew CG, Jackson AP, Stewart GS (2017) Mutations in DONSON disrupt replication fork stability and cause microcephalic dwarfism. Nat Genet 49(4):537–549. https://doi.org/10.1038/ng.3790

Scott EM, Halees A, Itan Y, Spencer EG, He Y, Azab MA, Gabriel SB, Belkadi A, Boisson B, Abel L, Clark AG, Greater Middle East Variome Consortium, Alkuraya FS, Casanova JL, Gleeson JG (2016) Characterization of greater middle eastern genetic variation for enhanced disease gene discovery. Nat Genet 48(9):1071–1076. https://doi.org/10.1038/ng.3592

Smith TF, Waterman MS (1981) Identification of common molecular subsequences. J Mol Biol 147:195–197. https://doi.org/10.1016/0022-2836(81)90087-5. PMID 7265238

The Huntington's Disease Collaborative Research Group (1993) A novel gene containing a trinucleotide repeat that is expanded and unstable on Huntington's disease chromosomes. Cell 72(6):971–983. https://doi.org/10.1016/0092-8674(93)90585-E. PMID 8458085

Turnbull C, Scott RH, Thomas E, Jones L, Murugaesu N, Pretty FB, Halai D, Baple E, Craig C, Hamblin A, Henderson S, Patch C, O'Neill A, Devereaux A, Smith K, Martin AR, Sosinsky A, McDonagh EM, Sultana R, Mueller M, Smedley D, Toms A, Dinh L, Fowler T, Bale M, Hubbard T, Rendon A, Hill S, Caulfield MJ, 100 000 Genomes Project (2018) The 100 000 Genomes Project: bringing whole genome sequencing to the NHS. BMJ 361:k1687. https://doi.org/10.1136/bmj.k1687

Chapter 4
Genomic Revolution-Driven Cancer Research

Meganathan P. Ramakodi and Muthukrishnan Eaaswarkhanth

Contents

M. P. Ramakodi
CSIR-National Environmental Engineering Research Institute, Hyderabad Zonal Centre, Hyderabad, India
e-mail: pr.meganathan@neeri.res.in

M. Eaaswarkhanth (✉)
Genetics and Bioinformatics, Dasman Diabetes Institute, Dasman, Kuwait
e-mail: eaaswar.muthukrishna@dasmaninstitute.org

N. A. Shaik et al. (eds.), *Essentials of Bioinformatics, Volume II*,
https://doi.org/10.1007/978-3-030-18375-2_4

4.1 Introduction

The increasing potential of genomic research has integrated it into the global mainstream healthcare systems. Evidently, the UK 100,000 Genomes Project (UK10K Consortium et al. 2015), the Personal Genome Project Canada (Reuter et al. 2018), and precision medicine initiatives of the USA (Collins and Varmus 2015) and China (Li 2016) are heading the way to personalized healthcare. At this exciting era of genomics, the affordability of next-generation sequencing (NGS) technologies geared up the data generation from whole genomes and exomes that play a decisive role in clinical diagnostics (Dewey et al. 2014; Hegde et al. 2017; Lionel et al. 2017; Posey et al. 2016; Taylor et al. 2015). Here comes the imperative participation of computational genomicists or, in broad terms, bioinformaticians that critically carry out the downstream or post-sequencing analysis to extract relevant information from the sequenced genomes (Oliver et al. 2015). As such, the application of NGS in clinical interventions is continuing to provide more facts on the genetic susceptibility to diseases (Taylor et al. 2015), unravel the basis of common and rare genetic disorders (Lee et al. 2014a), track the spreading of infectious diseases (Metsky et al. 2017), and categorize the subtypes of cancer (Foley et al. 2015; Müllauer 2017) and most importantly, prenatal as well as newborn screening (Stavropoulos et al. 2016). Day by day hundreds of thousands of genome sequences are being generated and deposited in high-throughput storages awaiting computational explorations. This is an inspirational opportunity for any level of candidate aspiring to acquire the required computational skills involved in NGS analysis. Considering the extent and importance of NGS in the clinic setting for health improvement and the need to develop the related bioinformatics skill set, in this chapter, we have put together most of the needed information on the step-by-step analysis methodologies with examples.

4.2 All About Next-Generation Sequencing

Although the history of sequencing dates back to 1960s, the incessant DNA sequencing technique was established in 1977 by Fred Sanger (Heather and Chain 2016). The iconic Human Genome Project was successfully completed using the Sanger and shotgun sequencing techniques that belong to the first generation of sequencing technologies. Nearly after three decades, since 2007, the second-generation sequencers from Roche, Applied Biosystems, and Illumina were extensively employed in genetic research as next-generation sequencers. Eventually, the new or third-generation sequencing took a new dimension with the advent of Oxford Nanopore Technologies (Clarke et al. 2009; Jain et al. 2016) that has recently made giant strides in sequencing and assembling the human genome filling some gaps missed out by other technologies (Jain et al. 2018). Heather and Chain, in their review, recapitulate the rich history and progress of DNA sequencing technologies in detail (Heather and Chain 2016). A brief overview on the widely used NGS platforms is presented in Table 4.1.

Table 4.1 A brief overview on the widely used NGS platforms

Platform[a]	Sequencing method	Sequencing chemistry	Read type	Average read length	Error type	References
New or third-generation sequencers						
Oxford Nanopore	Single-molecule nanopore sequencing	DNA molecule traverses the pore	Template, complement & two direction	Variable up to 900 kb	GC bias	Jain et al. (2018), Lu et al. (2016), Madoui et al. (2015)
Ion Torrent	Semiconductor-based sequencing by synthesis	Detection of released proton	Single-read and paired-end	200–400 bp	INDEL	Quail et al. (2012)
Pacific Biosystems	Single-molecule real-time (SMRT) sequencing by synthesis	Fluorescently labeled nucleotides	Single-read	10–20 kb	INDEL	Carneiro et al. (2012), Koren et al. (2012), Quail et al. (2012), Salmela and Rivals (2014)
Second-generation sequencers						
Illumina	Sequencing by synthesis	Reversible dye terminators	Single-read and paired-end	100–500 bp	Substitution	Ross et al. (2013)
AB SOLiD	Sequencing by ligation	Oligonucleotides chained ligation	Single-read and paired-end	100 bp	A/T bias	Glenn (2011)
Roche 454	Pyrosequencing	Pyrosequencing	Single-read and paired-end	700 bp	INDEL	Gilles et al. (2011)

[a] Models and series of the sequencing machine names not specified here

4.2.1 Applications of Next-Generation Sequencing Methods in Genomic Research

It is obvious that over the past decade, there has been swift development of NGS technologies (Goodwin et al. 2016) that completely transformed genetic and genomic research applications (Koboldt et al. 2013). The most common NGS applications include DNA sequencing, RNA sequencing (RNA-Seq), chromatin immunoprecipitation sequencing (ChIP-Seq), and methylation sequencing (Methyl-Seq) or whole genome bisulfite sequencing (WGBS). The DNA sequencing refers to whole genome sequencing (WGS), whole exome sequencing (WES), and target region or gene sequencing, which are scrutinized to detect single-nucleotide variants (SNVs), genomic structural variants (SVs) like copy number variants (CNVs), small to big range of insertions and deletions (INDELs), and duplications and transversions associated with human phenotypes and diseases. The gene expression profiles could be extracted using RNA-Seq to derive information about novel transcripts including the genomic "dark matter," long non-coding RNAs (lncRNAs). As the name implies, ChIP-Seq is employed to determine the modifications associated with chromatin and identify the transcription factor binding sites in the genome level. With Methyl-Seq, the methylation patterns across the genome regions especially CpG, CHH, and CHG can be studied. Typically, all applications exploring every single genomic alteration are focused toward protection and treatment of diseases.

4.3 Next-Generation Sequencing Analysis Workflow

Among various sequencing analysis methods, in this review, we concentrate on the WGS- and WES-based workflow for post-sequencing processing, variant detection, and related clinical implications. A common pipeline for WES analysis involves data preprocessing, alignment, variant calling, annotation, and prioritization. WES experimental and computational workflow is shown in Fig. 4.1.

4.3.1 Data Preprocessing

The initial step of NGS computational analysis post-sequencing is to perform the quality checks (QC) of raw reads that are in FASTQ (Cock et al. 2010) file format. Then QC is followed by filtering, trimming, or correcting reads that do not fit the defined quality standards. The common errors expected in sequencing data include base-calling errors, INDELs, poor-quality reads, and adaptor cross contamination (Dai et al. 2010). These errors occur due to failures in instrument hardware, optical sensors, and varied sequencing chemistry (Cox et al. 2010; Dohm et al. 2008). It should be noted here that many NGS downstream analysis pipelines are not built to

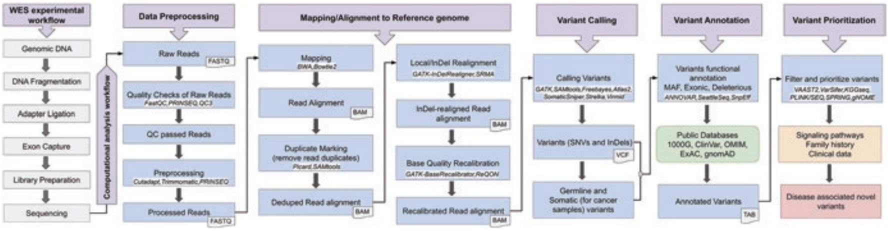


Fig. 4.1 The computational analysis workflow of whole exome sequencing data∗. ∗This is a general workflow for the downstream analysis of whole exome sequences. Sequencing experiment followed by five major steps of computational analysis is shown. The file formats are presented within the flowchart *document* shape and relevant description given in glossary. The program tools employed at each step are in italics and highlighted in white color

deal with poor- or low-quality sequence reads, and so raw data QC and preprocessing step become imperative to avoid any false-positive inferences. Collectively, the following are conducted as preprocessing steps: (1) visualization of the distribution of Phred-scaled base quality scores along the reads, GC content, read length, and sequence duplication level, (2) trimming of base reads, and (3) read filtering or adaptor clipping based on Phred score and sequence properties like primer or adaptor contaminations, N content, and GC bias. Some of the open-source tools available to perform these jobs are FastQC (Andrews 2010), Cutadapt (Martin 2011), and Trimmomatic (Bolger et al. 2014), while PRINSEQ (Schmieder and Edwards 2011) and QC3 (Guo et al. 2014) are package suites providing preprocessing functions.

4.3.2 Alignment

After QC and preprocessing, standard-quality sequence reads are available to map and align against the reference genome. The read alignment enables comparison of the sequenced data with the reference genome to determine the genomic variations. The largely used versions, GRCh37 and GRCh38, of human reference genome can be downloaded from the NCBI website (https://www.ncbi.nlm.nih.gov/genome/guide/human/). BWA (Li et al. 2009) and Bowtie2 (Wu and Nacu 2010) are the two popular short-read alignment tools that apply Burrows-Wheeler transformation (BWT) compression technique algorithm. The QC measures followed in the alignment against the reference sequence include the proportion of all aligned reads, the ratio of unique aligned reads, and the number of reads aligned at a specific locus. Further, to reduce the possible artifacts affecting the accuracy of subsequent variant calling step, the following three processing steps are performed: (1) duplicate marking or removal of read duplicates, (2) local or INDEL realignment, and (3) base quality recalibration. During the alignment process, some of the reads aligned exactly with the mapping coordinates are known as "read duplicates." These read duplicates may either be real DNA materials or PCR artifacts. It is very difficult to determine the real case with the alignment information alone. Therefore, in case of WES analysis, it is essential to remove the read duplicates before variant calling so as to rule out PCR-introduced artifact from the uneven DNA amplification. Several tools such as Picard MarkDuplicates (http://broadinstitute.github.io/picard/) and SAMtools (Li et al. 2009) efficiently detect the read duplicates based on the orientation on the genome. Once the read duplicates are removed, the next step is to detect the genomic regions with INDELs and improve the alignment quality in the specified region. This is because, in comparison to the regions containing only SNVs, the INDEL regions are likely to be noisy, which requires improvement of gapped alignment. IndelRealigner from the Genome Analysis Toolkit (GATK) (McKenna et al. 2010) and SRMA (Homer and Nelson 2010) are mostly used to perform improved local realignment to frame the consensus sequence for INDEL discovery. Followed by the local or indel realignment, the base quality recalibration is performed. As the Phred-scaled base quality is an essential factor for precise variant detection in the downstream analysis, the base quality recalibration step is recommended before proceeding to calling variants. This base quality recalibration is commonly done using BaseRecalibrator from GATK (McKenna et al. 2010) and ReQON (Cabanski et al. 2012). GATK BaseRecalibrator recalibrates the base scores of the alignment files from multiple sequencing runs (McKenna et al. 2010), whereas ReQON along with recalibration provides range of diagnostic data as well as plots prior and post-recalibration to demonstrate the improved accuracy (Cabanski et al. 2012).

4.3.3 Variant Calling

From the previous steps, standard-quality sequence reads, mapped and aligned against the reference genome, are available to detect the genomic variants like SNVs and SVs (CNVs and INDELs). This is done by calling variants that differ from the reference sequence. The accuracy of variant calling step largely depends on the higher read depth that helps in detecting rare genetic variants. For example, WES requires 100× read depth for heterozygous SNV detection, while WGS requires 35× and 60× read depths for detecting genotype and INDELs, respectively (Sims et al. 2014). Regarding SNVs, two types of variants such as germline and somatic can be called separately according to the need. The germline variants are inherited from the parents and exist in every cell, whereas somatic variants occur during the lifetime of an individual. The contributions of germline variants are generally studied for complex diseases like diabetes. Some of the largely used programs for germline variant calling include GATK (McKenna et al. 2010), SAMtools (Li et al. 2009), FreeBayes (Garrison and Marth 2012), and Atlas2 (Challis et al. 2012). GATK implements mainly two variant programs, UnifiedGenotyper and HaplotypeCaller, to detect SNVs and INDELs. The former one identifies SNVs and INDELs separately assuming that every single variant locus is independent, whereas the latter calls SNVs, INDELs, and some SV classes concurrently. SAMtools package consists of a battery of utilities to manipulate the aligned sequence reads in the SAM or BAM format and call the SNVs and INDELs. FreeBayes is a haplotype-based program tool that simultaneously detects SNVs, INDELs, multi-base mismatches, polyallelic sites, polyploidy, and CNVs in a single sample, pooled multiple samples, or mixed populations (Garrison and Marth 2012). Atlas2 implements logistic regression models trained on validated WES data to detect SNVs and INDELs from the data generated by the SOLiD^{TM} platform. Also, this tool is used to analyze the Illumina data using logistic regression models to call INDELs and a combination of logistic regression and a Bayesian model to call SNVs (Challis et al. 2012). Several variant calling programs are being developed and evaluated, so it is recommended to critically choose the suitable one depending on the need of the study (detailed evaluation by Hwang et al. 2015; Sandmann et al. 2017).

The somatic variants associated with disease state are studied for nonheritable diseases like some cancers by comparing tumor and normal samples. A number of somatic variant caller tools are available (Cai et al. 2016; Krøigård et al. 2016), and the following tools SomaticSniper (Larson et al. 2012), Strelka (Saunders et al. 2012), and Virmid (Kim et al. 2013) are discussed here. For calling the somatic variants, SomaticSniper compares the diploid genotype likelihood in the tumor and normal pair (Larson et al. 2012). Strelka implements a Bayesian model-based algorithm to derive a score from the combined probability of a somatic variant and a specific genotype in the normal samples for variant calling and computes the allele frequency variation in samples at any level without requiring an estimation of tumor purity (Saunders et al. 2012). In contrast to Strelka, Virmid considers the level of impurity in the sample and utilizes a similar Bayesian model and the maximum likelihood estimation (Kim et al. 2013). Virmid also accounts for various other noise types including sequencing errors, mapping bias, and CNV stage (Kim et al. 2013).

4.3.4 Variant Annotation

Further to the detection of different classes of genomic variants, annotation is crucial to understand their functional attributes such as synonymous, non-synonymous, loss-of-function (LoF), and the like. Majority of disease genetic studies concentrate on the non-synonymous SNVs, LoF variants, and INDELs in the exonic regions that are mostly associated with Mendelian and complex diseases. It is also important to consider synonymous SNVs to estimate the background mutation rate in the genome. Apart from these basic annotations, there are several programs that integrate public databases to provide supplementary information of the variants such as minor allele frequency (MAF) in normal global populations, experimental evidence from clinical studies, deleterious effect prediction of variant function, and collection of variants and genes in disease studies. ANNOVAR is one of the widely used variant annotation programs that annotate in three modes, gene-based, region-based, and filter-based (Wang et al. 2010). This program integrates about 4000 public databases to detect variants reported especially in dbSNP (Sherry et al. 2001), 1000 Genomes Project (Auton et al. 2015), NHLBI ESP6500 (http://evs.gs.washington.edu/EVS/), ClinVar (Landrum et al. 2014), and ExAC (Lek et al. 2016). In addition, ANNOVAR combines various deleterious function prediction tools, namely, PolyPhen-2 (Adzhubei et al. 2010), Sorting Intolerant From Tolerant (SIFT) (Kumar et al. 2009), and the Combined Annotation Dependent Depletion (CADD) (Kircher et al. 2014), to provide deleterious scores of the annotated variants. Some of the other annotation programs used in common are snpEff (Cingolani et al. 2012) and the Ensembl Variant Effect Predictor (VEP) (McLaren et al. 2016). PharmGKB (Whirl-Carrillo et al. 2012) database can be used to annotate the variants of pharmacogenetic importance.

4.3.5 Variant Prioritization

This is the prominent decision-making step, which aids in identifying causal variant for disease of interest. During the study of Mendelian, rare, and complex diseases, it is challenging to discern the disease-causing variants among tens of thousands of annotated variants. Notably, a large number of variants are called for all study designs, ranging from single individual, trio (affected child and parents), family (affected and unaffected individuals), disease vs normal tissue (e.g., cancer) to unrelated case-control cohort, requiring different statistical and data processing pipelines. On average, typical WGS experiment generates approximately 1–1.5 million variants, and WES yields about 50,000 variants (O'Rawe et al. 2013). Therefore, in the direction of detecting functional impact variants, it is indispensable to filter out the unreliable variants and prioritize the ones that likely cause the disease for further investigation. The filtering criteria include removal of variants (1) with low coverage and quality, strand bias, and low-confidence read alignment, (2) with common and low frequency, and (3) deviating from Hardy-Weinberg equilibrium. Ultimately, variants that change the amino acid and have functional effect are prioritized from the filtered variant list. Many tools are available to filter, evaluate, and prioritize thousands of variants collectively and systematically, considering annotation outcomes, patient familial information, phenotypes, and disease subtype information. VAAST2 is one such tool that generates variant lists with ranking and sorting according to its importance for the disease (Hu et al. 2013). This is very helpful in the analysis of complex genetic and rare Mendelian diseases. The other publicly available tools are VarSifter (Teer et al. 2012), KGGseq (Li et al. 2012), PLINK/SEQ (https://atgu.mgh.harvard.edu/plinkseq/), SPRING (Wu et al. 2014), and gNOME (Lee et al. 2014b). Comprehensive review on variant prioritization pipelines have been published recently (Eilbeck et al. 2017; Jalali Sefid Dashti and Gamieldien 2017) for further reading.

So far, NGS computational analysis workflow (Fig. 4.1) has been outlined generally. The following section will discuss the common applications of NGS sequencing approach in clinical research of head and neck squamous cell carcinoma (HNSCC).

4.4 Clinical and Research Applications of Next-Generation Sequencing Technology in Head and Neck Squamous Cell Carcinoma

Head and neck squamous cell carcinoma (HNSCC)—cancers of oral cavity, oropharynx, and larynx—is the sixth most common cancer type worldwide. The major risk factors associated with HNSCC are tobacco and alcohol usage and human papilloma virus (HPV) (Ragin et al. 2007). In addition, recent studies have shown genetics to be a significant factor associated with HNSCC (Ragin et al. 2007; Ramakodi et al. 2016, 2017). Pertaining to the importance of genetic factors, the high-throughput sequence approach is preferred in HNSCC studies as parallel sequencing method yields large data and could provide more details than traditional approach. In addition, the decreasing cost of DNA sequencing has enabled the broad use of NGS techniques to study the genetic changes in HNSCC.

4.4.1 Molecular Characterization and Subtypes in Head and Neck Squamous Cell Carcinoma

HNSCC is a complex and heterogeneous disease which is attributed to many etiological factors. The genomic studies based on NGS technologies have enhanced our knowledge about the molecular characteristics of HNSCC types and their clinical implications. In general, HNSCC could be broadly classified into HPV(+) and HPV(−) based on the HPV status. The exome sequence analyses have shown that HPV(+) tumors are different from HPV(−) tumors at molecular level. The analyses by Nichols et al. (2012) showed HPV(−) tumors to have more somatic mutations as compared to HPV(+) tumor. In contrast, the studies by Seiwert et al. (2015) noted that HPV(−) tumor has a similar mutational burden as HPV(+) tumors. However, both the studies have demonstrated distinct genomic characteristics of HPV(−) and HPV(+) tumors. Especially, the studies by Seiwert et al. (2015) showed HPV(−) tumors to harbor more mutations in TP53, CDKN2A, MLL2, CUL3, NSD1, PIK3CA, and NOTCH genes, while HPV(+) tumors had mutations in DDX3X and FGFR2/FGFR3 and abnormalities in PIK3CA, KRAS, MLL2/MLL3, and NOTCH1. The recent analyses by The Cancer Genome Atlas (TCGA) also revealed a distinct genomic alterations in HPV(−) tumors as compared to HPV(+) tumors (Cancer Genome Atlas 2015). HPV(+) tumors were noted to have recurrent deletions and truncating mutations of TRAF3. In addition HPV(+) tumors had amplifications of E2F1 and intact 9p21.3 chromosomal region, whereas HPV(−) tumors had co-amplifications of 11q13 and 11q22. Also, HPV(−) tumors had novel alterations in NSD1 and tumor suppressor genes along with recurrent amplifications of receptor tyrosine kinases. Apart from the molecular differences between HPV(+) and HPV(−) tumors, the studies utilizing genomic technologies also helped to characterize the HNSCC into various subclasses such as basal, mesenchymal, atypical, and classical (Walter et al. 2013; Cancer Genome Atlas 2015).

4.4.2 Mutational Landscapes of Head and Neck Squamous Cell Carcinoma

The exome sequencing approach was utilized to obtain a comprehensive knowledge on the underlying genetic alterations in HNSCC. The analyses by Agrawal et al. (2011) revealed the genes TP53, NOTCH1, CDKN2A, PIK3CA, FBXW7, and HRAS to be frequently mutated in their study cohort. Especially, NOTCH1 was found to be the most frequently mutated gene in the dataset. A similar exome sequence analyses by another group found 39 genes including TP53, CDKN2A, PTEN, PIK3CA, HRAS, NOTCH1, IRF6, and TP63 to be frequently mutated (Stransky et al. 2011). An integrated genomic analysis by TCGA revealed several novel genomic characteristics of HNSCC tumors (Cancer Genome Atlas 2015). In addition, the TCGA analyses also suggested many tumor suppressor genes, oncogenes, PI3- Kinases, and receptor tyrosine kinases as candidate genes for therapeutic targets in HNSCC. Another independent study focused on oral squamous cell carcinoma (OSCC) found TP53, FAT1, EPHA2, CDKN2A, NOTCH1, CASP8, HRAS, RASA, PIK3CA, CHUK, and ELAVL1 to be frequently mutated in OSCC (Su et al. 2017). Likewise, other high-throughput sequence-based studies also improved our knowledge on the mutational landscapes of HNSCC (India Project Team of the International Cancer Genome 2013; Pickering et al. 2013; Lin et al. 2014; Pickering et al. 2014). Overall, the sequence-based studies have enlightened our knowledge about the mutational landscape of HNSCC.

4.4.3 Association Between Genetic Polymorphism and Head and Neck Squamous Cell Carcinoma Risk

Earlier studies used a limited number of markers to analyze the HNSCC risk associated with individual genetics. However, the time and cost-effectiveness of NGS technologies have enabled the completion of many large population-based genomic studies such as the International HapMap project (International HapMap 2003) and

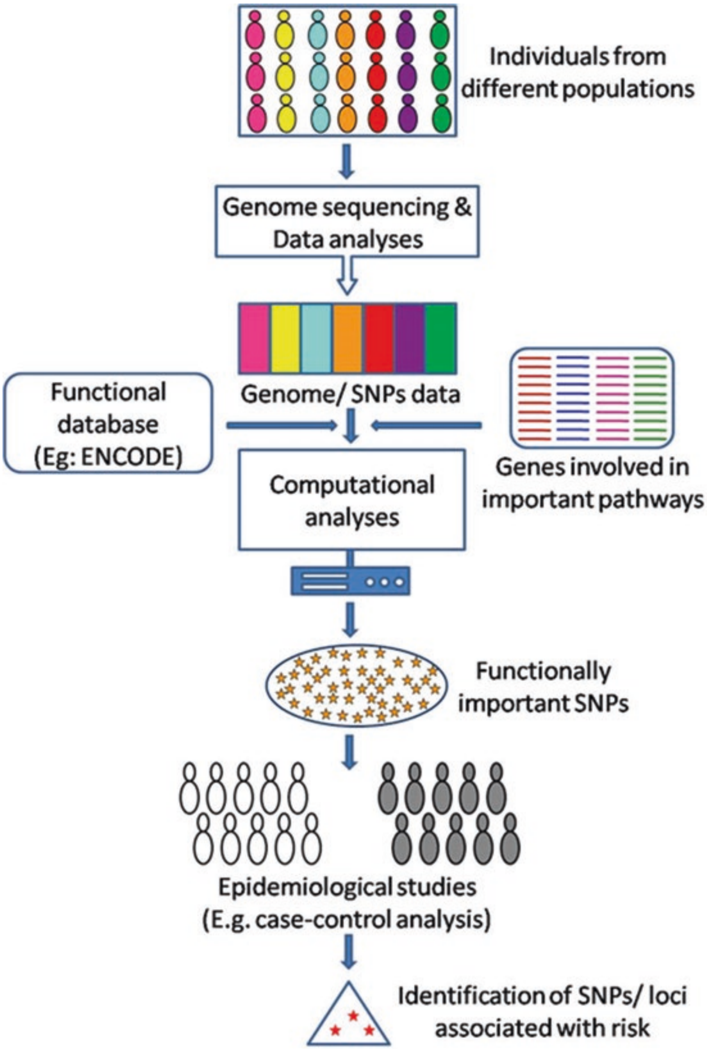

Individuals from different populations
Genome sequencing & Data analyses
Genome/ SNPs data
Functional database (Eg: ENCODE)
Genes involved in important pathways
Computational analyses
Functionally important SNPs
Epidemiological studies (E.g. case-control analysis)
Identification of SNPs/ loci associated with risk

Fig. 4.2 A schematic diagram illustrating the integrative approach to identify/select candidate markers to study genetic risk associated with disease

the 1000 Genomes Project (Auton et al. 2015), and the data of such large projects are freely available for research use. These large population-based data along with other functional datasets have helped the researchers to identify and select a comprehensive list of genetic polymorphisms in and/or around the genes involved in important pathways to evaluate the association between genetic and HNSCC risk. A schematic diagram to illustrate a bioinformatics approach to select the single-nucleotide polymorphisms (SNPs) for genetic studies is shown in Fig. 4.2.

Table 4.2 Studies reported HNSCC risk-related SNPs

Gene/locus	SNPs	Type	References
MIR548H4	rs7834169	All	Wilkins et al. (2017)
	rs16914640, rs1134367, rs7306991, rs1373756	OC	
HADH	rs221347	LA	
5p15.33	rs4975616		
KIT	rs6554198, rs2237025, rs17084687	All	Hang et al. (2017)
SOCS3	rs2280148, rs8064821	All	Hang et al. (2016)
miR-605	rs2043556	OC	Miao et al. (2016b)
miR-196a2	rs11614913		
COX-2	rs689466	All	Leng et al. (2016)
miR-101	rs578481, rs705509	OC	Miao et al. (2016a)
ERCC1	rs3212986, rs11615	All	Ding et al. (2015)
EGFR	rs12535536, rs2075110, rs1253871, rs845561, rs6970262, rs2072454	All	Fung et al. (2015)

All squamous cell carcinoma of oral cavity, larynx, and oropharynx, *OC* oral cavity, *LA* larynx

Several studies showed that germline genetic polymorphisms in genes involved in tobacco metabolism, nicotine addiction, xenobiotic metabolism, and excretion of active metabolites/carcinogens are associated with HNSCC risk (Jourenkova et al. 1998; Olshan et al. 2000; Ying et al. 2012). Similarly, the SNPs in genes involved in DNA repair and cancer oncogenesis were also found to be related to HNSCC risk (Huang et al. 2005; Al-Hadyan et al. 2012; Zhang et al. 2013). The list of SNPs that were found to be associated with HNSCC risk in some of the recent literatures (from 2015) is given in Table 4.2. These studies suggest the important role of genetics in HNSCC. The Genome Wide Association Studies (GWAS) in HNSCC have also identified several genetic loci to be associated with HNSCC risk (Wei et al. 2014; Lesseur et al. 2016). Although the exact mechanism of action of these SNPs in HNSCC development are not known, recent analyses showed that these germline polymorphisms could act as expression quantitative trait loci (eQTLs) and affect expression of genes thereby could be associated with HNSCC progression or survival (Hang et al. 2017; Ramakodi et al. 2017).

4.4.3.1 Genetic Association Study: A Case-Control Analysis

The genetic association studies are vital in healthcare research to identify the genetic basis and risk associated with a disease. Many epidemiological approaches including prospective, retrospective, and case-control analysis are followed to conduct the genetic association study. Among those various epidemiological methods, case-control approach is being widely used. In this section, we present an example of basic workflow involved in a case-control study using hypothetical genotype data.

The first step is to identify genes of interest, which could be done through literature survey or experimental procedure. Subsequently, putative functional SNPs in the gene of interest need to be identified using public database or custom array techniques. In this example, we have taken TP53 gene that is involved in DNA-damage repair mechanism and is one of the important cancer drivers. Several mutations and/or polymorphisms in TP53 are associated with various types of cancer. The steps involved in identifying the putative functional SNPs in TP53 gene and executing the genetic association study are presented as follows:

(a) Identification of Single-Nucleotide Polymorphisms in TP53

Primarily, TP53 gene was searched on the 1000 Genomes Project web browser available at https://www.ncbi.nlm.nih.gov/variation/tools/1000genomes/, and its genetic polymorphism in different populations was discerned. The search results showed TP53 gene located in chromosome 17 (position: 7,571,720–7,590,868) to have 622 polymorphisms. The rs IDs of these 622 polymorphisms were obtained for functional analysis. Here, it should be noted that using the 1000 Genomes Project Browser is optional. Alternatively, one can download the entire genotype dataset for all the populations and obtain the population-wise SNPs information for many genes computationally.

(b) Identification of Putative Functional Single-Nucleotide Polymorphisms in TP53

The rs IDs of 622 SNPs present in TP53 were searched on the web interface tool Variant Effect Predictor (VEP) available at http://grch37.ensembl.org/Tools/VEP (McLaren et al. 2016) to classify the functional characteristics of these SNPs. The part of the result as obtained from VEP is shown in Fig. 4.3. Alternative to web interface, standalone VEP software is also available, and usage of this will be efficient while analyzing large datasets. The web-based VEP classified the SNPs as intron variant, UTR variant, missense, stop_gained, synonymous, etc. Based on this information, one could select the SNPs of their interest for further study. This computational-based method has been utilized effectively to identify the causal SNPs associated with disease in minimal cost and time. For this example analysis, we ascertained 53 SNPs classified as "upstream_gene_variant," considering the fact that an upstream variant could be involved in regulation of gene expression and

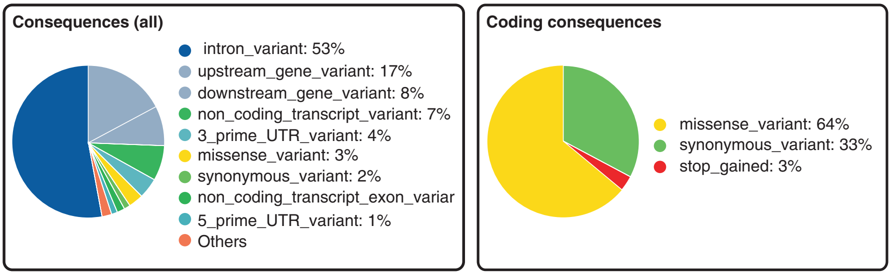


Fig. 4.3 Functional analysis results for SNPs present in *TP53* as obtained from Variant Effect Predictor (VEP)∗. ∗Only a part of the results obtained from VEP is shown in the figure

possibly could act as eQTL. Among these 53 SNPs, rs2287499 was selected as a marker for the case-control analysis as it is an upstream variant for TP53 gene and also a missense variant for WRAP53 gene. Thus, we could frame a hypothesis that this SNP is associated with HNSCC risk and will be statistically tested in the following case-control analysis.

(c) Case-Control Analysis

As stated before, for this analysis, we will use hypothetical genotype data of the TP53 gene variant rs2287499. However, it is imperative to know the methodology of generating the genotype data for desired analysis. The first step is to identify the HNSCC cases and healthy controls following the epidemiological principles. Accordingly, for our analysis, let us assume that we have 500 HNSCC cases and 500 controls. Then, biological samples like blood and/or saliva were collected from cases and controls to extract DNA and perform genotyping for rs2287499. Assume that all the genotypes passed quality-control evaluation. As the TP53 gene variant rs2287499 has two alleles "C" and "G," the following three possible genotypes CC, CG, and GG can be observed in the cases and controls. We generated hypothetical genotype counts for cases and controls for further statistical investigation. This genotype data is presented in Table 4.3.

To measure the disease risk associated with exposure in case-control analysis, we calculated odds ratio applying logistic regression model. As a result, GG genotype was found to be significant based on the *p*-value 1.44e-08 with the odds ratio of 2.37. This indicates the association of rs2287499 with HNSCC risk; especially, the individuals carrying GG genotype are at high risk of HNSCC by twofold as compared to others. The *p*-values and odds ratio of each genotype are tabulated and shown in Table 4.4.

It is to be noted that the above significant observations are subject to change when we consider the confounding factors such as age, sex, population, and the like, involved in a case-control study. The appropriate confounder should be adjusted when performing the logistic regression analysis to identify the true effect of the variants under study. As we dealt with the hypothetical data in this example analysis, adjusting for confounders are not shown. We suggest further reading of epidemiological principles and statistical analysis-related literature for in-depth understanding of case-control studies and appropriate statistical calculations.

Table 4.3 Hypothetical genotype data of rs2287499 for cases and controls

Genotype	Case	Control	Total
CC	162	220	382
CG	118	154	272
GG	220	126	346
Total	500	500	1000

Table 4.4 Results of odds ratio calculation following logistic regression approach. The results indicate that the genotype GG is associated with increased risk as compared to other genotypes

Genotype	Odds ratio (OR)	95% Confidence interval		*p*-value
CC	1.00	–	–	–
CG	1.04	0.76	1.42	0.804
GG	2.37	1.76	3.19	<0.001

4.4.4 Genetics and Head and Neck Squamous Cell Carcinoma Survival

The NGS technology is also utilized to unveil the genetics associated with HNSCC survival. Liu and colleagues (Liu et al. 2016) investigated the effect of somatic mutations and genetic variants of NOTCH1 on HNSCC occurrence and development using exome sequencing approach. The study revealed that patients with somatic mutations in NOTCH1 had higher 5-year relapse-free recurrence and lower survival proportions. Another exome sequence analyses showed the amplification of PIK3CA and mutations in RAS to be associated with poorer prognosis (Chau et al. 2016). Also, an integrative genomic analysis using the data generated from exome sequences along with other functional datasets identified several eQTLs and enabled to understand how the genetics could be associated with HNSCC survival (Ramakodi et al. 2017).

4.4.5 Genetics of Head and Neck Squamous Cell Carcinoma Disparity

The HNSCC incidence and survival rates differ among different populations. For example, African Americans (Afr-Amr) have higher incidence and lower survival rates as compared to Caucasian Americans (Cau-Amr) (Walker et al. 1995; Gourin and Podolsky 2006; Jiron et al. 2014). Interestingly, the HNSCC genetics differ between Afr-Amr and Cau-Amr patients. In addition, recent studies based on data derived from exome sequences from TCGA suggest that genetics could be involved in the HNSCC disparity observed between Afr-Amr and Cau-Amr. The mutational landscape analyses of laryngeal cancer showed different mutation burdens between Afr-Amr and Cau-Amr patients (Ramakodi et al. 2016). In addition, the significantly mutated genes were found to be different in Afr-Amr as compared to Eur-Amr patients. For example, PIK3CA, one of the important driver gene, was significantly differently mutated between Afr-Amr and Cau-Amr patients. The exome sequence data was also used to understand the functional importance of genetics in HNSCC and to uncover the association between ancestral genetics and HNSCC disparity. The functional analyses by Ramakodi et al. (2017) have identified many eQTLs, and their study explained the effect of population-specific allele on HNSCC survival disparity. Thus, the exome sequence data helped to uncover how genetic ancestry could be associated with increased HNSCC risk/lower HNSCC survival in Afr-Amr.

4.5 Conclusion

The NGS technologies and the algorithms to analyze the sequence data are continuously evolving. Also, the time and cost of sequencing the genomes are currently coming down. In addition, many web-based bioinformatics platforms such as Galaxy (https://usegalaxy.org/) are readily available to analyze the large NGS dataset for research purpose. Most importantly, today several online blogs are available to get clarifications or suggestions on NGS-related questions. These rapid developments of NGS technologies and advancements in bioinformatics simplified the use of NGS in clinical medicine and other scientific area. Indeed, the sequence-based analyses have improved our knowledge about the genetics of various types of cancer including HNSCC. The sequence data helped the researchers to understand the functionally important genetic factors in cancers. The sequence-based analyses also elucidated the important pathways involved in disease development and progression and helped to identify therapeutic targets to be used in precision medicine. Today, the high-throughput sequencing technologies have been adopted for personalized medicine in the developed countries, but the NGS technologies are not often used for personalized medicine in the developing or underdeveloped countries. Nonetheless, the continuing decrease in the cost of NGS technologies and the improvements of web-based analyses tools will benefit the developing and underdeveloped countries to use NGS technologies in personalized medicine to improve the quality of life. In summary, the NGS technologies play an important role in clinical medicine and hold a broad and promising future in medical discipline.

Glossary

BAM Binary Alignment Map, a compressed binary format for storing large nucleotide sequence alignments.

FASTQ The text-based format for storing both a DNA sequence and its corresponding quality scores.

Paired-end This sequencing procedure involves sequencing both the ends of the DNA fragments in a library and aligning the forward and reverse reads as read pairs.

Phred Q scores The base calling converts the signals into actual sequence data with this quality scores.

Read The WGS or WES procedure involves shearing DNA into hundreds of thousands of small fragments, and every single fragment is called a "read."

Read depth The average number of times that a given nucleotide in the genome has been read in a sequencing experiment. For instance, a 40× read depth means that each base is present in an average of 40 reads.

SAM Sequence Alignment Map, a genetic format for storing large nucleotide sequence alignments.

Single-read This sequencing procedure involves sequencing DNA from only one end.

TAB The text-based tab-delimited file format.

VCF Variant Calling Format, a text file format containing meta-information lines, a header line, and then data lines, each containing information about a position in the genome.

References

Adzhubei IA, Schmidt S, Peshkin L, Ramensky VE, Gerasimova A, Bork P et al (2010) A method and server for predicting damaging missense mutations. Nat Methods 7(4):248–249

Agrawal N, Frederick MJ, Pickering CR, Bettegowda C, Chang K, Li RJ et al (2011) Exome sequencing of head and neck squamous cell carcinoma reveals inactivating mutations in NOTCH1. Science 333(6046):1154–1157

Al-Hadyan KS, Al-Harbi NM, Al-Qahtani SS, Alsbeih GA (2012) Involvement of single-nucleotide polymorphisms in predisposition to head and neck cancer in Saudi Arabia. Genet Test Mol Biomarkers 16(2):95–101

Andrews S. (2010). FastQC: a quality control tool for high throughput sequence data. Available online at: http://www.bioinformatics.babraham.ac.uk/projects/fastqc

Auton A, Brooks LD, Durbin RM, Garrison EP, Kang HM, Korbel JO et al (2015) A global reference for human genetic variation. Nature 526(7571):68–74

Bolger AM, Lohse M, Usadel B (2014) Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics 30(15):2114–2120

Cabanski CR, Cavin K, Bizon C, Wilkerson MD, Parker JS, Wilhelmsen KC et al (2012) ReQON: a Bioconductor package for recalibrating quality scores from next-generation sequencing data. BMC Bioinformatics 13:221

Cai L, Yuan W, Zhang Z, He L, Chou K-C (2016) In-depth comparison of somatic point mutation callers based on different tumor next-generation sequencing depth data. Sci Rep 6:36540

Cancer Genome Atlas N (2015) Comprehensive genomic characterization of head and neck squamous cell carcinomas. Nature 517(7536):576–582

Carneiro MO, Russ C, Ross MG, Gabriel SB, Nusbaum C, DePristo MA (2012) Pacific biosciences sequencing technology for genotyping and variation discovery in human data. BMC Genomics 13:375

Challis D, Yu J, Evani US, Jackson AR, Paithankar S, Coarfa C et al (2012) An integrative variant analysis suite for whole exome next-generation sequencing data. BMC Bioinformatics 13:8

Chau NG, Li YY, Jo VY, Rabinowits G, Lorch JH, Tishler RB et al (2016) Incorporation of next-generation sequencing into routine clinical care to direct treatment of head and neck squamous cell carcinoma. Clin Cancer Res 22(12):2939–2949

Cingolani P, Platts A, Wang LL, Coon M, Nguyen T, Wang L et al (2012) 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

Clarke J, Wu H-C, Jayasinghe L, Patel A, Reid S, Bayley H (2009) Continuous base identification for single-molecule nanopore DNA sequencing. Nat Nanotechnol 4(4):265–270

Cock PJA, Fields CJ, Goto N, Heuer ML, Rice PM (2010) The Sanger FASTQ file format for sequences with quality scores, and the Solexa/Illumina FASTQ variants. Nucleic Acids Res 38(6):1767–1771

Collins FS, Varmus H (2015) A new initiative on precision medicine. N Engl J Med 372(9):793–795

Cox MP, Peterson DA, Biggs PJ (2010) SolexaQA: at-a-glance quality assessment of Illumina second-generation sequencing data. BMC Bioinformatics 11:485

Dai M, Thompson RC, Maher C, Contreras-Galindo R, Kaplan MH, Markovitz DM et al (2010) NGSQC: cross-platform quality analysis pipeline for deep sequencing data. BMC Genomics 11(Suppl 4):S7

Dewey FE, Grove ME, Pan C, Goldstein BA, Bernstein JA, Chaib H et al (2014) Clinical interpretation and implications of whole-genome sequencing. JAMA 311(10):1035–1045

Ding YW, Gao X, Ye DX, Liu W, Wu L, Sun HY (2015) Association of ERCC1 polymorphisms (rs3212986 and rs11615) with the risk of head and neck carcinomas based on case-control studies. Clin Transl Oncol 17(9):710–719

Dohm JC, Lottaz C, Borodina T, Himmelbauer H (2008) Substantial biases in ultra-short read data sets from high-throughput DNA sequencing. Nucleic Acids Res 36(16):e105

Eilbeck K, Quinlan A, Yandell M (2017) Settling the score: variant prioritization and Mendelian disease. Nat Rev Genet 18(10):599–612

Foley SB, Rios JJ, Mgbemena VE, Robinson LS, Hampel HL, Toland AE et al (2015) Use of whole genome sequencing for diagnosis and discovery in the cancer genetics clinic. EBioMedicine 2(1):74–81

Fung C, Zhou P, Joyce S, Trent K, Yuan JM, Grandis JR et al (2015) Identification of epidermal growth factor receptor (EGFR) genetic variants that modify risk for head and neck squamous cell carcinoma. Cancer Lett 357(2):549–556

Garrison E, Marth G (2012) Haplotype-based variant detection from short-read sequencing [Internet]. arXiv [q-bio.GN]. Available from: http://arxiv.org/abs/1207.3907

Gilles A, Meglécz E, Pech N, Ferreira S, Malausa T, Martin J-F (2011) Accuracy and quality assessment of 454 GS-FLX Titanium pyrosequencing. BMC Genomics 12:245

Glenn TC (2011) Field guide to next-generation DNA sequencers. Mol Ecol Resour 11(5):759–769

Goodwin S, McPherson JD, McCombie WR (2016) Coming of age: ten years of next-generation sequencing technologies. Nat Rev Genet 17(6):333–351

Gourin CG, Podolsky RH (2006) Racial disparities in patients with head and neck squamous cell carcinoma. Laryngoscope 116(7):1093–1106

Guo Y, Zhao S, Sheng Q, Ye F, Li J, Lehmann B et al (2014) Multi-perspective quality control of Illumina exome sequencing data using QC3. Genomics 103(5–6):323–328

Hang D, Yin Y, Wang L, Yuan H, Du J, Zhu M et al (2016) Effects of potentially functional polymorphisms in suppressor of cytokine signaling 3 (SOCS3) on the risk of head and neck squamous cancer. J Oral Pathol Med 46(8):598–602

Hang D, Yuan H, Liu L, Wang L, Miao L, Zhu M et al (2017) KIT polymorphisms were associated with the risk for head and neck squamous carcinoma in Chinese population. Mol Carcinog 56(1):232–237

Heather JM, Chain B (2016) The sequence of sequencers: the history of sequencing DNA. Genomics 107(1):1–8

Hegde M, Santani A, Mao R, Ferreira-Gonzalez A, Weck KE, Voelkerding KV (2017) Development and validation of clinical whole-exome and whole-genome sequencing for detection of germline variants in inherited disease. Arch Pathol Lab Med 141(6):798–805

Homer N, Nelson SF (2010) Improved variant discovery through local re-alignment of short-read next-generation sequencing data using SRMA. Genome Biol 11(10):R99

Hu H, Huff CD, Moore B, Flygare S (2013) VAAST 2.0: improved variant classification and disease-gene identification using a conservation-controlled amino acid substitution matrix. Genetic [Internet] Wiley Online Library. Available from: http://onlinelibrary.wiley.com/doi/10.1002/gepi.21743/full

Huang WY, Olshan AF, Schwartz SM, Berndt SI, Chen C, Llaca V et al (2005) Selected genetic polymorphisms in MGMT, XRCC1, XPD, and XRCC3 and risk of head and neck cancer: a pooled analysis. Cancer Epidemiol Biomark Prev 14(7):1747–1753

Hwang S, Kim E, Lee I, Marcotte EM (2015) Systematic comparison of variant calling pipelines using gold standard personal exome variants. Sci Rep 5:17875

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

International HapMap C (2003) The international HapMap project. Nature 426(6968):789–796

Jain M, Olsen HE, Paten B, Akeson M (2016) The Oxford Nanopore MinION: delivery of nanopore sequencing to the genomics community. Genome Biol 17(1):239

Jain M, Koren S, Miga KH, Quick J, Rand AC, Sasani TA et al (2018) Nanopore sequencing and assembly of a human genome with ultra-long reads. Nat Biotechnol [Internet] 29:338; Available from. https://doi.org/10.1038/nbt.4060

Jalali Sefid Dashti M, Gamieldien J (2017) A practical guide to filtering and prioritizing genetic variants. BioTechniques 62(1):18–30

Jiron J, Sethi S, Ali-Fehmi R, Franceschi S, Struijk L, van Doorn LJ et al (2014) Racial disparities in Human Papillomavirus (HPV) associated head and neck cancer. Am J Otolaryngol 35(2):147–153

Jourenkova N, Reinikainen M, Bouchardy C, Dayer P, Benhamou S, Hirvonen A (1998) Larynx cancer risk in relation to glutathione S-transferase M1 and T1 genotypes and tobacco smoking. Cancer Epidemiol Biomark Prev 7(1):19–23

Kim S, Jeong K, Bhutani K, Lee J, Patel A, Scott E et al (2013) Virmid: accurate detection of somatic mutations with sample impurity inference. Genome Biol 14(8):R90

Kircher M, Witten DM, Jain P, O'Roak BJ, Cooper GM, Shendure J (2014) A general framework for estimating the relative pathogenicity of human genetic variants. Nat Genet 46(3):310–315

Koboldt DC, Steinberg KM, Larson DE, Wilson RK, Mardis ER (2013) The next-generation sequencing revolution and its impact on genomics. Cell 155(1):27–38

Koren S, Schatz MC, Walenz BP, Martin J, Howard JT, Ganapathy G et al (2012) Hybrid error correction and de novo assembly of single-molecule sequencing reads. Nat Biotechnol 30(7):693–700

Krøigård AB, Thomassen M, Lænkholm A-V, Kruse TA, Larsen MJ (2016) Evaluation of nine somatic variant callers for detection of somatic mutations in exome and targeted deep sequencing data. PLoS One 11(3):e0151664

Kumar P, Henikoff S, Ng PC (2009) Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm. Nat Protoc 4(7):1073–1081

Landrum MJ, Lee JM, Riley GR, Jang W, Rubinstein WS, Church DM et al (2014) ClinVar: public archive of relationships among sequence variation and human phenotype. Nucleic Acids Res 42(Database issue):D980–D985

Larson DE, Harris CC, Chen K, Koboldt DC, Abbott TE, Dooling DJ et al (2012) SomaticSniper: identification of somatic point mutations in whole genome sequencing data. Bioinformatics 28(3):311–317

Lee H, Deignan JL, Dorrani N, Strom SP, Kantarci S, Quintero-Rivera F et al (2014a) Clinical exome sequencing for genetic identification of rare Mendelian disorders. JAMA 312(18):1880–1887

Lee I-H, Lee K, Hsing M, Choe Y, Park J-H, Kim SH et al (2014b) Prioritizing disease-linked variants, genes, and pathways with an interactive whole-genome analysis pipeline. Hum Mutat 35(5):537–547

Lek M, Karczewski KJ, Minikel EV, Samocha KE, Banks E, Fennell T et al (2016) Analysis of protein-coding genetic variation in 60,706 humans. Nature 536(7616):285–291

Leng WD, Wen XJ, Kwong JS, Huang W, Chen JG, Zeng XT (2016) COX-2 rs689466, rs5275, and rs20417 polymorphisms and risk of head and neck squamous cell carcinoma: a meta-analysis of adjusted and unadjusted data. BMC Cancer 16:457

Lesseur C, Diergaarde B, Olshan AF, Wunsch-Filho V, Ness AR, Liu G et al (2016) Genome-wide association analyses identify new susceptibility loci for oral cavity and pharyngeal cancer. Nat Genet 48(12):1544–1550

Li H (2016) Cancer precision medicine in China. Genomics Proteomics Bioinformatics 14(5):325–328

Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N et al (2009) The sequence alignment/map format and SAMtools. Bioinformatics 25(16):2078–2079

Li M-X, Gui H-S, Kwan JSH, Bao S-Y, Sham PC (2012) A comprehensive framework for prioritizing variants in exome sequencing studies of Mendelian diseases. Nucleic Acids Res 40(7):e53

Lin DC, Meng X, Hazawa M, Nagata Y, Varela AM, Xu L et al (2014) The genomic landscape of nasopharyngeal carcinoma. Nat Genet 46(8):866–871

Lionel AC, Costain G, Monfared N, Walker S, Reuter MS, Hosseini SM et al (2017) Improved diagnostic yield compared with targeted gene sequencing panels suggests a role for whole-genome sequencing as a first-tier genetic test. Genet Med [Internet] 3:435; Available from. https://doi.org/10.1038/gim.2017.119

Liu YF, Chiang SL, Lin CY, Chang JG, Chung CM, Ko AM et al (2016) Somatic mutations and genetic variants of NOTCH1 in head and neck squamous cell carcinoma occurrence and development. Sci Rep 6:24014

Lu H, Giordano F, Ning Z (2016) Oxford Nanopore MinION sequencing and genome assembly. Genomics Proteomics Bioinformatics 14(5):265–279

Madoui M-A, Engelen S, Cruaud C, Belser C, Bertrand L, Alberti A et al (2015) Genome assembly using Nanopore-guided long and error-free DNA reads. BMC Genomics 16:327

Martin M (2011) Cutadapt removes adapter sequences from high-throughput sequencing reads. EMBnet.journal 17(1):10–12

McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A et al (2010) The genome analysis toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res 20(9):1297–1303

McLaren W, Gil L, Hunt SE, Riat HS, Ritchie GRS, Thormann A et al (2016) The ensembl variant effect predictor. Genome Biol 17(1):122

Metsky HC, Matranga CB, Wohl S, Schaffner SF (2017) Genome sequencing reveals Zika virus diversity and spread in the Americas. bioRxiv [Internet]. biorxiv.org. Available from https://www.biorxiv.org/content/early/2017/04/23/109348.abstract

Miao L, Wang L, Yuan H, Hang D, Zhu L, Du J et al (2016a) MicroRNA-101 polymorphisms and risk of head and neck squamous cell carcinoma in a Chinese population. Tumour Biol 37(3):4169–4174

Miao L, Wang L, Zhu L, Du J, Zhu X, Niu Y et al (2016b) Association of microRNA polymorphisms with the risk of head and neck squamous cell carcinoma in a Chinese population: a case-control study. Chin J Cancer 35(1):77

Müllauer L (2017) Next generation sequencing: clinical applications in solid tumours. Memo 10(4):244–247

Nichols AC, Chan-Seng-Yue M, Yoo J, Xu W, Dhaliwal S, Basmaji J et al (2012) A pilot study comparing HPV-positive and HPV-negative head and neck squamous cell carcinomas by whole exome sequencing. ISRN Oncol 2012:809370

O'Rawe J, Jiang T, Sun G, Wu Y, Wang W, Hu J et al (2013) Low concordance of multiple variant-calling pipelines: practical implications for exome and genome sequencing. Genome Med 5(3):28

Oliver GR, Hart SN, Klee EW (2015) Bioinformatics for clinical next generation sequencing. Clin Chem 61(1):124–135

Olshan AF, Weissler MC, Watson MA, Bell DA (2000) GSTM1, GSTT1, GSTP1, CYP1A1, and NAT1 polymorphisms, tobacco use, and the risk of head and neck cancer. Cancer Epidemiol Biomark Prev 9(2):185–191

Pickering CR, Zhang J, Yoo SY, Bengtsson L, Moorthy S, Neskey DM et al (2013) Integrative genomic characterization of oral squamous cell carcinoma identifies frequent somatic drivers. Cancer Discov 3(7):770–781

Pickering CR, Zhang J, Neskey DM, Zhao M, Jasser SA, Wang J et al (2014) Squamous cell carcinoma of the oral tongue in young non-smokers is genomically similar to tumors in older smokers. Clin Cancer Res 20(14):3842–3848

Posey JE, Rosenfeld JA, James RA, Bainbridge M, Niu Z, Wang X et al (2016) Molecular diagnostic experience of whole-exome sequencing in adult patients. Genet Med 18(7):678–685

Quail MA, Smith M, Coupland P, Otto TD, Harris SR, Connor TR et al (2012) A tale of three next generation sequencing platforms: comparison of Ion Torrent, Pacific Biosciences and Illumina MiSeq sequencers. BMC Genomics 13:341

Ragin CCR, Modugno F, Gollin SM (2007) The epidemiology and risk factors of head and neck cancer: a focus on human papillomavirus. J Dent Res 86(2):104–114

Ramakodi MP, Kulathinal RJ, Chung Y, Serebriiskii I, Liu JC, Ragin CC (2016) Ancestral-derived effects on the mutational landscape of laryngeal cancer. Genomics 107(2–3):76–82

Ramakodi MP, Devarajan K, Blackman E, Gibbs D, Luce D, Deloumeaux J et al (2017) Integrative genomic analysis identifies ancestry-related expression quantitative trait loci on DNA polymerase beta and supports the association of genetic ancestry with survival disparities in head and neck squamous cell carcinoma. Cancer 123(5):849–860

Reuter MS, Walker S, Thiruvahindrapuram B, Whitney J, Cohn I, Sondheimer N et al (2018) The Personal Genome Project Canada: findings from whole genome sequences of the inaugural 56 participants. CMAJ 190(5):E126–E136

Ross MG, Russ C, Costello M, Hollinger A, Lennon NJ, Hegarty R et al (2013) Characterizing and measuring bias in sequence data. Genome Biol 14(5):R51

Salmela L, Rivals E (2014) LoRDEC: accurate and efficient long read error correction. Bioinformatics 30(24):3506–3514

Sandmann S, de Graaf AO, Karimi M, van der Reijden BA, Hellström-Lindberg E, Jansen JH et al (2017) Evaluating variant calling tools for non-matched next-generation sequencing data. Sci Rep 7:43169

Saunders CT, Wong WSW, Swamy S, Becq J, Murray LJ, Cheetham RK (2012) Strelka: accurate somatic small-variant calling from sequenced tumor-normal sample pairs. Bioinformatics 28(14):1811–1817

Schmieder R, Edwards R (2011) Quality control and preprocessing of metagenomic datasets. Bioinformatics 27(6):863–864

Seiwert TY, Zuo Z, Keck MK, Khattri A, Pedamallu CS, Stricker T et al (2015) Integrative and comparative genomic analysis of HPV-positive and HPV-negative head and neck squamous cell carcinomas. Clin Cancer Res 21(3):632–641

Sherry ST, Ward MH, Kholodov M, Baker J, Phan L, Smigielski EM et al (2001) dbSNP: the NCBI database of genetic variation. Nucleic Acids Res 29(1):308–311

Sims D, Sudbery I, Ilott NE, Heger A, Ponting CP (2014) Sequencing depth and coverage: key considerations in genomic analyses. Nat Rev Genet 15(2):121–132

Stavropoulos DJ, Merico D, Jobling R, Bowdin S, Monfared N, Thiruvahindrapuram B et al (2016) Whole-genome sequencing expands diagnostic utility and improves clinical management in paediatric medicine. NPJ Genom Med. The Author(s); 1:15012

Stransky N, Egloff AM, Tward AD, Kostic AD, Cibulskis K, Sivachenko A et al (2011) The mutational landscape of head and neck squamous cell carcinoma. Science 333(6046):1157–1160

Su SC, Lin CW, Liu YF, Fan WL, Chen MK, Yu CP et al (2017) Exome sequencing of Oral squamous cell carcinoma reveals molecular subgroups and novel therapeutic opportunities. Theranostics 7(5):1088–1099

Taylor JC, Martin HC, Lise S, Broxholme J, Cazier J-B, Rimmer A et al (2015) Factors influencing success of clinical genome sequencing across a broad spectrum of disorders. Nat Genet 47(7):717–726

Teer JK, Green ED, Mullikin JC, Biesecker LG (2012) VarSifter: visualizing and analyzing exome-scale sequence variation data on a desktop computer. Bioinformatics 28(4):599–600

UK10K Consortium, Walter K, Min JL, Huang J, Crooks L, Memari Y et al (2015) The UK10K project identifies rare variants in health and disease. Nature 526(7571):82–90

Walker B, Figgs LW, Zahm SH (1995) Differences in cancer incidence, mortality, and survival between African Americans and whites. Environ Health Perspect 103(Suppl 8):275–281

Walter V, Yin X, Wilkerson MD, Cabanski CR, Zhao N, Du Y et al (2013) Molecular subtypes in head and neck cancer exhibit distinct patterns of chromosomal gain and loss of canonical cancer genes. PLoS One 8(2):e56823

Wang K, Li M, Hakonarson H (2010) ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. Nucleic Acids Res 38(16):e164

Wei Q, Yu D, Liu M, Wang M, Zhao M, Liu M et al (2014) Genome-wide association study identifies three susceptibility loci for laryngeal squamous cell carcinoma in the Chinese population. Nat Genet 46(10):1110–1114

Whirl-Carrillo M, McDonagh EM, Hebert JM, Gong L, Sangkuhl K, Thorn CF et al (2012) Pharmacogenomics knowledge for personalized medicine. Clin Pharmacol Ther 92(4):414–417

Wilkins OM, Titus AJ, Gui J, Eliot M, Butler RA, Sturgis EM et al (2017) Genome-scale identification of microRNA-related SNPs associated with risk of head and neck squamous cell carcinoma. Carcinogenesis 38:986

Wu TD, Nacu S (2010) Fast and SNP-tolerant detection of complex variants and splicing in short reads. Bioinformatics 26(7):873–881

Wu J, Li Y, Jiang R (2014) Integrating multiple genomic data to predict disease-causing nonsynonymous single nucleotide variants in exome sequencing studies. PLoS Genet 10(3):e1004237

Ying XJ, Dong P, Shen B, Xu CZ, Xu HM, Zhao SW (2012) Glutathione S-transferase M1 gene polymorphism and laryngeal cancer risk: a meta-analysis. PLoS One 7(8):e42826

Zhang E, Cui Z, Xu Z, Duan W, Huang S, Tan X et al (2013) Association between polymorphisms in ERCC2 gene and oral cancer risk: evidence from a meta-analysis. BMC Cancer 13:594

Chapter 5
Chromosomal Microarray in the New High-Throughput Technological and Bioinformatic Era

Susan Mathew

Contents

Abbreviations

aCGH	Array comparative genomic hybridization
ASD	Autism spectrum disorders
CMA	Chromosomal microarray analysis
CNV	Copy number variants
DD	Developmental delay
FISH	Fluorescence in situ hybridization
ID	Intellectual disability
LCSH	Long contiguous stretches of homozygosity
LOH	Loss of heterozygosity
MCA	Multiple congenital anomalies
NGS	Next-generation sequencing
VISL	Variants in susceptibility loci
VOUS	Variants of uncertain significance

Prof. S. Mathew (✉)
Cytogenetics Laboratory, Department of Pathology and Laboratory Medicine,
Weill Cornell Medicine, New York, NY, USA
e-mail: sum2001@med.cornell.edu

N. A. Shaik et al. (eds.), *Essentials of Bioinformatics, Volume II*,
https://doi.org/10.1007/978-3-030-18375-2_5

5.1 Introduction

The introduction of microarray technology in mid-1990s allowed scientists to profile and analyze the human genome simultaneously. Since then, a number of microarray platforms allowing high volume automated analysis of DNA, RNA, and protein on a microchip-based testing platform have evolved. DNA microarray reveals genetic imbalances involving many genes and markers at multiple regions of the genome in all chromosomes.

Conventional cytogenetics through karyotyping has been the gold standard for detecting structural and numerical chromosomal abnormalities including losses, gains, inversions, deletions, duplications, and translocations in prenatal diagnosis as well as in postnatal diagnosis (in individuals with dysmorphic features, mental retardation), in products of conceptions and in cancer. Chromosomal microarray analysis (CMA) is a whole genome high-resolution genetic test that can identify chromosomal abnormalities that cannot be detected by conventional karyotyping and fluorescence in situ hybridization (FISH) assays. CMA not only detects abnormalities that are detected by conventional cytogenetics but also reveals microdeletions and microduplications. This novel technology was termed "molecular karyotyping" (Rauch et al. 2004; Vermeesch et al. 2007). The resolution of conventional cytogenetics is about 5–10 Mb in size whereas CMA has enhanced the ability to detect genome-wide imbalances in <1 kb, demonstrating its advantage over karyotyping and FISH.

CMA is a powerful diagnostic tool for prenatal, postnatal, and cancer diagnosis. CMA offers a much higher diagnostic yield (15%–20%) for genetic testing of individuals with unexplained developmental delay (DD), intellectual disability (ID), autism spectrum disorders (ASD), and multiple congenital anomalies (MCA) than conventional cytogenetics excluding Down syndrome and other known cytogenetics syndromes (Shaffer et al. 2006; Sagoo et al. 2009, Cooper et al. 2011; Kaminsky et al. 2011; Mefford et al. 2012). This has led to the recommendation to use CMA as the first-tier testing for children with DD, ID, ASD, and MCA (Manning and Hudgins 2010; Miller et al. 2010). In cancer diagnostics, earlier studies focused on hematological malignancies; however, much progress has been made recently for solid neoplasms. An overview of the microarray technology and its applications in prenatal, postnatal, and cancer diagnostics will be discussed in this chapter.

5.2 Microarray Technology

Currently, different types of DNA-microarray platforms are available. CMA can detect microdeletions and microduplications of chromosome segments (referred to as copy number variants, CNV) which are too small to be visible by conventional karyotyping. There are two major microarray platforms used for identifying the

CMAs: array comparative genomic hybridization (aCGH) and single nucleotide polymorphism (SNP) arrays. In aCGH, DNA from the patient is compared directly or indirectly with a reference genome (normal DNA). For aCGH, both the reference and patient DNAs are labeled with different fluorochromes and hybridized to multiple probes representing sequences across the genome on the microarray. The findings are reported as a signal ratio in the two color assays. The aCGH consists of CNV probes and only provides CNV information.

SNP-based arrays use SNP probes in a single color dye and instead of using a control sample as a reference in every run, reference intensity data from a population of normal samples is used as a reference for the patient sample—in silico reference (Coughlin et al. 2012). SNP arrays provide both SNP genotype and CNV information. The CNV information is generated using the signal intensity of the probes (Wang et al. 2005; Carter 2007; Gresham et al. 2008). The signal intensities of the fluorochromes are captured by a scanner. Loss and gains of genomic regions are compared with the reference DNA by the differences in the signal intensities of the probes. If there is no loss or gain, the expected copy number is "0" and the copy number is expressed as a $\log_2$ ratio. The log ratios for duplication will be >0 and for deletions will be <0. SNP arrays allow simultaneous detection of DNA copy number changes and absence of heterozygosity (AOH) due to loss of heterozygosity (LOH), hemizygosity, or homozygosity. The combined use of CNV and SNP probes is ideal for maximum coverage and high resolution in detection of these variants (Carter 2007). Both aCGH and SNP arrays detect CNVs, whereas SNP array can detect triploidy, uniparental disomy, mosaicism >25%, maternal cell contamination, parent of origin, and consanguinity, but aCGH cannot detect triploidy. Arrays using oligonucleotides, oligonucleotide plus single nucleotide polymorphism (SNP), and SNP are the most commonly available arrays (Agilent technologies, Affymetrix, Ilumina, etc.).

Analysis and Reporting It is important to understand the terminology used in microarray testing and is imperative how to interpret the findings obtained from microarray. A CNV is defined as a segment of DNA at least 1 kb in size that differs in copy number compared with a representative reference genome. Interpretation of CNVs should be provided as clearly as possible. CNVs can be benign or pathogenic depending on clinical relevance and can be interpreted as duplication or deletion to clarify the nature of the CNV. American College of Medical Genetics (ACMG) guidelines help in promoting consistency in interpretation and reporting (Cooley et al. 2013). Deletions less than 200 kb and duplications less than 400 kb are not reported, unless they involve regions of the genome with clear or suspected clinical significance. Regions of long contiguous stretches of homozygosity (LCSH) are reported when they are greater than 10 Mb on a single chromosome or when the total LCSH is greater than 2% of the autosomal genome.

If a CNV is clinically significant and has been reported in multiple peer-reviewed publications, it will be reported as pathogenic even if there is variable penetrance and expressivity. CNVs at the time of reporting are not clearly pathogenic or benign should be reported as CNV of uncertain clinical significance. Uncertain clinical significant CNVs can be any one of the following categories: (a) uncertain clinical significance, likely pathogenic (based on a single case report with well-defined breakpoints and phenotype or if a gene within the CNV has a functional impact on the phenotype of the patient); (b) uncertain clinical significance likely benign (when no genes in the CNV interval are mapped as well as a small number of cases in the database); or (c) uncertain clinical significance with no sub-classification, containing genes with unknown function, and/or multiple case reports with contradictory results and no concrete conclusions. A CNV is considered as benign, if multiple peer-reviewed publications or databases annotated it as a benign variant or as a common polymorphism.

According to the guidelines for CMA reporting, CNVs especially in the postnatal constitutional setting, every report follow the current International system for Human Cytogenomic nomenclature (ISCN) and should include the cytogenetic location, dosage (copy number gains or losses), CNV size and linear co-ordinates with specified genome build, clear statement of clinical significance, list of relevant genes in the CNV interval, and recommendations for appropriate clinical follow-up. In cases with uncertain clinical significance, the report should include recommendations for continued surveillance through regular medical literature searches for new information (McGowan-Jordan et al. 2016). An ideal report preferably have an integrated cytogenetic and CMA results with details as mentioned above.

Advantages CMA is a fast and a highly sensitive test and enables the genome-wide detection of imbalances by one assay in an unbiased manner. It allows genome-wide screening of samples lacking fresh tissues, where chromosomal analysis is not possible. When compared to conventional karyotyping which needs culturing of cells, formalin-fixed paraffin-embedded (FFPE) slides can be used for CMA. CMA has a rapid turnaround time. CMA also defines the regions of imbalance if an abnormality is identified. In addition, CMA can detect most of the numerical abnormalities (monosomy, trisomy, triploidy, tetraploidy, etc.), and most unbalanced chromosomal rearrangements (unbalanced translocations, large deletions, and duplications). In addition to identifying copy number changes, copy neutral abnormalities such as LSCH can also be identified. Extended regions of homozygosity (AOH or LOH) with a total homozygosity of >3 Mb in all autosomes can be associated with uniparental disomy or consanguinity. AOH or LOH may pose increased risk for autosomal recessive conditions or imprinting disorders (Papenhausen et al. 2011). Microarray analysis also helps to characterize translocations at the molecular level. Translocations that are apparently balanced at the microscopic level may be revealed by molecular analysis to be unbalanced. About 20% of individuals with apparently balanced translocations (de novo or familial) have loss or gain of genetic material (Astbury et al. 2004; Sismani et al. 2008). In addition, mosaicism greater than 20%–25% can also be detected by CMA testing. Majority of CNVs are benign and clinically insignificant. However, the impact of CNVs is significant when it involves a critical region within a gene that has relevant phenotypic features associated with patients. If a gene is involved in the critical region of imbalance, CMA makes it possible to correlate the clinical features to the gene. Small gains and losses seen in structural abnormalities help to define clinical consequences (Astbury et al. 2004; Shanske et al. 2004; Simovich et al. 2007; Higgins et al. 2008; Tabet et al. 2015).

Limitations It is well recognized that CMA has many advantages over conventional cytogenetics and FISH assays but it also has many limitations. CMA does not detect small changes in the genome (point mutations, methylation status), and duplications within a single gene, low-level mosaicism below 20%–25%, or balanced rearrangements such as translocations, inversions, and insertions. In addition, CMA does not explain the chromosomal mechanism of a genetic imbalance (South et al. 2013). Also, CMA cannot differentiate between a free trisomy of an acrocentric chromosome and an unbalanced Robertsonian translocation. It is important to differentiate between these two entities as the recurrence risk is different (Fruhman and Van den Veyver 2010). In such cases, karyotype is recommended to rule out whether the abnormality is inherited or de novo in nature. Some CMA platforms do not detect triploidy and other ploidy levels. CMA cannot characterize clonal and subclonal populations in neoplastic samples. CMA is not recommended for post therapy follow-up or detection of minimal residual disease. Low-level mosaicism may not be detected by CMA. CNVs with incomplete penetrance and variable expression are significant challenges particularly in the prenatal setting. In addition, the limitations for detection of small CNVs depend on the probe coverage, software, and the platform used in each laboratory. Some factors which influence microarray are the quality of the DNA sequences on the array, the size of the DNA clones, density of the regions of interest, and the controls. In an ideal scenario to circumvent these limitations, CMA should be used in conjunction with other cytogenetic techniques. The advantages and limitations of the different technologies are summarized in Table 5.1.

5.3 CMA in Postnatal Diagnosis

Genetic testing including cytogenetic analysis through karyotyping has been the gold standard for patients with DD, ID, ASD, and MCA. In the general population, incidence of DD/ID is about 3% (Shevell et al. 2003) and ASD affects about 1 in 150 individuals (Autism and Developmental Monitoring Network Surveillance Year 2000 Principal Investigators 2007; Newschaffer et al. 2007). Since the introduction of CMA, a number of significantly relevant CNVs have been identified in about 15%–20% of cases (Rauch et al. 2004; de Vries et al. 2005; Hochstenbach et al. 2006; Vermeesch and Rauch 2006; Hoyer et al. 2007; Vermeesch et al. 2007; Miller et al. 2010; Mefford et al. 2011). The findings of these studies resulted in recommending CMA as the first-tier test for postnatal evaluation of individuals with DD/ID disorder, global developmental delay, ASD, and/or MCA (Kearney et al. 2011). American College of Medical Genetics

Table 5.1 Comparison of various molecular cytogenetic technologies

Type of technology	Type of cells	Advantages	Limitations
Conventional cytogenetics: karyotyping	Live cells needed to culture	1. Global view of numerical and structural chromosomal abnormalities (balanced and unbalanced translocations, inversions, marker chromosomes, homogenously staining regions (HSRs), double minutes (dmins), large deletions and duplications, and aneuploidy 2. Mosaicism can be detected 3. Sensitivity to detect ~5 Mb deletions and duplications	1. Contamination of cultures 2. Unable to define the marker chromosomes, HSRs, dmins 3. Small imbalances (<3–5 Mb) cannot be detected 4. Not sensitive to detect cryptic translocations especially at telomeric regions 5. Low level mosaicism cannot be detected
FISH	Cultured and uncultured cells	1. Specific and sensitive molecular technique 2. Can detect aneuploidy, deletion, amplifications, and gene rearrangements 3. Deletions below 1 MB can be detected	1. Only few loci can be evaluated at a time 2. Highly focused and prior knowledge of the gene(s) is required
aCGH	Cultured and uncultured cells, FFPE	1. Genome-wide detection of deletions, duplications, amplification, and aneuploidy 2. Can detect small deletions and duplications 3. Can detect small unbalanced structural abnormalities (~2 Mb)	1. Cannot detect balanced imbalances (translocations and inversions) 2. Does not give the mechanism of an imbalance 3. Low level mosaicism can be missed 4. Cannot detect UPD and LOH 5. Can detect origin of aneuploidy if parents are tested 6. Cannot detect copy neutral LOH
SNP CMA	Live cells, fixed samples including FFPE samples	1. Genome-wide deletions, duplications, and aneuploidy 2. Detection of UPD, LOH, and consanguinity 3. Copy neutral LOH can be detected	1. Cannot detect balanced imbalances (translocations and inversions) 2. Does not give the mechanism of an imbalance 3. Low level mosaicism can be missed

FFPE formalin-fixed paraffin-embedded samples, *UPD* uniparental disomy, *LOH* loss of heterozygosity

(ACMG) has developed standards and guidelines to educate the laboratory personnel to provide quality clinical services with reference to this spectrum of diseases (Cooley et al. 2013).

A review of 20 studies in patients with isolated congenital heart disease (CHDs) with or without other related defects showed clinically relevant CNVs in 3%–25% patients (Lander and Ware 2014). The most common submicroscopic CNV associated with CHD is a deletion of the 22q11.2 region, occurring in about 1 in 4000 live births (Fig. 5.1). This 22q11.2 CNV is associated with DiGeorge syndrome and other abnormalities including immune deficiency, hypocalcemia, and other neurodevelopmental disorders (McDonald-McGinn and Sullivan 2011). This abnormality cannot be identified by conventional karyotyping, but FISH detects the deletions using specific probes covering the 22q11.2 region. CMA is also recommended for individuals with multiple congenital anomalies and epilepsy. Studies have shown that most of these children do not have dysmorphic features that can be recognized as part of a syndrome but showed duplications and deletions. Since the inception of CMA as the first-tier test for the detection of imbalances, a number of new microdeletion and microduplication syndromes have been described. Some of the syndromes described include regions involving 1q21.1, 15q24, 17q21.31 and 17q23.1q23.2 (Ballif et al. 2008; Koolen et al. 2006; Sharp et al. 2006; Sharp et al. 2007; Shaw-Smith et al. 2006). Several recurrent genetic imbalances associated with incomplete penetrance and highly variable expressivity have also been recog-

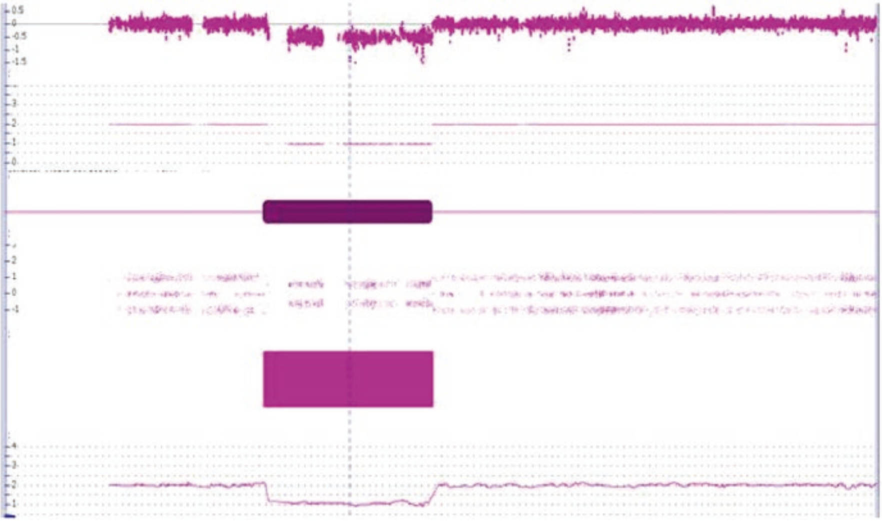


Fig. 5.1 SNP array showing 22q11.2 deletion in a patient with DiGeorge syndrome

nized (Mefford 2009). The microdeletion of 15q13.3 region has been associated with intellectual disability, epilepsy, or schizophrenia (Stefansson et al. 2008; Sharp et al. 2008; Helbig et al. 2009; International Schizophrenia Consortium 2008). Largest deletion identified at 16p11.2-p12.2 ranges from 7 to 9 Mb including the *SH2B1* gene that is associated with DD/and obesity (Bachmann-Gagescu et al. 2010).

With the introduction of CMA, many CNVs resulting in microduplication syndromes have also been described. Clinically relevant CNVs resulting in microduplication syndromes were seen at 1q21.1, 2q31, 3q29, 5q35, 7q11.23, 11p15 (Beckwith-Wiedemann syndrome), 15q11-13, 15q13.3, 15q24, 16p13.3, 16p13.11, 16p11.2, 17p13.3, 17p11.2 (Charcot-Marie Tooth type 1A disease), 17p11.2 (Potocki-Lupski syndrome), 17q21.31, 22q11.2, and 22q13 (Brunetti-Pierri et al. 2008; Mefford et al. 2008; Lisi et al 2008; Kantaputra et al. 2010; Cukier et al. 2012; Mullegama et al. 2015; Ballif et al. 2008; Goobie et al. 2008; Lisi et al. 2008; Franco et al. 2010; Zhang et al. 2011; Sanders et al. 2011; Berg et al. 2007; Baker et al. 1994; Bolton et al. 2001; Piard et al. 2010; van Bon et al. 2009; Stewart et al. 2011). A representative image showing a microduplication of 15q21.3 region is given in Fig. 5.2. When compared to microdeletion syndromes, the phenotype of microduplication syndromes is often less defined. Microduplication syndromes in general are less pathogenic and can also be inherited from normal parents, suggesting incomplete penetrance in some of these syndromes.

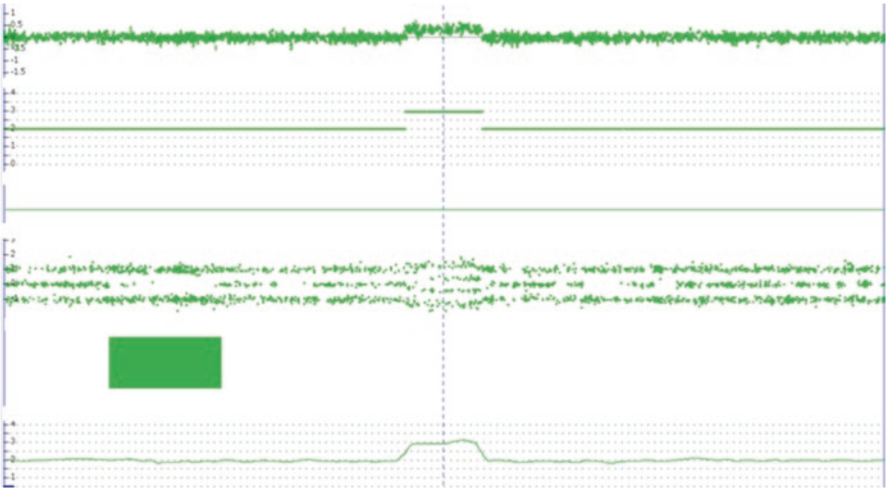


Fig. 5.2 SNP array showing duplication in chromosome 15

5.4 CMA in Prenatal Diagnosis

Microarray technology has revolutionized the practice of medical genetics in prenatal diagnosis. Structural abnormalities too small to be seen by conventional cytogenetics can now be detected by CMA. CMA detects common aneuploidies like trisomy 13, trisomy 18, and trisomy 21 in prenatal samples with 100% accuracy when compared to karyotyping (Wapner et al. 2012; Breman et al. 2012; Callaway et al. 2013). CMA detects about 5–7% of cases with abnormal ultrasound findings with clinically significant CNVs over conventional karyotyping (Shaffer et al. 2012; Wapner et al. 2012; South et al. 2013; Donnelly et al. 2014). In addition, in patients with advanced maternal age with positive serum screening, CMA detects about 1.7% of imbalances over standard karyotyping (Wapner et al. 2012; Srebniak et al. 2018). About 6% of abnormal fetuses with a normal karyotype may have pathogenic CNVs or likely pathogenic CNVs (Wapner et al. 2012; Shaffer et al. 2012; Srebniak et al. 2018). Even though CMA is recommended as the first tier of clinical diagnostic test for individuals with developmental disabilities or congenital anomalies (Manning and Hudgins 2010; Miller et al. 2010), prenatal CMA has become the standard of care in fetuses with congenital malformations (Wapner et al. 2012). Even though CMA has not yet substituted conventional cytogenetics through karyotyping for all indications, American College of Obstetrics and Gynecology (ACOG) and the Society for Maternal-Fetal Medicine (SMFM) (2015) recommended to consider CMA as a first-tier test in pregnancies with ultrasound abnormalities (Faas et al. 2010; Hillman et al. 2013; Vanakker et al. 2014; American College of Obstetrics and Gynecologists Committee on Genetics 2013; Dugoff et al. 2016; Walser et al. 2016; Wou et al. 2016).

Although advantage of CMA in structurally abnormal fetuses is well accepted, its utility in structurally normal fetuses is still a matter of some debate. There has been a trend to have CMA for patients who undergo invasive prenatal testing including cases with structurally normal fetuses. Pathogenic CNVs have been reported in about 1% of structurally normal fetuses (Van Opstal et al. 2015). In such low-risk pregnancies, the frequency of pathogenic CNVs reported varied considerably from 0.4% to 2% (Van Opstal et al. 2015; Wapner et al. 2015; Bornstein et al. 2017). Therefore, the ACOG and SMFM advocate that, in patients with a structurally normal fetus undergoing invasive testing (chorionic villi sampling or amniocentesis), fetal karyotyping or CMA may be performed. CMA also detects variants of uncertain clinical significance (VUSs) at a rate of approximately about 1.6%–4.2% (Wapner et al. 2012; Hillman et al. 2013; Westerfield et al. 2014). The possibility of finding CNVs of uncertain clinical significance, incomplete penetrance, or variable expressivity is significant, with associated phenotypic abnormalities ranging from normal to severely affected (Martin et al. 2015). A recent study showed that the overall risk for a pregnant woman to have a clinically significant cytogenetic abnormality is higher than 1 in 180 (Srebniak et al. 2018).

In the prenatal setting, interpretation is more challenging especially to predict the postnatal outcome in cases with incomplete penetrance and variable expressivity (Westerfield et al. 2014). Prenatal CMA detected low penetrance neurosusceptibility loci and created dilemmas in genetic counseling (Brabbing-Goldstein et al. 2018). Moreover, genetic counseling can also be challenging as VUS can be reclassified as benign or pathogenic variants as more and more cases are published over time (Werner-Lin et al. 2016). In spite of CMA being superior to the recently described non-invasive prenatal testing techniques, CMA tests dropped considerably for amniocentesis and chorionic villi samples over the years (Chan et al. 2015; Chetty et al. 2013; Williams 3rd et al. 2015; Brynn and Wapner 2018).

The CMA can be performed on DNA from uncultured cells (chorionic villi, amniotic fluid, and fetal blood) which results in a faster turnaround time in reporting the results. Due to the better and higher resolution of CMA over the conventional karyotyping, there is a greater likelihood of identifying VOUS, CNV containing genes with incomplete penetrance (variants in susceptibility loci, VISLs) (Oneda et al. 2014; Rosenfeld et al. 2013; Armengol et al. 2012; Cavalli et al. 2012), CNVs signifying a predisposition to late-onset diseases (Pichert et al. 2011), and CNVs that are relevant for future pregnancies only, for example, X-linked CNVs in a female fetus (Oneda et al. 2014).

Pre and post-test genetic counseling should be considered for CMA testing and should convey the advantages and limitations of the array. Even though balanced rearrangements (translocations and inversions) will be missed by CMA and do not have any clinical significance for the patient, it is important for the future pregnancies as the risk cannot be calculated if one of the parents is carrier of a balanced translocation. In such cases, karyotyping is necessary for identifying the balanced rearrangements. Genetic counselors should inform the patients of the potential finding of a clinically relevant CNV as well as CNVs of uncertain clinical significance. They should also discuss the phenotypic heterogeneity, variable penetrance, variable expressivity, potential identification of consanguinity, and non-paternity (Wapner et al. 2012; Hillman et al. 2013). Major professional societies like ACMG, ACOG, Canadian College of Medical Genetics, and Italian Society of Human Genetics do not encourage replacing prenatal karyotyping with CMA but recommend it as an adjunct test in specific cases only (ACOG Committee 2009; Duncan and Langolis 2011; Novelli et al. 2012).

5.5 CMA in Cancer Diagnosis

Although morphology is still the gold standard for cancer diagnostics, cell surface markers, immunohistochemistry, cytogenetics through karyotyping, FISH assays, real-time PCR, and Sanger sequencing have paved ways to better understand cancer and aided in the classification of neoplasms (Gresham et al. 2008; Paxton et al. 2015). However, there are some limitations, for example, morphology does not give information about the important factors like the genes involved and the clonal evolution. Cytogenetics has the advantage of not only detecting large chromosomal gains and losses, balanced and unbalanced rearrangements, but can also detect related and unrelated clones in a sample. However, cytogenetic analysis can only be performed on dividing cells and mature cells like plasma cells do not divide unless stimulated by specific mitogens. For some cases, karyotyping may not be possible due to the absence of dividing cells and in such cases CMA is extremely useful in detecting abnormalities. Moreover, cytogenetic analysis is a time-consuming process and the analysis depends on many factors like the skill of the technologist, quality of the metaphase cells, complexity of abnormalities, etc. Application of CMA in clinical oncology has circumvented most of these problems. However, high-resolution CMA on neoplastic samples is challenging because of the multiple abnormalities seen at the gene and chromosomal level.

Microarray testing was initially used for hematological malignancies (Golub et al. 1999; Alizadeh et al. 2000; Ebert and Golub 2004; Bullinger et al. 2004). Using class discovery studies Golub et al. (1999) were able to reveal diagnostic classes of acute myeloid leukemia and acute lymphoid leukemia especially when morphology suggested differential diagnosis. Alizadeh et al. (2000) in their study of diffuse large B-cell lymphoma (DLBCL) were able to distinguish two types of DLBCL which were previously unknown. They identified genes involved in B-cell activation and in germinal center formation and called these groups as "germinal center B-like DLBCL" and "activated B-like DLBCL." The two entities are biologically different but had significant prognostic values. The overall survival at 5 years for germinal center B-like DLBCL after anthracycline-based chemotherapy was 78%, whereas the overall survival was 16% for activated B-like DLBCL. Rapid integration and the clinical utility of microarray in the diagnostic laboratories lead to guidelines for the application of the microarray technique, quality control, and interpretation and reporting of array results (Cooley et al. 2013; Schoumans et al. 2016).

CMA has enhanced our understanding of diverse genetic abnormalities including gain and losses of genetic material, loss of heterozygosity (LOH), and other changes in hematological malignancies and has immensely helped in the diagnosis, prognosis, and management of cancer patients (Armengol et al. 2010; Gunnarsson et al. 2008; Okamoto et al. 2010; Slovak et al. 2011; Jung et al. 2017; Swerdlow et al. 2017; Taylor et al. 2017). A recent review of microarray studies on hematological malignancies emphasized the benefits of using microarray in myelodysplastic syndrome (MDS), B-lymphoblastic leukemia/lymphoma (B-cell ALL), chronic lymphocytic leukemia/small lymphocytic lymphoma (CLL/SLL), and Burkitt-like lymphoma with 11q aberration (Peterson et al. 2018). Studies on MDS patients showed that CMA not only confirmed or clarified chromosomal abnormalities seen by cytogenetics and FISH but also detected cryptic aberrations including deletions and copy neutral LOH (Kolquist et al. 2011; Stevens-Kroef et al. 2017). However, balanced rearrangements and low-level mosaicism were not identified by the microarrays. Studies on normal cases with MDS and cases with no analyzable karyotypes also showed recurrent CNVs (Thiel et al. 2011; Arenillas et al. 2013).

Microarray using SNP probes on B-ALL samples helped in distinguishing the pseudo-hyperdiploidy (due to doubling of near-haploid or low hypodiploid clones) from hyperdiploidy as the prognosis of these two abnormalities differs significantly (Nachman et al. 2007). Some of the other abnormalities which could be identified by CMA are intrachromosomal amplification of chromosome 21 (iAMP21), *ETV6* and *RB1* deletions, and *PAR1* deletions resulting in *P2RY8-CRLF2* fusions (Baughn et al. 2015). CMA can also differentiate iAMP21 from gains of chromosome 21. In summary, depending on the type of abnormalities seen in different hematological malignancies, microarray should be applied as a complementary technology to conventional cytogenetics, FISH, or RT-PCR. However, for deletions, gains, and amplification and ploidy levels with clinical significance, CMA is more suitable than conventional cytogenetics and FISH. There is substantial evidence that complex or increased CNVs and/or CN-LOH predict shortened overall survival in CLL/SLL (Ouillette et al. 2011). Laurie et al. (2014) compared the SNP array results of CLL patients and found that late-stage CLL has recurrent acquired abnormalities that do not occur in precursor conditions or in the general population. SNP-based arrays on plasma cell neoplasm (multiple myeloma) identified not only the abnormalities observed by FISH but all also identified prognostic relevant CNV-A to V (Stevens-Kroef et al. 2016, 2017; Agnelli et al. 2009). Additional prognostic relevant abnormalities include loss of 1p, 13q, and 17p. CMA in plasma cell neoplasm has helped in differentiating a near-tetraploid clone from a hyperdiploid clone (Stevens-Kroef et al. 2012). Significance of this finding is that a near-tetraploid clone has intermediate prognosis, whereas a hyperdiploid clone has a very favorable prognosis. Microarray testing should be used as a complementary test in hematological malignancies to detect copy number alterations, and in situations where normal and complex karyotypes reported, culture failure, and no analyzable metaphase cells are encountered and also to differentiate pseudodiploidy from heperdiploidy, detection of iAMP21, submicroscopic deletions, and amplifications of genes (Simons et al. 2012; Peterson et al. 2018).

As in hematological malignancies, microarray has also been used for detecting CNVs in solid tumors. However, genome-wide analysis of solid tumors is technically challenging due to various reasons. Even though DNA can be extracted from fresh tissue, typically, in many instances only available source of DNA is from formalin-fixed paraffin-embedded (FFPE) samples. FFPE samples represent about 80–90% of all archived solid tumors (Blow 2007). Different fixation timings, deterioration of DNA, and small amount of DNA can lead to assay failure and subsequent misinterpretation of results (Lewis et al. 2001). Another major obstacle in obtaining homogenous tumor DNA for any study is contamination of normal DNA that can hinder in getting the accurate LOH and copy number variant calls. A number of microarray platforms have been developed to evaluate cancer at the genomic level. One such array is Oncoscan array (Affymetrix, USA), used for FFPE samples. The assay is optimized for whole genome-wide copy number (CN), LOH, and somatic mutation (SM) from highly degraded FFPE samples. The assay utilizes the molecular inversion probe (MIP) technology (Coughlin et al. 2012). The assay covers about 900 cancer genes of which 74 clinically actionable SM can be detected. The assay requires less than <80 ng of DNA. The ability of genetic profiling of solid tumors using FFFP samples provides valuable information for diagnosis and prediction of treatment outcomes. Microarray studies on various tumor types have been reported in the literature, and this review does not represent all the studies in solid neoplasm.

Using high-resolution oligonucleotide array, Hawthorn and Cowell (2011) in a series Wilms tumor samples showed LOH events in about 45% of tumors. In their analysis of CNVs by tumor stage showed relatively stable karyotypes in stage 1 tumors and more complex array profiles in tumors for stages 3–5. SNP microarray provides a valuable insight on genetic aberrations in brain tumors and assists in stratification of patients for prognosis and guiding specific treatment choices. The embryonal tumors, in particular medulloblastoma (MB) and primitive neuroectodermal tumors (PNET), showed loss of 17p in more than 40% of cases of MB due to a gain of 17q (seen as isochromosome of 17q) (Inda et al. 2005; Kagawa et al. 2006). Combined analysis of loss of heterozygosity and copy number revealed no copy number alteration indicating the presence of copy number neutral LOH (cnLOH) in about half of the cases in glioblastoma multiforme (GBM) (Kuga et al. 2008). A recent study on two cases of clear cell papillary renal carcinoma identified neutral LOH of 10q11.22 (Alexiev and Zou 2014). Copy neutral LOH is the occurrence of LOH in the absence of allelic loss (copy number ≥ 2) and has been associated with the duplication of oncogenic mutations with concomitant loss of the normal allele. Increased copy number events were observed in ductal carcinoma (Gorringe et al. 2015). They also showed increased frequency of ERBB2 gene amplification, 20q gain, and 15q loss in recurrence ductal carcinoma in situ (DCIS), suggesting copy number changes to provide prognostic information for DCIS recurrence.

The first generation of DNA-microarray studies in human cancer focused on detecting differences in gene-expression profiles between tumors of different types and grades. Even though CMA is significantly superior to conventional cytogenetics and FISH in identifying cryptic imbalances, CNVs, and CN-LOH, CMA cannot detect balanced rearrangements, or detect evolving and existing clones below 20–30% of cells. CMA cannot be used to detect minimal residual disease. In MDS and acute leukemia, balanced translocations and inversions are quite common and certain balanced rearrangements are negative prognostic indicators, for example, inv(3)/t(3;3), t(9;22), t(6;9), and 11q23 translocations. Whole genome analysis using microarray may identify unrecognized clinically relevant molecular subsets that can help in identifying specific markers for personalized therapy. It is important to understand that the differential expression of genes does not indicate causality but microarray provides an important first step in target identification which can be followed by functional studies. The ability to detect and accurately define regions of variation across the genome will continue to be an important aspect of precision medicine efforts.

5.6 Conclusion

The detection of CNVs in a broad spectrum of disorders in prenatal and neonatal cases helps in early diagnosis, timely interventions, and targeted clinical management. Microarray studies have improved the diagnosis of cancer and prediction of clinical outcome, in turn have guided and optimized the treatment options in a number of hematological and solid malignancies. Although next-generation sequencing (NGS) technology has helped in detecting somatic variants including SNPs and indels, there are limitations to identify CNV information when compared to microarray. In a review of literature comparing the utility of a variety of techniques in MDS, Song et al. (2017) concluded that no single technology provides all necessary information for clinicians to plan the treatment protocols and that a combination of techniques is required. In future, combination of routine cytogenetics, FISH, SNP and CGH microarray and other high-throughput technologies (NGS, whole exome expression profiling) with powerful computational biology tools will strengthen the diagnostic specificity and sensitivity of the screening methods and in turn will result in better prognosis and treatment options of human diseases at the individual level (precision medicine).

Databases for References

Database of genotype and phenotype at NCBI (dbGaP): https://www.ncbi.nlm.nih.gov/gap

International standard Cytogenomic array (ISCA) https://www.iscaconsortium.org/

Cancer Genomics Consortium (formerly called Cancer Cytogenomics Microarray Consortium): https://www.cancergenomics.org/

UCSC genome browser: https://genome.ucsc.edu/

ENSEMBL: https://asia.ensembl.org/index.html

Database of genomic variants (DGV): http://dgv.tcag.ca/dgv/app/links

DECIPHER (Database of Chromosomal Imbalance and Phenotype in Humans using Ensembl Resources): https://decipher.sanger.ac.uk/

References

ACOG Committee (2009) Opinion No. 446: array comparative genomic hybridization in prenatal diagnosis. Obstet Gynecol 114(5):1161–1163. https://doi.org/10.1097/AOG.0b013e3181c33cad

Agnelli L, Mosca L, Fabris S, Lionetti M, Andronache A, Kwee I, Todoerti K, Verdelli D, Battaglia C, Bertoni F, Deliliers GL, Neri A (2009) A SNP microarray and FISH-based procedure to detect allelic imbalances in multiple myeloma: an integrated genomics approach reveals a wide gene dosage effect. Genes Chromosomes Cancer 48(7):603–614. https://doi.org/10.1002/gcc.20668

Alexiev BA, Zou YS (2014) Clear cell papillary renal cell carcinoma: a chromosomal microarray analysis for two cases using a novel Molecular Inversion Probe (MIP) technology. Pathol Res Pract 210:1049–1053

Alizadeh AA, Eisen MB, Davis RE, Ma C, Lossos IS, Rosenwald A, Boldrick JC, Sabet H, Tran T, Yu X, Powell JI, Yang L, Marti GE, Moore T, Hudson J Jr, Lu L, Lewis DB, Tibshirani R, Sherlock G, Chan WC, Greiner TC, Weisenburger DD, Armitage JO, Warnke R, Levy R, Wilson W, Grever MR, Byrd JC, Botstein D, Brown PO, Staudt LM (2000) Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling. Nature 403:503–511

American College of Obstetricians and Gynecologists Committee on Genetics (2013) Committee Opinion No. 581: the use of chromosomal microarray analysis in prenatal diagnosis. Obstet Gynecol 122(6):1374–1377. https://doi.org/10.1097/01.AOG.0000438962.16108.d1

Arenillas L, Mallo M, Ramos F, Guinta K, Barragán E, Lumbreras E, Larráyoz MJ, De Paz R, Tormo M, Abáigar M, Pedro C, Cervera J, Such E, José Calasanz M, Díez-Campelo M, Sanz GF, Hernández JM, Luño E, Saumell S, Maciejewski J, Florensa L, Solé F (2013) Single nucleotide polymorphism array karyotyping: a diagnostic and prognostic tool in myelodysplastic syndromes with unsuccessful conventional cytogenetic testing. Genes Chromosomes Cancer 52(12):1167–1177. https://doi.org/10.1002/gcc.22112

Armengol G, Canellas A, Alvarez Y, Bastida P, Toledo JS, Pérez-Iribarne Mdel M, Camós M, Tuset E, Estella J, Coll MD, Caballín MR, Knuutila S (2010) Genetic changes including gene copy number lterations and their relation to prognosis in childhood acute myeloid leukemia. Leuk Lymphoma 51:114–124

Armengol L, Nevado J, Serra-Juhé C, Plaja A, Mediano C, García-Santiago FA, García-Aragonés M, Villa O, Mansilla E, Preciado C, Fernández L, Ángeles Mori M, García-Pérez L, Lapunzina PD, Pérez-Jurado LA (2012) Clinical utility of chromosomal microarray analysis in invasive prenatal diagnosis. Hum Genet 131:513–523. https://doi.org/10.1007/s00439-011-1095-5

Astbury C, Christ LA, Aughton DJ, Cassidy SB, Kumar A, Eichler EE, Schwartz S (2004) Detection of deletions in de novo "balanced" chromosome rearrangements: further evidence for their role in phenotypic abnormalities. Genet Med 6:81–89

Autism and Developmental Monitoring Network Surveillance Year (2000) Principal Investigators (2007) Prevalence of autism spectrum disorders–autism and developmental disabilities monitoring network, six sites, United States. MMWR Surveill Summ 56:1–11

Bachmann-Gagescu R, Mefford HC, Cowan C, Glew GM, Hing AV, Wallace S, Bader PI, Hamati A, Reitnauer PJ, Smith R, Stockton DW, Muhle H, Helbig I, Eichler EE, Ballif BC, Rosenfeld J, Tsuchiya KD (2010) Recurrent 200-kb deletions of 16p11.2 that include the SH2B1 gene are associated with developmental delay and obesity. Genet Med 12(10):641–647. https://doi.org/10.1097/GIM.0b013e3181ef4286

Baker P, Piven J, Schwartz S, Patil S (1994) Brief report: duplication of chromosome 15q11-13 in two individuals with autistic disorder. J Autism Dev Disord 24:529

Ballif BC, Theisen A, McDonald-McGinn DM, Zackai EH, Hersh JH, Bejjani BA, Shaffer LG (2008) (2Identification of a previously unrecognized microdeletion syndrome of 16q11.2q12.2.008). Clin Genet. 74(5):469–75. https://doi.org/10.1111/j.1399-0004.2008.01094.x.PMID:18811697

Baughn LB, Biegel JA, South ST, Smolarek TA, Volkert S, Carroll AJ, Heerema NA, Rabin KR, Zweidler-McKay PA, Loh M, Hirsch B (2015). Integration of cytogenomic data for furthering the characterization of pediatric B-cell acute lymphoblastic leukemia: a multi-institution, multi-platform microarray study. Cancer Genet. 208(1-2):1–18. https://doi.org/10.1016/j.cancergen.2014.11.003.

Berg JS, Brunetti-Pierri N, Peters SU, Kang SH, Fong CT, Salamone J, Freedenberg D, Hannig VL, Prock LA, Miller DT, Raffalli P, Harris DJ, Erickson RP, Cunniff C, Clark GD, Blazo MA, Peiffer DA, Gunderson KL, Sahoo T, Patel A, Lupski JR, Beaudet AL, Cheung SW (2007) Speech delay and autism spectrum behaviors are frequently associated with duplication of the 7q11.23 Williams-Beuren syndrome region. Genet Med 9(7):427–441. doi: 10.1097GIM.0b013e3180986192

Blow N (2007) Tissue preparation: Tissue issues. Nature 448(7156):959–963

Bolton PF, Dennis NR, Browne CE, Thomas NS, Veltman MW, Thompson RJ, Jacobs P (2001) The phenotypic manifestations of interstitial duplications of proximal 15q with special reference to the autistic spectrum disorders. Am J Med Genet 105(8):675–685

Bornstein E, Berger S, Cheung SW, Maliszewski KT, Patel A, Pursley AN, Lenchner E, Bacino C, Beaudet AL, Divon MY (2017) Universal prenatal chromosomal microarray analysis: additive value and clinical dilemmas in fetuses with a normal karyotype. Am J Perinatol 34(4):340–348. https://doi.org/10.1055/s-0036-1586501

Brabbing-Goldstein D, Reches A, Svirsky R, Bar-Shira A, Yaron Y (2018) Dilemmas in genetic counseling for low-penetrance neuro-susceptibility loci detected on prenatal chromosomal microarray analysis. Am J Obstet Gynecol 218(2):247.e1–247.e12. https://doi.org/10.1016/j.ajog.2017.11.559

Breman A, Pursley AN, Hixson P, Bi W, Ward P, Bacino CA, Shaw C, Lupski JR, Beaudet A, Patel A, Cheung SW, Van den Veyver I (2012) Prenatal chromosomal microarray analysis in a diagnostic laboratory; experience with >1000 cases and review of the literature. Prenat Diagn 32(4):351–361. https://doi.org/10.1002/pd.3861

Brunetti-Pierri N, Berg JS, Scaglia F, Belmont J, Bacino CA, Sahoo T, Lalani SR, Graham B, Lee B, Shinawi M, Shen J, Kang SH, Pursley A, Lotze T, Kennedy G, Lansky-Shafer S, Weaver C, Roeder ER, Grebe TA, Arnold GL, Hutchison T, Reimschisel T, Amato S, Geragthy MT, Innis JW, Obersztyn E, Nowakowska B, Rosengren SS, Bader PI, Grange DK, Naqvi S, Garnica AD, Bernes SM, Fong CT, Summers A, Walters WD, Lupski JR, Stankiewicz P, Cheung SW, Patel A (2008) Recurrent reciprocal 1q21.1 deletions and duplications associated with microcephaly or macrocephaly and developmental and behavioral abnormalities. Nat Genet 40(12):1466–1471. https://doi.org/10.1038/ng.279

Brynn L, Wapner R (2018) Prenatal diagnosis by chromosomal microarray analysis. Fertil Steril 109(2):201–212

Bullinger L, Döhner K, Bair E, Fröhling S, Schlenk RF, Tibshirani R, Döhner H, Pollack JR (2004) Use of gene-expression profiling to identify prognostic subclasses in adult acute myeloid leukemia. N Engl J Med 350(16):1605–1616

Callaway JL, Shaffer LG, Chitty LS, Rosenfeld JA, Crolla JA (2013) The clinical utility of microarray technologies applied to prenatal cytogenetics in the presence of a normal conventional karyotype: a review of the literature. Prenat Diagn 233(12):1119–1123. https://doi.org/10.1002/pd.4209

Carter NP (2007) Methods and strategies for analyzing copy number variation using DNA microarrays. Nat Genet 39:S16–S21

Cavalli P, Cavallari U, Novelli A (2012) Array CGH in routine prenatal diagnosis practice. Prenat Diagn 32:708–709, author reply 11–2

Chan YM, Leung WC, Chan WP, Leung TY, Cheng YK, Sahota DS (2015) Women's uptake of non-invasive DNA testing following a high-risk screening test for trisomy 21 within a publicly funded healthcare system: findings from a retrospective review. Prenat Diagn 35(4):342–347. https://doi.org/10.1002/pd.4544

Chetty S, Garabedian MJ, Norton ME (2013) Uptake of noninvasive prenatal testing (NIPT) in women following positive aneuploidy screening. Prenat Diagn 33:542–546

Cooley LD, Lebo M, Li MM, Slovak ML, Wolff DJ, Working Group of the American College of Medical Genetics and Genomics (ACMG) Laboratory Quality Assurance Committee (2013) American College of Medical Genetics and Genomics technical standards and guidelines: microarray analysis for chromosome abnormalities in neoplastic disorders. Genet Med 15(6):484–494. https://doi.org/10.1038/gim.2013.49

Cooper GM, Coe BP, Girirajan S, Rosenfeld JA, Vu TH, Baker C, Williams C, Stalker H, Hamid R, Hannig V, Abdel-Hamid H, Bader P, McCracken E, Niyazov D, Leppig K, Thiese H, Hummel M, Alexander N, Gorski J, Kussmann J, Shashi V, Johnson K, Rehder C, Ballif BC, Shaffer LG, Eichler EE (2011) A copy number variation morbidity map of developmental delay. Nat Genet 43(9):838–846. https://doi.org/10.1038/ng.909

Coughlin CR II, Scharer GH, Shaikh TH (2012) Clinical impact of copy number variation analysis using high-resolution microarray technologies: advantages, limitations and concerns. Genome Med 4:80

Cukier HN, Lee JM, Ma D, Young JI, Mayo V, Butler BL, Ramsook SS, Rantus JA, Abrams AJ, Whitehead PL, Wright HH, Abramson RK, Haines JL, Cuccaro ML, Pericak-Vance MA, Gilbert JR (2012) The expanding role of MBD genes in autism: identification of a MECP2 duplication and novel alterations in MBD5, MBD6, and SETDB1. Autism Res 5(6):385–397. https://doi.org/10.1002/aur.1251

de Vries BB, Pfundt R, Leisink M, Koolen DA, Vissers LE, Janssen IM, Sv R, Nillesen WM, Huys EH, Nd L, Smeets D, Sistermans EA, Feuth T, van Ravenswaaij-Arts CM, van Kessel AG, Schoenmakers EF, Brunner HG, Veltman JA (2005) Diagnostic genome profiling in mental retardation. Am J Hum Genet 77(4):606–616

Donnelly JC, Platt LD, Rebarber A, Zachary J, Grobman WA, Wapner RJ (2014) Association of copy number variants with specific ultrasonographically detected fetal anomalies. Obstet Gynecol 124(1):83–90. https://doi.org/10.1097/AOG.0000000000000336

Duncan A, Langlois S, SOGC Genetics Committee, CCMG Prenatal Diagnosis Committee (2011) Use of array genomic hybridization technology in prenatal diagnosis in Canada. J Obstet Gynaecol Can 33(12):1256–1259

Dugoff L, Norton ME, Kuller JA. (2016) The use of chromosomal microarray for prenatal diagnosis. Society of Maternal Fetal medicine (SMFM) Consult Series #41 L Published by Elsevier Inc. http://dx.doi.org/10.1016/j.ajog.2016.07.016

Ebert BL, Golub TR (2004) Genomic approaches to hematologic malignancies. Blood 104:923–932

Faas BH, van der Burgt I, Kooper AJ, Pfundt R, Hehir-Kwa JY, Smits AP, de Leeuw N (2010) Identification of clinically significant, submicroscopic chromosome alterations and UPD in fetuses with ultrasound anomalies using genome-wide 250k SNP array analysis. J Med Genet 47(9):586–594. https://doi.org/10.1136/jmg.2009.075853

Franco LM, de Ravel T, Graham BH, Frenkel SM, Van Driessche J, Stankiewicz P, Lupski JR, Vermeesch JR, Cheung SW (2010) A syndrome of short stature, microcephaly and speech delay is associated with duplications reciprocal to the common Sotos syndrome deletion. Eur J Hum Genet 18(2):258–261. https://doi.org/10.1038/ejhg.2009.164

Fruhman G, Van den Veyver IB (2010) Applications of array comparative genomic hybridization in obstetrics. Obstet Gynecol Clin N Am 37:71–85

Golub TR, Slonim DK, Tamayo P, Huard C, Gaasenbeek M, Mesirov JP, Coller H, Loh ML, Downing JR, Caligiuri MA, Bloomfield CD, Lander ES (1999) Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. Science 286(5439):531–537

Goobie S, Knijnenburg J, Fitzpatrick D, Sharkey FH, Lionel AC, Marshall CR, Azam T, Shago M, Chong K, Mendoza-Londono R, den Hollander NS, Ruivenkamp C, Maher E, Tanke HJ, Szuhai K, Wintle RF, Scherer SW (2008) Molecular and clinical characterization of de novo and familial cases with microduplicatio 3q29: guidelines for copy number variation case reporting. Cytogeneti Genome Res 123:65–68. https://doi.org/10.1159/000184693

Gorringe KL, Hunter SM, Pang JM, Opeskin K, Hill P, Rowley SM, Choong DY, Thompson ER, Dobrovic A, Fox SB, Mann GB, Campbell IG (2015) Copy number analysis of ductal carcinoma in situ with and without recurrence. Mod Pathol 28(9):1174–1184. https://doi.org/10.1038/modpathol.2015.75

Gresham D, Dunham MJ, Botstein D (2008) Comparing whole genomes using DNA microarrays. Nat Rev Genet 9:291–302

Gunnarsson R, Staaf J, Jansson M, Ottesen AM, Göransson H, Liljedahl U, Ralfkiaer U, Mansouri M, Buhl AM, Smedby KE, Hjalgrim H, Syvänen AC, Borg A, Isaksson A, Jurlander J, Juliusson G, Rosenquist R (2008) Screening for copy-number alterations and loss of heterozygosity in chronic lymphocytic leukemia -a comparative study of four differently designed, high resolution microarray platforms. Genes Chromosomes Cancer 47(8):697–711. https://doi.org/10.1002/gcc.20575

Hawthorn L, Cowell JK (2011) Analysis of Wilms Tumors using SNP mapping array based comparative genomic hybridization. PLoS One 6(4):e18941

Helbig I, Mefford HC, Sharp AJ, Guipponi M, Fichera M, Franke A, Muhle H, de Kovel C, Baker C, von Spiczak S, Kron KL, Steinich I, Kleefuss-Lie AA, Leu C, Gaus V, Schmitz B, Klein KM, Reif PS, Rosenow F, Weber Y, Lerche H, Zimprich F, Urak L, Fuchs K, Feucht M, Genton P, Thomas P, Visscher F, de Haan GJ, Møller RS, Hjalgrim H, Luciano D, Wittig M, Nothnagel M, Elger CE, Nürnberg P, Romano C, Malafosse A, Koeleman BP, Lindhout D, Stephani U, Schreiber S, Eichler EE, Sander T (2009) 15q13.3 microdeletions increase risk of idiopathic generalized epilepsy. Nat Genet 41(2):160–162. https://doi.org/10.1038/ng.292

Higgins AW, Alkuraya FS, Bosco AF, Brown KK, Bruns GA, Donovan DJ, Eisenman R, Fan Y, Farra CG, Ferguson HL, Gusella JF, Harris DJ, Herrick SR, Kelly C, Kim HG, Kishikawa S, Korf BR, Kulkarni S, Lally E, Leach NT, Lemyre E, Lewis J, Ligon AH, Lu W, Maas RL, MacDonald ME, Moore SD, Peters RE, Quade BJ, Quintero-Rivera F, Saadi I, Shen Y, Shendure J, Williamson RE, Morton CC (2008) Characterization of apparently balanced chromosomal rearrangements from the developmental genome anatomy project. Am J Hum Genet 82:712–722. https://doi.org/10.1016/j.ajhg.2008.01.011

Hillman SC, McMullan DJ, Hall G, Togneri FS, James N, Maher EJ, Meller CH, Williams D, Wapner RJ, Maher ER, Kilby MD (2013) Use of prenatal chromosomal microarray: prospective cohort study and systematic review and meta-analysis. Ultrasound Obstet Gynecol 41(6):610–620. https://doi.org/10.1002/uog.12464

Hochstenbach R, van Amstel PHK, Poot M (2006) Microarray-based genome investigation: molecular karyotyping or segmental aneuploidy profiling? Eur J Hum Genet 14:262–265

Hoyer J, Dreweke A, Becker C, Göhring I, Thiel CT, Peippo MM, Rauch R, Hofbeck M, Trautmann U, Zweier C, Zenker M, Hüffmeier U, Kraus C, Ekici AB, Rüschendorf F, Nürnberg P, Reis A, Rauch A (2007) Molecular karyotyping in patients with mental retardation using 100K single-nucleotide polymorphism arrays. J Med Genet 44(10):629–636

Inda MM, Perot C, Guillaud-Bataille M, Danglot G, Rey JA, Bello MJ, Fan X, Eberhart C, Zazpe I, Portillo E, Tuñón T, Martínez-Peñuela JM, Bernheim A, Castresana JS (2005) Genetic heterogeneity in supratentorial and infratentorial primitive neuroectodermal tumours of the central nervous system. Histopathology 47(6):631–637

International Schizophrenia Consortium (2008) Rare chromosomal deletions and duplications increase risk of schizophrenia. Nature 455(7210):237–241. https://doi.org/10.1038/nature07239

Jung HS, Leffers JA, Tsongalis GJ (2017) Utilization of the oncoscan microarray assay in cancer diagnostics. Applied Cancer Research 37:1. https://doi.org/10.1186/s41241-016-0007-3

Kagawa N, Maruno M, Suzuki T, Hashiba T, Hashimoto N, Izumoto S, Yoshimine T (2006) Detection of genetic and chromosomal aberrations in medulloblastomas and primitive neuroectodermal tumors with DNA microarrays. Brain Tumor Pathol 23:41–47. https://doi.org/10.1007/s10014-006-0201-1

Kaminsky EB, Kaul V, Paschall J, Church DM, Bunke B, Kunig D, Moreno-De-Luca D, Moreno-De-Luca A, Mulle JG, Warren ST, Richard G, Compton JG, Fuller AE, Gliem TJ, Huang S, Collinson MN, Beal SJ, Ackley T, Pickering DL, Golden DM, Aston E, Whitby H, Shetty S, Rossi MR, Rudd MK, South ST, Brothman AR, Sanger WG, Iyer RK, Crolla JA, Thorland EC, Aradhya S, Ledbetter DH, Martin CL (2011) An evidence based approach to establish the functional and clinical significance of copy number variants in intellectual and developmental disabilities. Genet Med 13(9):777–784. https://doi.org/10.1097/GIM.0b013e31822c79f9

Kantaputra PN, Klopocki E, Hennig BP, Praphanphoj V, Le Caignec C, Isidor B, Kwee ML, Shears DJ, Mundlos S (2010) Mesomelic dysplasia Kantaputra type is associated with duplications of the HOXD locus on chromosome 2q. Eur J Hum Genet 18(12):1310–1314. https://doi.org/10.1038/ejhg.2010.116

Kearney HM, Thorland EC, Brown KK, Quintero-Rivera F, South ST, Working Group of the American College of Medical Genetics Laboratory Quality Assurance Committee (2011) American College of Medical Genetics standards and guidelines for interpretation and reporting of postnatal constitutional copy number variants. Genet Med 13(7):680–685. https://doi.org/10.1097/GIM.0b013e3182217a3a

Kolquist KA, Schultz RA, Furrow A, Brown TC, Han JY, Campbell LJ, Wall M, Slovak ML, Shaffer LG, Ballif BC (2011) Microarray-based comparative genomic hybridization of cancer targets reveals novel, recurrent genetic aberrations in the myelodysplatic syndromes. Cancer Genet 204(11):603–628. https://doi.org/10.1016/j.cancergen.2011.10.004

Koolen DA, Vissers LE, Pfundt R, de Leeuw N, Knight SJ, Regan R, Kooy RF, Reyniers E, Romano C, Fichera M, Schinzel A, Baumer A, Anderlid BM, Schoumans J, Knoers NV, van Kessel AG, Sistermans EA, Veltman JA, Brunner HG, de Vries BB (2006) A new chromosome 17q21.31 microdeletion syndrome associated with a common inversion polymorphism. Nat Genet 38:999–1001. https://doi.org/10.1038/ng1853

Kuga D, Mizoguchi M, Guan Y, Hata N, Yoshimoto K, Shono T, Suzuki SO, Kukita Y, Tahira T, Nagata S, Sasaki T, Hayashi K (2008) Prevalence of copy-number neutral LOH in glioblastomas revealed by genomewide analysis of laser-microdissected tissues. Neuro-Oncology 10(6):995–1003. https://doi.org/10.1215/15228517-2008-064

Lander J, Ware SM (2014) Copy number variation in congenital heart defects. Current Genetic Medicine Reports 2(3):168–178

Laurie CC, Laurie CA, Smoley SA, Carlson EE, Flinn I, Fridley BL, Greisman HA, Gribben JG, Jelinek DF, Nelson SC, Paietta E, Schaid D, Sun Z, Tallman MS, Weinshilboum R, Kay NE, Shanafelt TD (2014) Acquired chromosomal anomalies in chronic lymphocytic leukemia patients compared with more than 50,000 quasi-normal participants. Cancer Genet 207(1–2):19–30. https://doi.org/10.1016/j.cancergen.2014.01.004

Lewis F, Maughan NJ, Smith V, Hillan K, Quirke P (2001) Unlocking the archive--gene expression in paraffin-embedded tissue. J Pathol 195(1):66–71. https://doi.org/10.1002/1096-9896(200109)195:1<66::AID-PATH921>3.0.CO;2-F

Lisi EC, Hamosh A, Doheny KF, Squibb E, Jackson B, Galczynski R, Thomas GH, Batista DA (2008) 3q29 interstitial microduplication: a new syndrome in a three-generation family. Am J Med Genet A 146A(5):601–609. https://doi.org/10.1002/ajmg.a.32190

Manning M, Hudgins L (2010) Professional Practice and Guidelines Committee. Array-based technology and recommendations for utilization in medical genetics practice for detection of chromosomal abnormalities. Genet Med 12:742–745

Martin CL, Kirkpatrick BE, Ledbetter DH (2015) Copy number variants, aneuploidies, and human disease. Clin Perinatol 42:227–242

McDonald-McGinn DM, Sullivan KE (2011) Chromosome 22q11.2 deletion syndrome (DiGeorge syndrome/velocardiofacial syndrome). Medicine (Baltimore) 90:1–18

McGowan-Jordan J, Simons A, and Schmid M, Editors (2016). ISCN: An International System for Human Cytogenomic Nomenclature (2016). ISBN 978–3–318–05857–4

Mefford HC (2009) Genotype to phenotype-discovery and characterization of novel genomic disorders in a “genotype-first” era. Genet Med 11:836–842

Mefford HC, Sharp AJ, Baker C, Itsara A, Jiang Z, Buysse K, Huang S, Maloney VK, Crolla JA, Baralle D, Collins A, Mercer C, Norga K, de Ravel T, Devriendt K, Bongers EM, de Leeuw N, Reardon W, Gimelli S, Bena F, Hennekam RC, Male A, Gaunt L, Clayton-Smith J, Simonic I, Park SM, Mehta SG, Nik-Zainal S, Woods CG, Firth HV, Parkin G, Fichera M, Reitano S, Lo Giudice M, Li KE, Casuga I, Broomer A, Conrad B, Schwerzmann M, Räber L, Gallati S, Striano P, Coppola A, Tolmie JL, Tobias ES, Lilley C, Armengol L, Spysschaert Y, Verloo P, De Coene A, Goossens L, Mortier G, Speleman F, van Binsbergen E, Nelen MR, Hochstenbach R, Poot M, Gallagher L, Gill M, McClellan J, King MC, Regan R, Skinner C, Stevenson RE, Antonarakis SE, Chen C, Estivill X, Menten B, Gimelli G, Gribble S, Schwartz S, Sutcliffe JS, Walsh T, Knight SJ, Sebat J, Romano C, Schwartz CE, Veltman JA, de Vries BB, Vermeesch JR, Barber JC, Willatt L, Tassabehji M, Eichler EE (2008) Recurrent rearrangements of chromosome 1q21.1 and variable pediatric phenotypes. N Engl J Med 359(16):1685–1699. https://doi.org/10.1056/NEJMoa0805384

Mefford HC, Yendle SC, Hsu C, Cook J, Geraghty E, McMahon JM, Eeg-Olofsson O, Sadleir LG, Gill D, Ben-Zeev B, Lerman-Sagie T, Mackay M, Freeman JL, Andermann E, Pelakanos JT, Andrews I, Wallace G, Eichler EE, Berkovic SF, Scheffer IE (2011) Rare copy number variants are an important cause of epileptic encephalopathies. Ann Neurol 70(6):974–985. https://doi.org/10.1002/ana.22645

Mefford HC, Batshaw ML, Hoffman EP (2012) Genomics, intellectual disability, and autism. N Engl J Med 366:733–743

Miller DT, Adam MP, Aradhya S, Biesecker LG, Brothman AR, Carter NP, Church DM, Crolla JA, Eichler EE, Epstein CJ, Faucett WA, Feuk L, Friedman JM, Hamosh A, Jackson L, Kaminsky EB, Kok K, Krantz ID, Kuhn RM, Lee C, Ostell JM, Rosenberg C, Scherer SW, Spinner NB, Stavropoulos DJ, Tepperberg JH, Thorland EC, Vermeesch JR, Waggoner DJ, Watson MS, Martin CL, Ledbetter DH (2010) Consensus statement: chromosomal microarray is a first-tier clinical diagnostic test for individuals with developmental disabilities or congenital anomalies. Am J Hum Genet 86(5):749–764. https://doi.org/10.1016/j.ajhg.2010.04.006

Mullegama SV, Alaimo JT, Chen L, Elsea SH (2015) Phenotypic and molecular convergence of 2q23.1 deletion syndrome with other neurodevelopmental syndromes associated with autism spectrum disorder. Int J Mol Sci 16(4):7627–7643. https://doi.org/10.3390/ijms16047627

Nachman JB, Heerema NA, Sather H, Camitta B, Forestier E, Harrison CJ, Dastugue N, Schrappe M, Pui CH, Basso G, Silverman LB, Janka-Schaub GE (2007) Outcome of treatment in children with hypodiploid acute lymphoblastic leukemia. Blood 110:1112–1115. https://doi.org/10.1182/blood-2006-07-038299

Newschaffer CJ, Croen LA, Daniels J, Giarelli E, Grether JK, Levy SE, Mandell DS, Miller LA, Pinto-Martin J, Reaven J, Reynolds AM, Rice CE, Schendel D, Windham GC (2007) The epidemiology of autism spectrum disorders. Annu Rev Public Health 28:235–258. https://doi.org/10.1146/annurev.publhealth.28.021406.144007

Novelli A, Grati FR, Ballarati L, Bernardini L, Bizzoco D, Camurri L, Casalone R, Cardarelli L, Cavalli P, Ciccone R, Clementi M, Dalprà L, Gentile M, Gelli G, Grammatico P, Malacarne M, Nardone AM, Pecile V, Simoni G, Zuffardi O, Giardino D (2012) Microarray application in prenatal diagnosis: a position statement from the cytogenetics working group of the Italian Society of Human Genetics (SIGU), November 2011. Ultrasound Obstet Gynecol 39(4):384–388. https://doi.org/10.1002/uog.11092

Okamoto R, Ogawa S, Nowak D, Kawamata N, Akagi T, Kato M, Sanada M, Weiss T, Haferlach C, Dugas M, Ruckert C, Haferlach T, Koeffler HP (2010) Genomic profiling of adult acute lymphoblastic leukemia by single nucleotide polymorphism oligonucleotide microarray and comparison to pediatric acute lymphoblastic leukemia. Haematologica 95(9):1481–1488. https://doi.org/10.3324/haematol.2009.011114

Oneda B, Baldinger R, Reissmann R, Reshetnikova I, Krejci P, Masood R, Ochsenbein-Kölble N, Bartholdi D, Steindl K, Morotti D, Faranda M, Baumer A, Asadollahi R, Joset P, Niedrist D, Breymann C, Hebisch G, Hüsler M, Mueller R, Prentl E, Wisser J, Zimmermann R, Rauch A (2014) High-resolution chromosomal microarrays in prenatal diagnosis significantly increase diagnostic power. Prenat Diagn 34(6):525–533. https://doi.org/10.1002/pd.4342

Ouillette P, Collins R, Shakhan S, Li J, Peres E, Kujawski L, Talpaz M, Kaminski M, Li C, Shedden K, Malek SN (2011) Acquired genomic copy number aberrations and survival in chronic lymphocytic leukemia. Blood 118(11):3051–3061. https://doi.org/10.1182/blood-2010-12-327858

Papenhausen P, Schwartz S, Risheg H, Keitges E, Gadi I, Burnside RD, Jaswaney V, Pappas J, Pasion R, Friedman K, Tepperberg J (2011) UPD detection using homozygosity profiling with a SNP genotyping microarray. Am J Med Genet A 155A(4):757–768. https://doi.org/10.1002/ajmg.a.33939

Paxton CN, Rowe LR, South ST (2015) Observation of the genomic landscape beyond 1p19q deletions and EGFR amplification in glioma. Mol Cytogenet 8:60. https://doi.org/10.1186/s13039-015-0156-1. eCollection 2015

Peterson JF, Van Dyke DL, Hoppman NL, Kearney HM, Sukov WR, Greipp PT, Ketterling RP, Baughn LB (2018) The utilization of chromosomal microarray technologies for hematologic neoplasms: an ACLPS critical review. Am J Clin Pathol 150(5):375–384. https://doi.org/10.1093/ajcp/aqy076

Piard J, Philippe C, Marvier M, Beneteau C, Roth V, Valduga M, Béri M, Bonnet C, Grégoire MJ, Jonveaux P, Leheup B (2010) Clinical and molecular characterization of a large family with an interstitial 15q11q13 duplication. Am J Med Genet A 152A(8):1933–1941. https://doi.org/10.1002/ajmg.a.33521

Pichert G, Mohammed SN, Ahn JW, Ogilvie CM, Izatt L (2011) Unexpected findings in cancer predisposition genes detected by array comparative genomic hybridization: what are the issues? J Med Genet 48(8):535–539. https://doi.org/10.1136/jmg.2010.087593

Rauch A, Rüschendorf F, Huang J, Trautmann U, Becker C, Thiel C, Jones KW, Reis A, Nürnberg P (2004) Molecular karyotyping using an SNP array for genome wide genotyping. J Med Genet 41:916–922. https://doi.org/10.1136/jmg.2004.022855

Rosenfeld JA, Coe BP, Eichler EE, Cuckle H, Shaffer LG (2013) Estimates of penetrance for recurrent pathogenic copy-number variations. Genet Med 15(6):478–481. https://doi.org/10.1038/gim.2012.164

Sagoo GS, Butterworth AS, Sanderson S, Shaw-Smith C, Higgins JP, Burton H (2009) Array CGH in patients with learning disability (mental retardation) and congenital anomalies: updated systematic review and meta-analysis of 19 studies and 13,926 subjects. Genet Med 11(3): 139–146. https://doi.org/10.1097/GIM.0b013e318194ee8f

Sanders SJ, Ercan-Sencicek AG, Hus V, Luo R, Murtha MT, Moreno-De-Luca D, Chu SH, Moreau MP, Gupta AR, Thomson SA, Mason CE, Bilguvar K, Celestino-Soper PB, Choi M, Crawford EL, Davis L, Wright NR, Dhodapkar RM, DiCola M, DiLullo NM, Fernandez TV, Fielding-Singh V, Fishman DO, Frahm S, Garagaloyan R, Goh GS, Kammela S, Klei L, Lowe JK, Lund SC, McGrew AD, Meyer KA, Moffat WJ, Murdoch JD, O'Roak BJ, Ober GT, Pottenger RS, Raubeson MJ, Song Y, Wang Q, Yaspan BL, Yu TW, Yurkiewicz IR, Beaudet AL, Cantor RM, Curland M, Grice DE, Günel M, Lifton RP, Mane SM, Martin DM, Shaw CA, Sheldon M, Tischfield JA, Walsh CA, Morrow EM, Ledbetter DH, Fombonne E, Lord C, Martin CL, Brooks AI, Sutcliffe JS, Cook EH Jr, Geschwind D, Roeder K, Devlin B, State MW (2011) Multiple recurrent de novo CNVs, including duplications of the 7q11.23 Williams syndrome region, are strongly associated with autism. Neuron 70(5):863–885. https://doi.org/10.1016/j.neuron.2011.05.002

Schoumans J, Suela J, Hastings R, Muehlematter D, Rack K, van den Berg E, Berna Beverloo H, Stevens-Kroef M (2016) Guidelines for genomic array analysis in acquired haematological neoplastic disorders. Genes Chromosomes Cancer 55(5):480–491. https://doi.org/10.1002/gcc.22350

Shaffer LG, Kashork CD, Saleki R, Rorem E, Sundin K, Ballif BC, Bejjani BA (2006) Targeted genomic microarray analysis for identification of chromosome abnormalities in 1500 consecutive clinical cases. J Pediatr 149(4):98–102. https://doi.org/10.1016/j.jpeds.2006.02.006

Shaffer LG, Rosenfeld JA, Dabell MP, Coppinger J, Bandholz AM, Ellison JW, Ravnan JB, Torchia BS, Ballif BC, Fisher AJ (2012) Detection rates of clinically significant genomic alterations by microarray analysis for specific anomalies detected by ultrasound. Prenat Diagn 32(10):986–995. https://doi.org/10.1002/pd.3943

Shanske AL, Edelmann L, Kardon NB, Gosset P, Levy B (2004) Detection of an interstitial deletion of 2q21-22 by high resolution comparative genomic hybridization in a child with multiple congenital anomalies and an apparent balanced translocation. Am J Med Genet A 131(1):29–35. https://doi.org/10.1002/ajmg.a.30311

Sharp AJ, Hansen S, Selzer RR, Cheng Z, Regan R, Hurst JA, Stewart H, Price SM, Blair E, Hennekam RC, Fitzpatrick CA, Segraves R, Richmond TA, Guiver C, Albertson DG, Pinkel D, Eis PS, Schwartz S, Knight SJ, Eichler EE (2006) Discovery of previously unidentified genomic disorders from the duplication architecture of the human genome. Nat Genet 38(9):1038–1042

Sharp AJ, Selzer RR, Veltman JA, Gimelli S, Gimelli G, Striano P, Coppola A, Regan R, Price SM, Knoers NV, Eis PS, Brunner HG, Hennekam RC, Knight SJ, de Vries BB, Zuffardi O, Eichler EE (2007) Characterization of a recurrent 15q24 microdeletion syndrome. Hum Mol Genet 16(5):567–572. https://doi.org/10.1093/hmg/ddm016

Sharp AJ, Mefford HC, Li K, Baker C, Skinner C, Stevenson RE, Schroer RJ, Novara F, De Gregori M, Ciccone R, Broomer A, Casuga I, Wang Y, Xiao C, Barbacioru C, Gimelli G, Bernardina BD, Torniero C, Giorda R, Regan R, Murday V, Mansour S, Fichera M, Castiglia L, Failla P, Ventura M, Jiang Z, Cooper GM, Knight SJ, Romano C, Zuffardi O, Chen C, Schwartz CE, Eichler EE (2008) A recurrent 15q13.3 microdeletion syndrome associated with mental retardation and seizures. Nat Genet 40:322–328. https://doi.org/10.1038/ng.93

Shaw-Smith C, Pittman AM, Willatt L, Martin H, Rickman L, Gribble S, Curley R, Cumming S, Dunn C, Kalaitzopoulos D, Porter K, Prigmore E, Krepischi-Santos AC, Varela MC, Koiffmann CP, Lees AJ, Rosenberg C, Firth HV, de Silva R, Carter NP (2006) Microdeletion encompassing MAPT at chromosome 17q21.3 is associated with developmental delay and learning disability. Nat Genet 38(9):1032–1037. https://doi.org/10.1038/ng1858

Shevell M, Ashwal S, Donley D, Flint J, Gingold M, Hirtz D, Majnemer A, Noetzel M, Sheth RD, Quality Standards Subcommittee of the American Academy of Neurology, Practice Committee of the Child Neurology Society (2003) Practice parameter: evaluation of the child with global developmental delay: report of the Quality Standards Subcommittee of the American Academy of Neurology and The Practice Committee of the Child Neurology Society. Neurology 60(3):367–380

Simons A, Sikkema-Raddatz B, de Leeuw N, Konrad NC, Hastings RJ, Schoumans J (2012) Genome-wide arrays in routine diagnostics of hematological malignancies. Hum Mutat 33(6):941–948. https://doi.org/10.1002/humu.22057

Simovich MJ, Yatsenko SA, Kang SH, Cheung SW, Dudek ME, Pursley A, Ward PA, Patel A, Lupski JR (2007) Prenatal diagnosis of a 9q34.3 microdeletion by array-CGH in a fetus with an apparently balanced translocation. Prenat Diagn 27(12):1112–1117. https://doi.org/10.1002/pd.1841

Sismani C, Kitsiou-Tzeli S, Ioannides M (2008) Cryptic genomic imbalances in patients with de novo or familial apparently balanced translocations and abnormal phenotype. Mol Cytogenet 1:15

Slovak ML, Bedell V, Hsu YH, Estrine DB, Nowak NJ, Delioukina ML, Weiss LM, Smith DD, Forman SJ (2011) Molecular karyotypes of Hodgkin and Reed-Sternberg cells at disease onset reveal distinct copy number alterations in chemosensitive versus refractory Hodgkin lymphoma. Clin Cancer Res 17(10):3443–3454. https://doi.org/10.1158/1078-0432.CCR-10-1071

Society for Maternal-Fetal Medicine (SMFM), Dugoff L, Norton ME, Kuller JA (2015) The use of chromosomal microarray for prenatal diagnosis. Society for Maternal-Fetal Medicine (SMFM) Consult Series I #41. Am J Obstet Gynecol 215(4):B2–B9. https://doi.org/10.1016/j.ajog.2016.07.016

Song Q, Peng M, Chu Y, Huang S (2017) Techniques for detecting chromosomal aberrations in myelodysplastic syndromes. Oncotarget 8(37):62716–62729. https://doi.org/10.18632/oncotarget.17698

South ST, Lee C, Lamb AN, Higgins AW, Kearney HM (2013) Working Group for the American College of Medical Genetics and Genomics Laboratory Quality Assurance Committee. ACMG Standards and Guidelines for constitutional cytogenomic microarray analysis, including postnatal and prenatal applications: revision 2013. Genet Med 15:901–909

Srebniak MI, Joosten M, Knapen MFCM, Arends LR, Polak M, van Veen S, Go ATJI, Van Opstal D (2018) Frequency of submicroscopic chromosomal aberrations in pregnancies without increased risk for structural chromosomal aberrations: systematic review and meta-analysis. Ultrasound Obstet Gynecol 51(4):445–452. https://doi.org/10.1002/uog.17533

Stefansson H, Rujescu D, Cichon S, Pietiläinen OP, Ingason A, Steinberg S, Fossdal R, Sigurdsson E, Sigmundsson T, Buizer-Voskamp JE, Hansen T, Jakobsen KD, Muglia P, Francks C, Matthews PM, Gylfason A, Halldorsson BV, Gudbjartsson D, Thorgeirsson TE, Sigurdsson A, Jonasdottir A, Jonasdottir A, Bjornsson A, Mattiasdottir S, Blondal T, Haraldsson M, Magnusdottir BB, Giegling I, Möller HJ, Hartmann A, Shianna KV, Ge D, Need AC, Crombie C, Fraser G, Walker N, Lonnqvist J, Suvisaari J, Tuulio-Henriksson A, Paunio T, Toulopoulou T, Bramon E, Di Forti M, Murray R, Ruggeri M, Vassos E, Tosato S, Walshe M, Li T, Vasilescu C, Mühleisen TW, Wang AG, Ullum H, Djurovic S, Melle I, Olesen J, Kiemeney LA, Franke B, GROUP, Sabatti C, Freimer NB, Gulcher JR, Thorsteinsdottir U, Kong A, Andreassen OA, Ophoff RA, Georgi A, Rietschel M, Werge T, Petursson H, Goldstein DB, Nöthen MM, Peltonen L, Collier DA, St Clair D, Stefansson K (2008) Large recurrent microdeletions associated with schizophrenia. Nature 455(7210):232–236. https://doi.org/10.1038/nature07229

Stevens-Kroef M, Weghuis DO, Croockewit S (2012) High detection rate of clinically relevant genomic abnormalities in plasma cells enriched from patients with multiple myeloma. Genes Chromosomes Cancer 51(11):997–1006

Stevens-Kroef M, Weghuis DO, Wezenberg S, Croockewit S, Wessels H, van der Mespel M, Zweegman S, Poddighe PJ (2016) Superior identification of prognostic relevant copy number abnormalities by SNP-based genomic arrays as compared to interphase FISH in multiple myeloma. Blood 128:4426

Stevens-Kroef MJ, Olde Weghuis D, ElIdrissi-Zaynoun N, van der Reijden B, Cremers EMP, Alhan C, Westers TM, Visser-Wisselaar HA, Chitu DA, Cunha SM, Vellenga E, Klein SK, Wijermans P, de Greef GE, Schaafsma MR, Muus P, Ossenkoppele GJ, van de Loosdrecht AA, Jansen JH (2017) Genomic array as compared to karyotyping in myelodysplastic syndromes in a prospective clinical trial. Genes Chromosomes Cancer 56(7):524–534. https://doi.org/10.1002/gcc.22455

Stewart LR, Hall AL, Kang SH, Shaw CA, Beaudet AL (2011) High frequency of known copy number abnormalities and maternal duplication 15q11-q13 in patients with combined schizophrenia and epilepsy. BMC Med Genet 12:154. https://doi.org/10.1186/1471-2350-12-154

Swerdlow SH, Campo E, Harris NL, Jaffe ES, Pileri SA, Stein H, Thiele J (2017) WHO classification of tumours of haematopoietic and lymphoid tissues. Revised 4th ed. World Health Organization; International Agency for Research on Cancer, Lyon. ISBN:9789283244943; 9781234567897 NLM ID: 101716000

Tabet AC, Verloes A, Pilorge M, Delaby E, Delorme R, Nygren G, Devillard F, Gérard M, Passemard S, Héron D, Siffroi JP, Jacquette A, Delahaye A, Perrin L, Dupont C, Aboura A, Bitoun P, Coleman M, Leboyer M, Gillberg C, Benzacken B, Betancur C (2015) Complex nature of apparently balanced chromosomal rearrangements in patients with autism spectrum disorder. Mol Autism 6:19. https://doi.org/10.1186/s13229-015-0015-2. eCollection 2015

Taylor J, Xiao W, Abdel-Wahab O (2017) Diagnosis and classification of hematologic malignancies on the basis of genetics. Blood 130(4):410–423. https://doi.org/10.1182/blood-2017-02-734541

Thiel A, Beier M, Ingenhag D, Servan K, Hein M, Moeller V, Betz B, Hildebrandt B, Evers C, Germing U, Royer-Pokora B (2011) Comprehensive array CGH of normal karyotype myelodysplastic syndromes reveals hidden recurrent and individual genomic copy number alterations with prognostic relevance. Leukemia:387–399. https://doi.org/10.1038/leu.2010.293

van Bon BW, Mefford HC, Menten B, Koolen DA, Sharp AJ, Nillesen WM, Innis JW, de Ravel TJ, Mercer CL, Fichera M, Stewart H, Connell LE, Ounap K, Lachlan K, Castle B, Van der Aa N, van Ravenswaaij C, Nobrega MA, Serra-Juhé C, Simonic I, de Leeuw N, Pfundt R, Bongers EM, Baker C, Finnemore P, Huang S, Maloney VK, Crolla JA, van Kalmthout M, Elia M, Vandeweyer G, Fryns JP, Janssens S, Foulds N, Reitano S, Smith K, Parkel S, Loeys B, Woods CG, Oostra A, Speleman F, Pereira AC, Kurg A, Willatt L, Knight SJ, Vermeesch JR, Romano C, Barber JC, Mortier G, Pérez-Jurado LA, Kooy F, Brunner HG, Eichler EE, Kleefstra T, de Vries BB (2009) Further delineation of the 15q13 microdeletion and duplication syndromes: a clinical spectrum varying from non-pathogenic to a severe outcome. J Med Genet 46(8):511–523. https://doi.org/10.1136/jmg.2008.063412

Van Opstal D, de Vries F, Govaerts L, Boter M, Lont D, van Veen S, Joosten M, Diderich K, Galjaard RJ, Srebniak MI (2015) Benefits and burdens of using a SNP array in pregnancies at increased risk for the common aneuploidies. Hum Mutat 36(3):319–326. https://doi.org/10.1002/humu.22742

Vanakker O, Vilain C, Janssens K, Van der Aa N, Smits G, Bandelier C, Blaumeiser B, Bulk S, Caberg JH, De Leener A, De Rademaeker M, de Ravel T, Desir J, Destree A, Dheedene A, Gaillez S, Grisart B, Hellin AC, Janssens S, Keymolen K, Menten B, Pichon B, Ravoet M, Revencu N, Rombout S, Staessens C, Van Den Bogaert A, Van Den Bogaert K, Vermeesch JR, Kooy F, Sznajer Y, Devriendt K (2014) Implementation of genomic arrays in prenatal diagnosis: the Belgian approach tomeet the challenges. Eur J Med Genet 57(4):151–156. https://doi.org/10.1016/j.ejmg.2014.02.002

Vermeesch JR, Rauch A (2006) Reply to Hochstenbach et al. 'Molecular karyotyping'. Eur J Hum Genet 14:1063–1064

Vermeesch JR, Fiegler H, de Leeuw N, Szuhai K, Schoumans J, Ciccone R, Speleman F, Rauch A, Clayton-Smith J, Van Ravenswaaij C, Sanlaville D, Patsalis PC, Firth H, Devriendt K, Zuffardi O (2007) Guidelines for molecular karyotyping in constitutional genetic diagnosis. Eur J Hum Genet 15(11):1105–1114. https://doi.org/10.1038/sj.ejhg.5201896

Walser SA, Werner-Lin A, Russell A, Wapner RJ, Bernhardt BA (2016) "Something extra on chromosome 5": Parents' understanding of positive prenatal chromosomal microarray analysis (CMA) results. J Genet Couns 25(5):1116–1126. https://doi.org/10.1007/s10897-016-9943-z

Wang Y, Moorhead M, Karlin-Neumann G, Falkowski M, Chen C, Siddiqui F, Davis RW, Willis TD, Faham M (2005) Allele quantification using molecular inversion probes (MIP). Nucleic Acids Res 33(21):e183. https://doi.org/10.1093/nar/gni177

Wapner RJ, Martin CL, Levy B, Ballif BC, Eng CM, Zachary JM, Savage M, Platt LD, Saltzman D, Grobman WA, Klugman S, Scholl T, Simpson JL, McCall K, Aggarwal VS, Bunke B, Nahum O, Patel A, Lamb AN, Thom EA, Beaudet AL, Ledbetter DH, Shaffer LG, Jackson L (2012) Chromosomal microarray versus karyotyping for prenatal diagnosis. N Engl J Med 367(23):2175–2184. https://doi.org/10.1056/NEJMoa1203382

Wapner RJ, Zachary J, Clifton R (2015) Change in classification of prenatal microarray analysis copy number variants over time. Prenat Diagn 35(Suppl S1):8–3

Werner-Lin A, McCoyd JLM, Bernhardt BA (2016) Balancing genetics (science) and counseling (art) in prenatal chromosomal microarray testing. J Genet Couns 5:855–867

Westerfield L, Darilek S, van den Veyver I (2014) Counseling challenges with variants of uncertain significance and incidental findings in prenatal genetic screening and diagnosis. J Clin Med 3:1018–1032

Williams J 3rd, Rad S, Beauchamp S, Ratousi D, Subramaniam V, Farivar S, Pisarska MD (2015) Utilization of noninvasive prenatal testing: impact on referrals for diagnostic testing. Am J Obstet Gynecol 213(1):102.e1–102.e6. https://doi.org/10.1016/j.ajog.2015.04.005

Wou K, Levy B, Wapner RJ (2016) Chromosomal microarrays for the prenatal detection of microdeletions and microduplications. Clin Lab Med 36(2):261–276. https://doi.org/10.1016/j.cll.2016.01.017

Zhang H, Lu X, Beasley J, Mulvihill JJ, Liu R, Li S, Lee JY (2011) Reversed clinical phenotype due to a microduplication of Sotos syndrome region detected by array CGH: microcephaly, developmental delay and delayed bone age. Am J Med Genet A 155A(6):1374–1378. https://doi.org/10.1002/ajmg.a.33769

Chapter 6
Finding a Needle in a Haystack: Variant Effect Predictor (VEP) Prioritizes Disease Causative Variants from Millions of Neutral Ones

Yashvant Khimsuriya, Salil Vaniyawala, Babajan Banaganapalli, Muhammadh Khan, Ramu Elango, and Noor Ahmad Shaik

Contents

Y. Khimsuriya (✉) · S. Vaniyawala (✉)
SN Gene Laboratory and Research Centre, Surat, Gujarat, India
e-mail: salil@sngenelab.com

B. Banaganapalli (✉) · R. Elango · N. A. Shaik (✉)
Princess Al-Jawhara Center of Excellence in Research of Hereditary Disorders,
Department of Genetic Medicine, Faculty of Medicine,
King Abdulaziz University, Jeddah, Saudi Arabia
e-mail: bbabajan@kau.edu.sa; relango@kau.edu.sa; nshaik@kau.edu.sa

M. Khan
Department of Clinical Laboratory Sciences, College of Applied Medical Sciences,
King Saud University, Riyadh, Saudi Arabia
e-mail: imkhan@ksu.edu.sa

N. A. Shaik et al. (eds.), *Essentials of Bioinformatics, Volume II*,
https://doi.org/10.1007/978-3-030-18375-2_6

6.1 Introduction

An approximate number of 22,000 genes of the human genome encode all the functional proteins forming the protein-coding blueprint of the human proteome (International Human Genome Sequencing Consortium 2004). Any molecular defect, be it distinct or multiple abnormalities spanning a single or multiple genes in the genome, can become the basis of a genetic disease in humans.

A disease condition caused by a mutation in one of the identified genes is known as a monogenic or single gene disorder. After the completion of the human genome sequencing, researchers began to shift their efforts from monogenic to polygenic disorders, which are caused by mutations in multiple genes (Antonarakis and Beckmann 2006). There are as many as 1621 monogenic diseases for which identified genes are very uncommon. Consequently, researchers face complications in recognizing relations between mutation and the genetic syndrome and also to collect adequate amounts of genetic and clinical material for the evaluation of unaffected family participants. Moreover, biotechnology corporations, funding agencies, and pharmaceutical industries are often not interested in investing financial resources in researching rare genetic conditions (Stenson et al. 2003).

The determination of the human genome sequence has enabled scientists to generate sequence maps of all human chromosomes. The precise location of every gene is already mapped, and the polymorphic regions of the genome are identified. Out of these genomic variations, single-nucleotide polymorphisms (SNPs), which are single base pair polymorphic regions, were of special interest to the scientists and clinicians (Schmutz et al. 2004). On average, SNPs occur 1 per every 1000 base pairs in the genome (Sachidanandam et al. 2001). HapMap Project documents all the discovered SNPs along the length of the chromosome. SNPs can be used as biomarkers to map disease-associated genes (Consortium 2003). This information has been freely available to scientists worldwide that further developed the new fields of biology named bioinformatics and computational biology.

Studying rare disorders is always challenging because of the low occurrence and the inadequate penetrance of concerned alleles (Cirulli and Goldstein 2010). The whole-genome sequence (WGS) or whole-exome sequence (WES) of rare disease patients often generates a huge list of variants, running thousands to hundreds of thousands in number (Dewey et al. 2014). Filtering the real disease causative variants from the huge crowd of neutral variants helps to explore treatment possibilities and for personalized medicine. Effective filtration of neutral variants is key to significantly reduce the technical labor, economic cost, and time factors required for studying every single mutation generated by sequencing methods.

Bioinformatics scrutinizes genomic information to forecast gene-gene, protein-protein, and gene-protein interactions and functions. Additionally, the correlation of the sequence of a gene of unidentified function to the rest of the genome helps find similar genes with known functions. Based on the relationship between genes, scientists can often predict the function of the hypothetical protein encoded by these genes within a cell.

Advances in genetic techniques in the past decade, such as high-throughput technologies, has been widely applied throughout biological and biomedical fields of research. Moreover, WES is the most progressive genomic technique for sequencing all of the protein-coding genes in a genome. The human has almost 180,000 exons, creating about 1% of the human genome or nearly 30 million base pairs. The main approach is to identify genetic variants that alter protein sequences. Since these variants are most studied due to their protein-coding property, it is important to know pathogenicity of all those variants before it is studied on population (Yang et al. 2013).

Over the last decade, many tools and software are developed introduced to predict the functional and structural prioritizations of the variants. This is also known as computational analysis of genetic variants. However, the availability of multiple computational methods which operates on diverse principles to classify deleterious variants has further complicated the users to learn the input and output formats and interpretation of results for every computational tool. Moreover, analyzing variants on individual computational tools and preparing the prediction results in data sheets are very laborious as well as time-consuming. In this regard, the Variant Effect Predictor computational tool hosted by Ensembl acts as a powerful integrative platform which can be easily used by end users for entering the data and interpreting the prediction outcomes easily.

6.2 Ensemble Variant Effect Predictor (VEP)

The Ensembl Variant Effect Predictor (VEP) is a website which hosts a group of computational webservers (Table 6.1) used to study, annotate, and prioritize genomic variations in coding and noncoding regions. VEP is open-source and free and supports full reproducibility of results. It can very well accelerate the interpretation of the variants in a wide range of research projects (McLaren et al. 2016).

Online VEP offers access to a broad collection of tools for genomic annotation. The flexible interface could be set as per the demands of the study by configuring simple preferences. This helps to accommodate the diverse requirements of a study. The effect of the variations like SNPs or deletions or insertions on the genes or gene products or the regulatory sequences could be calculated using VEP.

Table 6.1 List of tools available in variant effect predictor (VEP)

Tool name	Pathogenic range	Principle	Web link
SIFT	Scores range from 0 to 1. The smaller the score, the more likely the SNP has damaging effect	A SIFT score predicts whether an amino acid substitution affects protein function	https://ionreporter.thermofisher.com/ionreporter/help/GUID-2097F236-C8A2-4E67-862D-0FB5875979AC.html
Polyphen2	The score ranges from 0 to 1. Most damaging (largest) [0.52844,0.89865]), "P" ("possibly damaging") "B" ("benign" HDIV score in [0,0.452] or rankscore in [0.02634,0.34268]) deleterious" if the HDIV score is larger than 0.5	It is a tool which predicts possible impact of an amino acid substitution on the structure and function of a human protein using straightforward physical and comparative considerations	http://genetics.bwh.harvard.edu/pph2/
LRT:	Ranges from 0 to 1. The scores range from 0.00162 to 0.84324	This LRT statistic approximately follows a chi-square distribution. To determine if the difference in likelihood scores between the two models is statistically significant, we next must consider the degrees of freedom. In the LRT, degrees of freedom is equal to the number of additional parameters in the more complex model	http://evomics.org/resources/likelihood-ratio-test/
MutationTaster_score	Ranges from 0 to 1. 0.08979– 0.81033.	The Grantham matrix does not provide values for an amino acid insertion/deletion, no *score* is given in such cases. The *score* is only displayed for informational reasons and does not influence the *MutationTaster* prediction as generated by our Bayes classifier	http://www.mutationtaster.org/info/documentation.html

Table 6.1 (continued)

MutationAssessor_pred:	H, N, L are 3.5, 1.935, and 0.8, respectively. The rankscore cutoffs between "H" and "M", "M" and "L", and "L" and "N" are 0.92922, 0.51944, and 0.19719, respectively		
FATHMM_pred:	If a FATHMMori score is ≤−1.5 the corresponding nsSNV is predicted as "D(AMAGING)"	Predicting the functional consequences of both coding variants, i.e., non-synonymous single nucleotide variants (nsSNVs), and noncoding variants	http://fathmm.biocompute.org.uk/
PROVEAN_pred	If PROVEANori ≤ −2.5 (rankscore ≥ 0.543), the corresponding nsSNV predicted as "D(amaging)"	Predicts whether an amino acid substitution or indel has an impact on the biological function of a protein	http://provean.jcvi.org/index.php
VEST3_score *VEST (Variant Effect Scoring Tool)*	VEST 3.0 score. Score ranges from 0 to 1. The larger the score, the more likely the mutation may cause functional change	It predicts the functional significance of missense mutations based on the probability that they are pathogenic	https://karchinlab.org/apps/appVest.html
MetaSVM_pred:	The rankscore cutoff between "D" and "T" is 0.82268	Achieved the highest discriminative power compared to all 18 existing deleteriousness prediction *scores*, which demonstrated the value of combining information from multiple orthologous approaches	http://wglab.org/members/15-member-detail/36-coco-dong
MetaLR_pred	The score cutoff between "D" and "T" is 0.5. The rankscore cutoff between "D" and "T" is 0.81113		
Reliability_index	Ranges from 1 to 10		
M-CAP_pred	The score cutoff between "D" and "T" is 0.025	Aims to misclassify no more than 5% of pathogenic variants while aggressively reducing the list of variants of uncertain significance	

(continued)

Table 6.1 (continued)

REVEL_score	Scores range from 0 to 1	Predicts the pathogenicity of missense variants on the basis of individual tools	https://omictools.com/revel-tool
MutPred_score	Scores range from 0 to 1. The larger the score, the more likely the SNP has damaging effect	Predicts the pathogenicity of amino acid substitutions and their molecular mechanisms	http://mutpred.mutdb.org/
MutPred_Top5features	MutPred_score >0.5 and $p < 0.05$ are referred to as actionable hypotheses MutPred_score >0.75 and $p < 0.05$ are referred to as confident hypotheses. MutPred_score >0.75 and $p < 0.01$ are referred to as very confident hypotheses		
CADD_phred	This is phred-like rankscore based on whole genome CADD raw scores. The larger the score, the more likely the SNP has damaging effect	It is a method that integrates the information from many various functional annotations and condenses this information into a single score	http://epilepsygenetics.net/2015/07/15/here-is-why-cadd-has-become-the-preferred-variant-annotation-tool/
DANN_score	Scores range from 0 to 1. A larger number indicates a higher probability to be damaging	Aims to recognize pathogenic variants by annotating genetic variants, and especially noncoding variants	https://omictools.com/dann-tool
Fathmm MKL_coding_score	Scores range from 0 to 1. SNVs with scores >0.5 are predicted to be deleterious, and those <0.5 are predicted to be neutral or benign. Scores close to 0 or 1 are with the highest confidence.	To predict the functional consequences of both coding and noncoding sequence variants	https://www.ncbi.nlm.nih.gov/pubmed/25583119
Fathmm-MKL_coding_pred	Fathmm-MKL_coding_score is >0.5 (or rankscore >0.28317) the corresponding nsSNV is predicted as “D(AMAGING)”		

Table 6.1 (continued)

Eigen-PC-raw_rankscore	The rankscore is the ratio of the rank of the score over the total number of Eigen-PC-raw scores in dbNSFP	Scoring variants which does not make use of labeled training data. It is useful in prioritizing likely causal variants in a region of interest when it is combined with population-level genetic data in the framework of a hierarchical model	https://omictools.com/eigen-tool
GenoCanyon_score_rankscore	The rankscore is the ratio of the rank of the score over the total number of GenoCanyon_score scores in dbNSFP	Predicts many of the known functional regions and its generalizable statistical framework	https://omictools.com/search?q=GenoCanyon
integrated_confidence_value	0 – highly significant scores (approx. $p < 0.003$); 1 – significant scores (approx. $p < 0.05$); 2 – informative scores (approx. $p < 0.25$); 3 – other scores (approx. $p > =0.25$)	Integrates functional assays (such as ChIP-Seq) with selective pressure inferred using the INSIGHT method. The result is a score ρ in the range [0.0–1.0] that indicates the fraction of genomic positions evincing a particular pattern (or “fingerprint”) of functional assay results that are under selective pressure	http://compgen.cshl.edu/fitCons/
GM12878_confidence_value	0 – highly significant scores (approx. $p < 0.003$); 1 – significant scores (approx. $p < 0.05$); 2 – informative scores (approx. $p < 0.25$); 3 – other scores (approx. $p > =0.25$)		
H1-hESC_confidence_value	0 – highly significant scores (approx. $p < 0.003$); 1 – significant scores (approx. $p < 0.05$); 2 – informative scores (approx. $p < 0.25$); 3 – other scores (approx. $p > =0.25$)		

(continued)

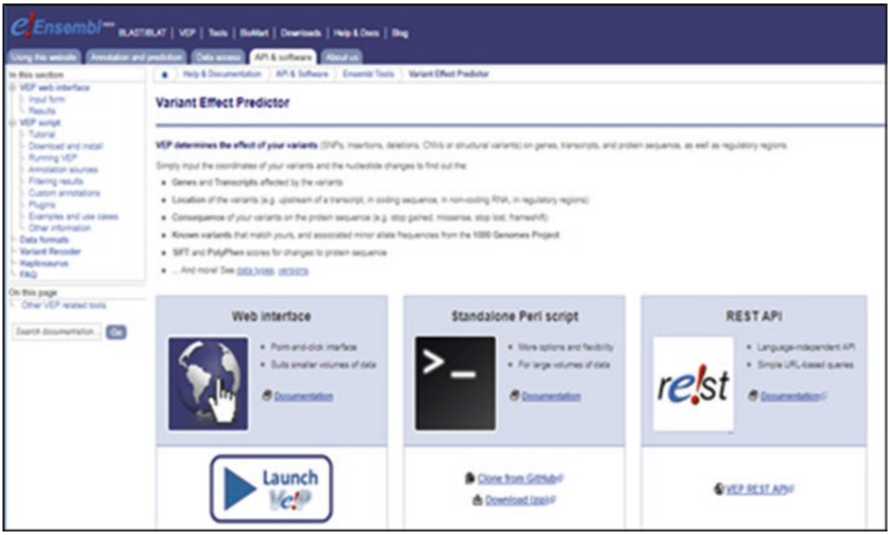
Table 6.1 (continued)

HUVEC_confidence_value	0 – highly significant scores (approx. $p < 0.003$); 1 – significant scores (approx. $p < 0.05$);
	2 – informative scores (approx. $p < 0.25$);
	3 – other scores (approx. $p > =0.25$).		

GERP++_RS	Scores range from −12.3 to 6.17	Identifies constrained elements in multiple alignments by quantifying substitution deficits	http://mendel.stanford.edu/SidowLab/downloads/gerp/
phyloP100way_vertebrate	Scores range from −20.0 to 10.003 in dbNSFP	Measures evolutionary conservation at individual alignment sites	https://ionreporter.thermofisher.com/ionreporter/help/GUID-03D1F68A-E646-4B49-AD59-AF2F51874BD2.html
phyloP20way_mammalian	Scores range from −13.282 to 1.199 in dbNSFP		
phastCons100way_vertebrate	Scores range from 0 to 1	Conservation scoring and identification of conserved elements	http://compgen.cshl.edu/phast/
SiPhy_29way_logOdds	Scores range from 0 to 37.9718 in dbNSFP	Identifies bases under selection from multiple alignment data via rigorous implemented statistical tests	https://omictools.com/siphy-tool

It is considers three different aspects: (A) web interface, (B) stand-alone Perl script, and (C) REST API (Fig. 6.1). The VEP is coded in Perl programming language and is available as an Ensembl API. To increase the speed of execution, the time-critical parts are coded in C and integrated into the API using the XS framework. Chronological blocks of variants are stored in an input memory buffer. All the variants are transformed into an Ensembl Variation Feature objects that point to a genetic location and the alleles. Variants in different file formats like tab enclosed or collision formats are changed directly to objects. HGVS annotation is mapped to their genomic location by removing the applicable reference feature like the protein or transcripts or chromosomes using the Ensembl API.

Preprocessing of the VCF input is done to justify the treatment of unbalanced substitutions and indels due to the dissimilarities in how VCF and Ensembl characterize them. The input buffer is divided among several sub-processes when using the VEP's diverging functionality. After performing the calculations, the result of each sub-process is then formatted into a combined output according to the instruction given in the input.



Variant Effect Predictor

VEP determines the effect of your variants (SNPs, insertions, deletions, CNVs or structural variants) on genes, transcripts, and protein sequence, as well as regulatory regions.

Simply input the coordinates of your variants and the nucleotide changes to find out the:

- Genes and Transcripts affected by the variants
- Location of the variants (e.g. upstream of a transcript, in coding sequence, in non-coding RNA, in regulatory regions)
- Consequence of your variants on the protein sequence (e.g. stop gained, missense, stop lost, frameshift)
- Known variants that match yours, and associated minor allele frequencies from the 1000 Genomes Project
- SIFT and PolyPhen scores for changes to protein sequence
- … And more! See data types, versions

Web interface

- Point-and-click interface
- Suits smaller volumes of data

Documentation

Standalone Perl script

- More options and flexibility
- For large volumes of data

Documentation

Clone from GitHub

Download (zip)

REST API

- Language-independent API
- Simple URL-based queries

Documentation

VEP REST API

Fig. 6.1 The VEP web page with three different aspects of using VEP

For recounting variant significance, the standardized sequence ontology (SO) terms are used. VEP results could be obtained in the VCF format. There is an ongoing effort to achieve a comprehensive variant annotation data exchange format within the Global Alliance for Genomic Health (GA4GH). Moreover, the GA4GH has described standards for demonstration of associations between variants and phenotypes, diseases, and traits. The VEP will accommodate these provisions when the alliance advances them. Present annotation tools are blind to the effects of multiple allele mergers through the multiple variant loci. This restriction that these tools annotate each input variant individually prohibits taking into account the effect of having multiple variants disturbing the matching codon or a change in the reading frame being modified by a downstream variant. In the future, such limitations are expected to be overcome since VEP is actively developed and maintained. New features are consistently additional to both the plugin library and the core VEP code. These expansions are driven by the emerging new interpretations of the datasets available for *H. sapiens.*

The Ensembl's VEP offers toolsets to methodically analyze, prioritize, and annotate variants in both large sequencing projects and minor analyses. By automating the process of annotation in a standardized manner, VEP reduces the required time for physical review. This, in turn, supports the management of many of the collective challenges related to SNVs' analysis, copy number variants, short insertions-deletions, and structural variants. The VEP annotates variants using various reference data, transcripts, citations, regulatory regions, clinical consequences, and estimates of the biophysical significance of variants. The characteristics of variant annotation gained depend on the choice of transcript set used. VEP offers multiple options to format result output and thereby decreases the number of variants requiring manual review. This increases efficiency in processing high volume numbers of variant annotations and transcript isoforms.

6.3 Biological Databases and Computational Methods Comprised in VEP

6.3.1 UniProt: The Universal Protein Knowledgebase

The UniProt is a database of protein sequences and related complete annotation. The knowledgebase comprises in excess of 60 million sequences, of which above half a million sequences have been curated by specialists who judgmentally review experimental and expected data for each protein (Apweiler et al. 2004).

6.3.2 TrEMBL

UniProtKB/TrEMBL is an automated annotated protein sequence database. It translates all coding sequences present in the EMBL/GenBank/DDBJ nucleotide sequence databases. Additionally, protein sequences extracted from the literature are submitted to UniProtKB/Swiss-Prot. This database automatically classifies and annotates the protein sequences added to it (Bairoch and Apweiler 2000).

6.3.3 UniParc

The UniProt Archive (UniParc) is a wide-ranging and nonredundant databank that encompasses most of the freely obtainable protein sequences. Proteins may occur in altered source databanks and several replicas in the same databank. UniParc can escape such severance by storing each distinctive sequence and giving it a steady and unique identifier (UPI), creating it likely to find the same protein from a diverse source of databases (Sharma 2013).

6.3.4 CSN

Clinical sequencing nomenclature (CSN) is a nomenclature developed by researchers to standardize the naming convention for the variations which is in accordance with the ideologies of the Human Genome Variation Society (HGVS) guidelines (Münz et al. 2015).

6.3.5 Pfam

The Pfam database is a big assortment of protein families, each signified by hidden Markov models (HMMs) and multiple sequence alignments (Finn et al. 2014).

6.3.6 PROSITE

PROSITE is a data storage of protein domains and families. It is mainly based on the scrutiny that, while there is enormous number of diverse proteins, most of them can be assembled, based on comparisons in their sequences, into a restricted number of families. Proteins or protein domains belonging to a specific family mostly share functional characteristics and are consequent from a common ancestor (Hulo et al. 2006).

6.3.7 InterPro

InterPro is a source that delivers a functional analysis of protein sequences by categorizing these sequences into families basing on the expected presence of domains and significant sites. To categorize proteins in this way, InterPro uses analytical models, known as signatures, provided by several different member databases that structure the InterPro consortium (Hunter et al. 2009).

6.3.8 Sift

Sorting Intolerant From Tolerant (SIFT) calculates the probable impact of the substitution of amino acid on the function of the protein based on a set of rules. SIFT analyzes the possible effect an amino acid substitution will have on protein function by calculating the sequence homology. These estimations are based on the hypothesis that within a given protein sequence, the significant positions are evolutionarily (Sim et al. 2012).

6.3.9 PolyPhen-2

PolyPhen-2 calculates the probable impact of an amino acid change on the performance of a human protein (Adzhubei et al. 2013). This tool predicts the position-specific independent count (PSIC) score for every variation and calculates the score variance between variants. The higher the PSIC score variance, the higher the efficient impacts of a particular amino acid replacement.

6.3.10 dbNSFP

dbNSFP is a tool developed for well-designed annotation and prediction of all possible non-synonymous single-nucleotide variants (nsSNVs) in the human genome (Liu et al. 2013). Its present edition is based on the Ensembl version 79/GENCODE release 22 and contains a total of 83, 422, 341 non-synonymous SNVs and splicing-site SNVs. It collects prediction scores from 20 prediction algorithms such as Polyphen2-HDIV, SIFT, MutationTaster2, Polyphen2-VAR (Schwarz et al. 2014), LRT, Mutation Assessor (Reva et al. 2011), MetaSVM (Glanzmann et al. 2016), FATHMM (Kim et al. 2017), MetaLR (Dong et al. 2015), VEST3 (Kircher et al. 2014), CADD (Carter et al. 2013), PROVEAN (Choi et al. 2015), fitCons (Gulko et al. 2015), FATHMM-MKL coding, DANN (Quang et al. 2015), Eigen coding (Lu et al. 2015), Eigen-PC, GenoCanyon (Ionita-Laza et al. 2016), M-CAP (Jagadeesh et al. 2016), MutPred (Ioannidis et al. 2016), REVEL (Pejaver et al. 2017). The dbNSFP also provides the detailed information about conservation scores (phastConsx2, PhyloPx2, SiPhyand GERP++) and other related evidence including allele frequencies perceived in the 1000 Genomes Project phase 3 data (Project T 1000 G et al. 2015), UK10K connections data (https://www.uk10k.org/), gnomAD data, ExAC consortium data (Karczewski et al. 2017) and the NHLBI Exome Sequencing Project ESP6500 data, functional descriptions of genes, various gene IDs from different databases, gene expression and gene interaction information.

6.3.11 Condel

Condel is a scheme to evaluate the consequence of non-synonymous SNVs using a Consensus Deleteriousness score that chains various tools (Mutation Assessor, FATHMM) (González-Pérez and López-Bigas 2011).

6.3.12 LoFtool

This tool arranges the loss-of-function (LoF) mutations based on their genomic context and their relevance to susceptibility to disease. The ordering is done based on the Exome Aggregation Consortium (ExAC) dataset for the candidate disease-causing gene (Fadista et al. 2017).

6.3.13 ExAc

The Exome Aggregation Consortium (ExAC) is an alliance of researchers which attempts cumulatively harmonizing the exome sequencing data from diverse resources of large-scale sequencing projects. The intention is to prepare a summarized data accessible to the broad scientific community. The dataset on this website contains 60,706 discrete individuals sequenced as part of several disease-specific and population genetic research studies (Karczewski et al. 2017).

6.3.14 MaxEntScan

MaxEntScan is based on the "maximum entropy principle" where the sequences of short motifs such as those involved in RNA splicing parallelly account for nonadjacent or non-neighboring as well as neighboring dependencies between sequences to build a model. This method simplifies the predictable probabilistic models of sequence motifs such as inhomogeneous Markov models and weight matrix models (Jian et al. 2014).

6.4 Variant Effect Predictor (VEP) Analysis by Web Interface

6.4.1 Description of Data Input Form

Once the user reaches the VEP web interface, an input form will be presented to enter data and alter various options and filters. Input form contains the following entries and selections:

(i) Species of the data

- Genomic data of 101 different species including human (*Homo sapiens*)

(ii) Name of the job

- Alphabetical and/or numerical letter (i.e., PAK3_rs121434612)

(iii) Data uploading

- Paste the data with any of the following formats (Ensembl default, VCF, variant identifiers, HGVS notations) (i.e., rs121434612)
- Or upload file with any of the abovementioned formats
- Or provide file URL of publically accessible address
- Select transcript database (e.g., Ensemble, GENCODE, RefSeq NCBI)

(iv) Identifier and variants of frequency data

- Gene symbol (as HGNC) of the gene to the output
- Consensus CDS identifier for a core set of Mouse and Human proteins
- Ensemble protein identifier
- UniProt for translated protein products from SWISSPROT, TREMBL, and UniParc
- HGVS for generate notation of coding sequence (HGVSc) and protein sequence (HGVSp)
- CSN for generating clinical sequencing nomenclature

(v) Frequency data for co-located variants

- This helps report known variants from the Ensemble variation database that overlaps with the input
- Allelic frequency data from major genotyping projects (i.e., 1000 Genomes global, 1000 Genomes continental, Exome Sequencing Project for African-American and European-American populations, Genome Aggregation Database)

(vi) Extra options (pathogenicity predictions; regulatory region consequences; amino acid conservation).

- Transcript biotype add equivalent to VEP script
- Protein domains, to report protein domains from Pfam, PROSITE, and InterPro tools
- Exon and intron numbers
- Transcript support level
- SIFT, based on the physical properties of amino acids, helps predict the possible substitutions of the amino acids which could affect the protein function
- PolyPhen predicts possible impact of an amino acid substitution on the protein structure and function using physical comparative considerations
- dbNSFP provides pathogenicity predictions for missense variants from various algorithms
- ConDel (Consensus Deleteriousness) scores for a missense mutation based on pre-calculated SIFT and PolyPhen scores
- LoFtool calculates, based on the ratio of loss-of-function to synonymous mutations in ExAC data, the rank of genic intolerance and following susceptibility to disease
- Regulatory data, to get regulatory consequences of variants that overlap regulatory features and transcription factor binding motifs
- dbscSNV, to retrieve data for splicing variants from a tabix-indexed dbscSNV file
- MaxEntScan, to predict sequence motifs and maximum entropy based splice sites consensus predictions
- BLOSUM62, to report amino acid conservation score

- Ancestral allele, to retrieve the ancestral allele for variants inferred from the Ensembl Compara Enredo-Pecan-Ortheus (EPO) pipeline

(vii) Other filtering options

- Filter by frequency, to exclude common variants to remove input variants that overlap with known variants that have a minor allele frequency greater than 1% in the 1000 Genomes Phase 1 combined population
- Use advance filtering to change the population, frequency threshold, and other parameters
- Return results for variants in coding regions only, excluding intronic and intergenic regions
- Restrict results by the severity of consequences that is determined subjectively by Ensembl (Fig. 6.2)

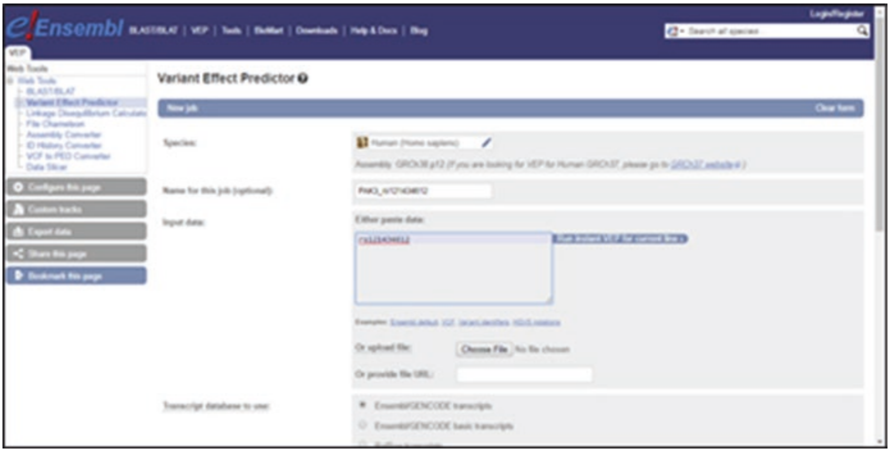
6.4.2 Description of Results and Output

The VEP displays both summary and detailed preview of results on the results page.

6.4.2.1 Summary Details

This panel gives basic statistics of the result, including a brief overview of the VEP job (Fig. 6.3a).

Statistics listed include:



Variant Effect Predictor

Species: Human (Homo sapiens)

Assembly: GRCh38.p12 (If you are looking for VEP for Human GRCh37, please go to GRCh37 website.)

Name for this job (optional): PAK3_rs121434612

Input data: Either paste data: rs121434612

Examples: Ensembl default, VCF, Variant identifiers, HGVS notations

Or upload file: Choose File No file chosen

Or provide file URL:

Transcript database to use:
● Ensembl/GENCODE transcripts
○ Ensembl/GENCODE basic transcripts
○ RefSeq transcripts

Fig. 6.2 New job entry in VEP web interface for PAK3 gene variant rs121434612

- Variants processed – any unprocessed variants are not included
- Variants remaining after filtering
- Novel/known variants – this shows the number and percentage of novel variants over the existing variants
- Number of overlaps found for genes, transcripts, and regulatory regions

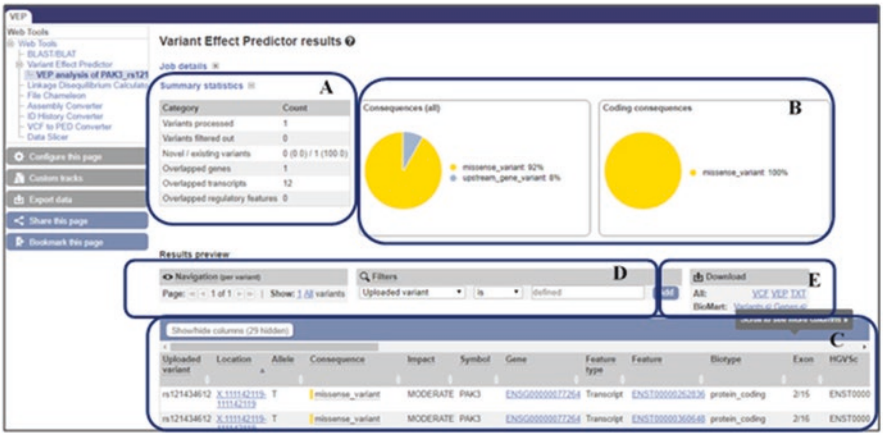
6.4.2.2 Information in Pie Charts' Preview

Pie charts display the proportion of consequence types called across all variants transversely in the results. The color scheme of the graph matches the colors used to display variants in detail view (Fig. 6.3b).

The results' page displays all of the columns by default. To hide columns, the "Show/hide columns" button, which is blue in color, could be clicked to select the user's choice. The user-selected columns could be recalled when viewing other jobs.

6.4.2.3 Results Description in the Preview Table

The table of results displays one raw per transcript and variant. The default setting shows all of the columns, but as described previously, the user can hide the columns. Column headers could be clicked to order sorting as per the user's need. The table



Category	Count
Variants processed	1
Variants filtered out	0
Novel / existing variants	0 (0.0) / 1 (100.0)
Overlapped genes	1
Overlapped transcripts	12
Overlapped regulatory features	0

Uploaded variant	Location	Allele	Consequence	Impact	Symbol	Gene	Feature type	Feature	Biotype	Exon	HGVSc
rs121434612	X:111142119-111142119	T	missense_variant	MODERATE	PAK3	ENSG00000077264	Transcript	ENST00000262836	protein_coding	2/15	ENST0000
rs121434612	X:111142119-111142119	T	missense_variant	MODERATE	PAK3	ENSG00000077264	Transcript	ENST00000360648	protein_coding	2/16	ENST0000

Fig. 6.3 Description of obtained results after VEP's web interface analysis for (**a**) PAK3 gene variant summary preview, (**b**) pie chart preview, (**c**) results in preview, (**d**) navigations of results' pages, (**e**) downloading the results

can be downloaded as a spreadsheet by clicking the top right corner spreadsheet icon (Fig. 6.3c).

Navigating Results

The result pages could be scrolled using the navigation panel. Five variants are displayed by default (Fig. 6.3d). It has to be noted that since there can be an overlap between variant and multiple transcripts, the table will often display more than five rows. The relevant link could be clicked to change the number of rows shown. As a caution, when a large input file is used, it is advised to filter the results before displaying them. This will avoid the unresponsiveness of the browser when it tries to load all the results given in the table. The arrow icons could be used for navigating through the results.

Downloading the Results

The VEP allows selecting and downloading full or filtered results (Fig. 6.3e):

- VCF: It is a portable format for variant data. This format stores the consequence data as a series of delimited strings
- VEP: This is the default VEP output format which gives one row per variant and transcript overlap
- TXT: This is a plain text format, which is the tab-delimited format. All the columns are present in the output irrespective of the selection made by the user. This format is useful for import into a spreadsheet like Microsoft Excel

6.5 Conclusion

In the age of medical and clinical genomics, SNV prioritization has become more important. This task can be performed by many computational tools separately or collectively. The Variant Effect Predictor (VEP) can now facilitate the accurate assessment of SNVs for clinical diagnostic as well as the genetic disease discovery programs. However, researchers who use VEP should comprehend how to interpret the prediction outcomes and limitations of the computational tools. Moreover, the predictions should be interpreted with knowledge regarding SNVs characteristics and properties. The results obtained from the VEP assessment need to correlate with previously defined clinical characters by translational research studies. Additionally, it is also likely to benefit the research studies currently underway on assessing the consequences of genomic variants for various cancers and genetic diseases with new insights on the medical relevance of SNVs.

References

Adzhubei I, Jordan DM, Sunyaev SR (2013) Predicting functional effect of human missense mutations using PolyPhen-2. Curr Protoc Hum Genet:1–41

Antonarakis SE, Beckmann JS (2006) Mendelian disorders deserve more attention. Nat Rev Genet 7:277–282

Apweiler R, Bairoch A, Wu CH, Barker WC, Boeckmann B, Ferro S et al (2004) UniProt: the universal protein knowledgebase. Nucleic Acids Res 32(Database issue):D115–D119. Available from: https://www.ncbi.nlm.nih.gov/pubmed/14681372

Bairoch A, Apweiler R (2000) The SWISS-PROT protein sequence database and its supplement TrEMBL in 2000. Nucleic Acids Res 28(1):45–48. Available from: https://www.ncbi.nlm.nih.gov/pubmed/10592178

Carter H, Douville C, Stenson PD, Cooper DN, Karchin R (2013) Identifying Mendelian disease genes with the variant effect scoring tool. BMC Genomics 14(Suppl 3):S3. Available from:https://www.ncbi.nlm.nih.gov/pubmed/23819870

Choi Y, Chan AP, Craig TJ (2015) PROVEAN web server: a tool to predict the functional effect of amino acid substitutions and indels. Bioinformatics 31:2745–2747

Cirulli ET, Goldstein DB (2010) Uncovering the roles of rare variants in common disease through whole-genome sequencing. Nat Rev Genet 11:415. Available from: https://doi.org/10.1038/nrg2779

Consortium IH (2003) The international HapMap project. Nature 426:789–796

Dewey FE, Grove ME, Pan C, Goldstein BA, Bernstein JA, Chaib H et al (2014) Clinical interpretation and implications of whole-genome sequencing. JAMA 311(10):1035–1045. Available from: https://doi.org/10.1001/jama.2014.1717

Dong C, Wei P, Jian X, Gibbs R, Boerwinkle E, Wang K et al (2015) Comparison and integration of deleteriousness prediction methods for nonsynonymous SNVs in whole exome sequencing studies. Hum Mol Genet 24(8):2125–2137. Available from: https://www.ncbi.nlm.nih.gov/pubmed/25552646

Fadista J, Oskolkov N, Hansson O, Groop L (2017) LoFtool: a gene intolerance score based on loss-of-function variants in 60 706 individuals. Bioinformatics 33(4):471–474. Available from: https://doi.org/10.1093/bioinformatics/btv602

Finn RD, Bateman A, Clements J, Coggill P, Eberhardt Y, Eddy SR et al (2014) Pfam : the protein families database. Nucleic Acids Res 42:222–230

Glanzmann B, Herbst H, Kinnear CJ, Möller M, Gamieldien J, Bardien S (2016) A new tool for prioritization of sequence variants from whole exome sequencing data. Source Code Biol Med 11(1):10. Available from: https://doi.org/10.1186/s13029-016-0056-8

González-Pérez A, López-Bigas N (2011) Improving the assessment of the outcome of nonsynonymous SNVs with a consensus deleteriousness score, Condel. Am J Hum Genet 88(4):440–449. Available from: https://www.ncbi.nlm.nih.gov/pubmed/21457909

Gulko B, Hubisz MJ, Gronau I, Siepel A (2015) A method for calculating probabilities of fitness consequences for point mutations across the human genome. Nat Genet 47(3):276–283. Available from: https://www.ncbi.nlm.nih.gov/pubmed/25599402

Hulo N, Bairoch A, Bulliard V, Cerutti L, De Castro E, Langendijk-Genevaux PS et al (2006) The PROSITE database. Nucleic Acids Res 34(suppl_1):D227–D230. Available from: https://doi.org/10.1093/nar/gkj063

Hunter S, Apweiler R, Attwood TK, Bairoch A, Bateman A, Binns D et al (2009) InterPro: the integrative protein signature database. Nucleic Acids Res 37(suppl_1):D211–D215. Available from: https://doi.org/10.1093/nar/gkn785

International Human Genome Sequencing Consortium (2004) Finishing the euchromatic sequence of the human genome. Nature 431(7011):931–945

Ioannidis NM, Rothstein JH, Pejaver V, Middha S, McDonnell SK, Baheti S et al (2016) REVEL: an ensemble method for predicting the pathogenicity of rare missense variants. Am J Hum Genet 99(4):877–885. Available from: https://www.ncbi.nlm.nih.gov/pubmed/27666373

Ionita-Laza I, McCallum K, Xu B, Buxbaum JD (2016) A spectral approach integrating functional genomic annotations for coding and noncoding variants. Nat Genet 48(2):214–220. Available from: https://www.ncbi.nlm.nih.gov/pubmed/26727659

Jagadeesh KA, Wenger AM, Berger MJ, Guturu H, Stenson PD, Cooper DN et al (2016) M-CAP eliminates a majority of variants of uncertain significance in clinical exomes at high sensitivity. Nat Genet 48:1581. Available from: https://doi.org/10.1038/ng.3703

Jian X, Boerwinkle E, Liu X (2014) In silico tools for splicing defect prediction: a survey from the viewpoint of end users. Genet Med 16(7):497–503. Available from: https://www.ncbi.nlm.nih.gov/pubmed/24263461

Karczewski KJ, Weisburd B, Thomas B, Solomonson M, Ruderfer DM, Kavanagh D et al (2017) The ExAC browser: displaying reference data information from over 60 000 exomes. Nucleic Acids Res 45(D1):D840–D845. Available from: https://www.ncbi.nlm.nih.gov/pubmed/27899611

Kim S, Jhong J-H, Lee J, Koo J-Y (2017) Meta-analytic support vector machine for integrating multiple omics data. BioData Min 10:2. Available from: https://www.ncbi.nlm.nih.gov/pubmed/28149325

Kircher M, Witten DM, Jain P, O'Roak BJ, Cooper GM, Shendure J (2014) A general framework for estimating the relative pathogenicity of human genetic variants. Nat Genet 46(3):310–315. Available from: https://www.ncbi.nlm.nih.gov/pubmed/24487276

Liu X, Jian X, Boerwinkle E (2013) dbNSFP v2.0: a database of human non-synonymous SNVs and their functional predictions and annotations. Hum Mutat 34(9):E2393–E2402. Available from: https://doi.org/10.1002/humu.22376

Lu Q, Hu Y, Sun J, Cheng Y, Cheung K-H, Zhao H (2015) A statistical framework to predict functional non-coding regions in the human genome through integrated analysis of annotation data. Sci Rep 5:10576. Available from: https://www.ncbi.nlm.nih.gov/pubmed/26015273

McLaren W, Gil L, Hunt SE, Riat HS, Ritchie GRS, Thormann A et al (2016) The Ensembl variant effect predictor. Genome Biol 17:1):1–1)14. Available from: https://doi.org/10.1186/s13059-016-0974-4

Münz M, Ruark E, Renwick A, Ramsay E, Clarke M, Mahamdallie S et al (2015) CSN and CAVA: variant annotation tools for rapid, robust next-generation sequencing analysis in the clinical setting. Genome Med 7(1):76. Available from: https://doi.org/10.1186/s13073-015-0195-6

Pejaver V, Urresti J, Lugo-martinez J, Kymberleigh A, Lin GN, Nam H et al (2017) MutPred2: inferring the molecular and phenotypic impact of amino acid variants, pp 1–28

Project T 1000 G, Auton A, Abecasis GR, Altshuler DM, Durbin RM, Abecasis GR et al (2015) A global reference for human genetic variation. Nature 526:68. Available from: https://doi.org/10.1038/nature15393

Quang D, Chen Y, Xie X (2015) DANN: a deep learning approach for annotating the pathogenicity of genetic variants. Bioinformatics 31(5):761–763. Available from: https://www.ncbi.nlm.nih.gov/pubmed/25338716

Reva B, Antipin Y, Sander C (2011) Predicting the functional impact of protein mutations: application to cancer genomics. Nucleic Acids Res 39(17):e118. Available from: https://doi.org/10.1093/nar/gkr407

Sachidanandam R, Weissman D, Schmidt SC, Kakol JM, Stein LD, Marth G et al (2001) A map of human genome sequence variation containing 1.42 million single nucleotide polymorphisms. Nature 409:928. Available from: https://doi.org/10.1038/35057149

Schmutz J, Wheeler J, Grimwood J, Dickson M, Yang J, Caoile C et al (2004) Quality assessment of the human genome sequence. Nature 429(6990):365–368

Schwarz JM, Cooper DN, Schuelke M, Seelow D (2014) MutationTaster2: mutation prediction for the deep-sequencing age. Nat Methods 11:361. Available from: https://doi.org/10.1038/nmeth.2890

Sharma R (2013) Birth defects in India: hidden truth, need for urgent attention. Indian J Hum Genet 19(2):125–129. Available from: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3758715&tool=pmcentrez&rendertype=abstract

Sim N, Kumar P, Hu J, Henikoff S, Schneider G, Ng PC (2012) SIFT web server : predicting effects of amino acid substitutions on proteins. Nucleic Acids Res 40:452–457

Stenson P, Ball E, Mort M, Phillips A, Shiel J, Thomas N et al (2003) Human gene mutation database (HGMD): 2003 update. Hum Mutat 21(6):577–581

Yang Y, Muzny DM, Reid JG, Bainbridge MN, Willis A, Ward PA et al (2013) Clinical whole-exome sequencing for the diagnosis of Mendelian disorders. N Engl J Med 369(16):1502–1511. Available from: https://doi.org/10.1056/NEJMoa1306555

Chapter 7
Clinical Strategies for Developing Next-Generation Cancer Precision Medicines

Chee Gee See

Contents

7.1 Introduction

The ability to counter the progression of disease by targeting the key drivers of disease progression itself is a hallmark of precision medicine. In many ways, there is nothing especially novel about this approach. In therapeutic areas like infectious disease, the main identified driver of disease is the infectious agent itself (e.g. bacteria or virus), and the approach has always been to discover ways of reducing its presence down to zero. This is extremely precise and extremely effective. Antibiotics targeted against specific bacteria work to eradicate its presence, and therefore its negative effects in the host. And in the parlance of precision medicine development, the bacteria (or whichever pathogen) becomes the diagnostic biomarker itself. Such an approach in infectious disease has always been envied by drug developers in other therapeutic areas, as this approach is eminently simplistic, and the readouts are readily measurable. But disease pathology in most therapeutic areas is

C. G. See (✉)
Life Science Expert at PA Consulting Ltd, 10 Bressenden Place, London, SW1E 5DN, UK

CGS Precision Medicine Consultancy, 5 Broadwater Avenue, Letchworth Garden City,
Hertfordshire, United Kingdom
e-mail: cheegee.see@paconsulting.com

N. A. Shaik et al. (eds.), *Essentials of Bioinformatics, Volume II*,
https://doi.org/10.1007/978-3-030-18375-2_7

notoriously complex, and disease classification had always been necessarily organ-centric, especially in the pre-molecular era. Therefore, in the case of lung cancer even up to the late 1970s, all patients were regarded as having just ‘one’ disease: lung cancer. The advent of advanced molecular tools and techniques from the 1970s onwards revolutionized what we could do to unpack disease pathophysiology and indeed provide significant insights into what constituted a heterogeneous mix of diseases, all previously united by their organ of origin.

Indeed, it was the advent of such advanced molecular techniques that provided the impetus for the precision medicine development arena we see today. We owe much of what we see in precision medicine today to the pioneers of these techniques and the visionaries who were able to extend their use to clinical utility and patient benefit. Herceptin (trastuzumab), arguably the most well-known of all precision medicine drugs had its origins in the wave of exciting new discoveries from molecular techniques in the 1970s. The discovery and identification of the HER2 gene as a major driver in metastatic breast cancer led researchers to postulate if it might be possible to knock out this single gene in the disease state, thereby potentially stopping breast cancer in its tracks. It was a daring hypothesis as it implied that it was possible to apply monogenic principles of gene function in a complex multifactorial disease. And as researchers began accumulating corroborating data, it became more evident that this hypothesis might just be right. It would, however, not be for another 20 years before Herceptin was approved by the FDA in 1998 as a therapy specifically against HER2-overexpressing breast cancer. This 1998 Herceptin FDA approval is a crystallizing moment in clinical drug development history. The year 1998 signals the true start of the clinical development timeline that specifically targeted or precision medicine is and can be a reality for tackling complex diseases such as cancer. Today, 20 years on from 1998, the precision medicine landscape has evolved and matured significantly. Precision medicine development is still largely powered by ever-improving molecular technologies and empowered by clinical visionaries and disease biology experts. We will always require these insights to have the upper hand in the battle against seriously debilitating disease. This review will look at the way in which our approach and strategy in precision medicine clinical development has evolved over the last 20 years. Identifying a key disease-driving gene and then producing a precision medicine product is one thing; how you improve on that product and manage the disease is another. This review will look at the example of the epidermal growth factor receptor (EGFR) gene as a major driver of non-small cell lung cancer (NSCLC) and its pivotal use as a target for next-generation precision medicine development in oncology. The strategies utilized in the generational development will also be examined.

7.2 The EGFR Inhibitor Approach in Non-Small Cell Lung Cancer Precision Medicine Development

The overexpression of the epidermal growth factor receptor or EGFR gene has been consistently implicated in the pathophysiology of different cancers for over 20 years (Salomon et al. 1995; Hirsch et al. 2009). The observation that the EGFR signalling pathway can also activate the MAPK, PI3K and JAK/STAT pathways amongst many other pathways suggests that EGFR may play a crucial early role in tumorigenesis (Grandis and Sok 2004; Lemmon and Schlessinger 2010). In NSCLC, EGFR overexpression in both premalignant and malignant tissues can be as high as 40–80% (Salomon et al. 1995; Grandis and Sok 2004; Merrick et al. 2006), and it therefore made perfect sense to drug developers that an EGFR inhibitor would be the natural answer to counter any one of these EGFR overexpressing cancers. The first EGFR inhibitors to be developed and explored for NSCLC, the so-called first-generation EGFR-tyrosine kinase inhibitor (TKIs), were the small molecules gefitinib and erlotinib and also the chimaeric monoclonal antibody cetuximab. The strategy used here in developing the first generation of NSCLC EGFR small-molecule TKIs was mainly medicinal chemistry in nature. Both gefitinib and erlotinib are reversible competitive inhibitors for ATP for the tyrosine kinase domain of EGFR, this being one of four ways in which TKIs operate (Posner et al. 1994). The working hypothesis was that knocking out EGFR, either through a small molecule or through monoclonal antibody, would result in the blockade of its downstream pathway and therefore its oncogenic consequence. Presumably, the related MAPK, PI3K and JAK/STAT pathways would be knocked out as well. And as the drug developers were careful to ensure that they mitigated against the safety aspect of complete EGFR knockout, a reversible inhibitor would allow for EGFR to at least function in a normal capacity, a sort of EGFR reset capacity. In this early development of EGFR-TKIs, it is almost hard to believe that the emphasis then was to diminish the overexpressing powers of EGFR rather than of the effects of any EGFR mutations that we now know. Hence, the early trials of EGFR-TKIs (such as gefitinib and erlotinib) were in unselected NSCLC populations (Fukuoka et al. 2003; Kris et al. 2003; Pérez-Soler et al. 2004; Pérez-Soler 2004a, b). The gefitinib IRESSA Pan-Asia (IPASS) study provides an excellent snapshot of the prevailing scientific understanding at this time juncture. IPASS was a randomized trial comparing first-line gefitinib against the chemotherapy doublet carboplatin/paclitaxel in 1217 NSCLC adenocarcinoma patients across multiple sites in Asia who were either non-smokers or previous light smokers. The median PFS (primary endpoint) was 5.7 and 5.8 months for gefitinib and carboplatin/paclitaxel respectively in IPASS (Reck et al. 2010). The only conclusion here from such a large dataset was to assert a non-inferiority label of gefitinib over carboplatin/paclitaxel. And in terms of clinical development, this result is not only disappointing, but a complete disaster. No commercial or scientific justification can or will be made by any company to continue developing a novel compound that is only just as good as the standard chemotherapy agents available. If these initial PFS results stood as a testament to the IPASS study, then the pharmaceutical industry EGFR overexpression hypothesis and the EGFR-TKI therapeutic option would be in serious jeopardy. The IPASS study needed more than just the PFS non-inferiority label to progress.

In these early studies, it was either a stroke of genius planning or ingenious luck that additional patient samples for *exploratory* sub-group analyses were taken. Clinical investigators needed to confirm the hypotheses of EGFR overexpression being a dominant driver for NSCLC and therefore the justification for using the

EGFR inhibitor route, but here was an opportunity to explore if other hitherto unknown factors might contribute to patient responses. As it turned out, these exploratory analyses revealed a crucial reason for patient responses. NSCLC patients with somatic EGFR mutations specifically within exons 18–21 seemed to respond better to the EGFR-TKIs than EGFR wild-type (WT) patients. In fact, there was a race at this point in time to be the first to report this finding. In the event, two groups in Boston and a third in New York (Lynch et al. 2004; Paez et al. 2004; Pao et al. 2004) reported in 2004 that mutations of the EGFR gene present in the tumours of NSCLC patients predispose them to better responses. This was the first hint that EGFR-TKIs for NSCLC should be targeting patients with EGFR tumour mutations rather than EGFR overexpressers per se. The mutations within the *golden mile* of exons 18–21 were therefore labelled as 'activating mutations', to distinguish them from other EGFR mutations which made little difference to EGFR-TKI efficacies. These EGFR-activating mutations and the direct link to clinical responses may also explain why selecting NSCLC patients based on EGFR overexpression techniques like immunohistochemistry or fluorescence in situ hybridization copy number did not generate the expected clinical response rates, even if it made biological sense at the time, and the large 40–80% EGFR overexpressing patient population (Salomon et al. 1995; Grandis and Sok 2004; Merrick et al. 2006) made this a very attractive patient stratification strategy. The IPASS clinical study also made a later reference that within the 1217 adenocarcinoma patient cohort, a sub-group analysis of patients with activating mutations showed superiority of gefitinib over carboplatin/paclitaxel (Reck et al. 2010). The median PFS in this IPASS sub-group was 9.5 and 6.3 months for gefitinib and carboplatin/paclitaxel respectively, making an emphatic rewording of the clinical study conclusion from non-inferiority to superiority for gefitinib (Mok et al. 2009).

Indeed, a later study comparing gefitinib against another chemotherapy agent validated the utility of selecting EGFR mutation-positive subjects. The INTEREST study reported improved PFS in EGFR-mutant NSCLC patients on gefitinib over docetaxel. The response rate was also improved by twofold, 42% versus 21% (Kim et al. 2008; Douillard et al. 2008).

An important lesson emerges from this first-generation clinical development of NSCLC EGFR inhibitors. Without the *exploratory* analyses of clinical samples being factored into the study protocols, it is arguable if the study investigators would have discovered the EGFR-activating mutations and its crucial link to improved clinical response. More importantly, the study investigators implemented the exploratory component of the study protocols expediently and explored other potential reasons other than to confirm the EGFR overexpression hypothesis. It is important to remember that no matter how plausible a current biological hypothesis is, and how unattractive implementing a programme of exploratory analysis might be in terms of additional time and budget resource, without actually physically undertaking these 'nice-to-have' analyses and being open to other hypotheses, it will not be possible to gain additional biological or clinical insights. What this demonstrates is that it is very desirable to have a parallel track of exploratory analyses running alongside the 'essential' clinical study. And in the first-generation EGFR-TKI development, it was the ‘nice-to-have’ exploratory analyses that effectively saved the ‘essential’ EGFR clinical programme as it provided the crucial evidence that not only was the NSCLC EGFR-TKI angle driven by mutations and not overexpression, but it was specific mutations within a certain region.

Despite this tranche of EGFR mutation evidence reaching the attention of clinical developers, it was clear that some developers and companies had already invested so heavily on the overexpression strategy that it was difficult to be immediately swayed by the mutation evidence. The EGFR overexpression population was so much bigger than the EGFR mutation population (80% versus 10–30%) that in commercial terms, this reduction in potential market share and sales would appear catastrophic. This genuine struggle to balance the original aspirations of the drug target profile and its commercial objectives is reflected in the final report of the Tarceva Lung Cancer Survival Treatment (TRUST) Phase IV study (Reck et al. 2010). Involving over 6500 patients, Reck et al. (2010) reported that ‘Although patients whose tumors have these mutations are likely to obtain a greater magnitude of benefit from EGFR-TKIs such as erlotinib, it is important to note that the absence of these mutations does not necessarily result in a lack of benefit with erlotinib therapy’. This clear pushback to the greater efficacy of EGFR-TKIs in mutation-positive patients in favour of an all-comers EGFR population is further evidenced by the use of the Disease Control Rate (DCR) measurement, defined as the sum of complete response, partial response *and* stable disease (CR + PR + SD). The DCR in the TRUST study was 69% (3705/6580), and the study authors conclude therefore that there was a favourable survival and safety profile of erlotinib in a global patient population and across a broad range of patient sub-groups. A second extremely important lesson emerges from this study report. It is important for clinical scientists and drug developers to be driven by actual scientific and clinical data and less on aspirations, especially from the commercial perspective. Whilst it is true that drug development has a very clear commercial angle, this must not take precedence over any actual clinical evidence or the emerging clinical picture. In fact, the use of the Disease Control Rate has been very contentious, and one report has even described the use of DCR as being ‘disingenuous’ without any meaningful reference to clinical endpoints (Sznol 2010). In this respect, a lot of time and effort was actually wasted in trying to make the case for an EGFR all-comers population rather than a mutation-positive EGFR population for EGFR-TKIs. This episode serves as a useful lesson and warning that clinical scientists and developers at the forefront of clinical trials who see the clinical data and analyse them must themselves be confident and strong enough to provide the evidence and make the right recommendations and decisions. Clinical scientists and drug developers clearly owe a duty of service to their parent pharmaceutical company, but they must hold fast to their first and foremost duty of care to patients. As it turned out, the case for the EGFR mutation population being more efficacious to EGFR-TKIs was convincingly made with the IPASS study, initiated firstly in 2006 and finally reported in 2011 (Fukuoka et al. 2011).

7.3 Development of Second-Generation EGFR-TKIs for NSCLC

Gefitinib and erlotinib were the first-generation EGFR-TKIs for NSCLC, and the experience with all targeted therapies is that they do work very well but within a short time, secondary resistance kicks in and there is generally relapse. From the IPASS study, the median duration of response to gefitinib was 9.6 months, with data based mainly on the Asian population where the incidence of EGFR mutations in NSCLC was particularly high (Mok et al. 2017). For the IRESSA Follow-Up Measure (IFUM) study, a commitment to the European Medicines Agency to address efficacy in non-Asian patients, the median duration of response to gefitinib was even shorter at 6.0 months when ascertained by a BICR (blinded independent central review) (Kazandjian et al. 2016).

What this means is that there was a huge motivation to develop the second generation of EGFR TKIs for NSCLC that may overcome some or all of the reasons for the limited duration of response. A median duration of response of between 6 and 9 months is not an exceptionally cost-effective value for an innovative high-cost precision medicine, and patients and payers would want to see greater improvements.

First-generation NSCLC EGFR-TKIs were all *reversible* inhibitors utilizing a core 4-anilinoquinazoline scaffold that reversibly inhibited both EGFR mutants and wild-type (WT) EGFR. In the thinking about developing second-generation inhibitors, there was discussion about the option for irreversible inhibitors as opposed to reversible inhibitors. The rationale behind this thinking was that the safety concern by inhibiting EGFR in a reversible manner may have taken off some of the drug potency required for a longer duration of response. This therefore led to a strategic rethink and refocus on the structural attributes of the ideal second-generation EGFR-TKI for NSCLC. If in reconstructing a second-generation EGFR-TKI into an irreversible inhibitor, the drug developers can maintain a clear safety profile, then this would make a compelling case. Additionally, if the new irreversible construct can improve the efficacy profile, then this would make a far greater clinical and commercial case. The second-generation development was therefore clearly led from the chemistry angle and would now contain a Michael acceptor moiety for binding covalently to the thiol group of Cys797 in the ATP-binding domain of EGFR (Castellanos and Horn 2015). By this time, NSCLC disease biology understanding through use of next-generation sequencing techniques had uncovered a series of mutational hotspots on the EGFR gene. The mutational hotspot discovery was to have a huge bearing on subsequent thinking around tackling this disease through the EGFR TKI route. Interestingly, the medicinal chemists employed in developing the second-generation EGFR-TKIs continued to use the anilinoquinazoline core as in the first-generation construction, which targeted both mutant and WT EGFR. There is essentially a hotspot mutational region within EGFR exons 18–21 that first-generation EGFR-TKIs like gefitinib and erlotinib were designed to hit. As more NSCLC patient DNA sequence information became available, it became clear that patients who developed resistance to these first-generation EGFR-TKIs were harbouring specific mutations, especially within exons 19–21. The implication was that these *are* the mutations that are the root cause of the resistance.

In the second-generation construct, attention was therefore focussed on these mutations from exons 19–21 that apparently did not respond well to the first-generation inhibitors. Examples of second-generation NSCLC EGFR-TKIs are afatinib, developed by Boehringer Ingelheim, and dacomitinib, developed by Pfizer. Boehringer Ingelheim conducted a very successful clinical development of afatinib, and it obtained its first FDA approval in July 2013. Boehringer Ingelheim also implemented the contemporaneous development of a companion diagnostic test together with the clinical development of afatinib, the drug-diagnostic *co-development* so favoured by the FDA. In this respect, the diagnostic company Qiagen was engaged to develop a companion diagnostic test at the same time. Ultimately, this led to afatinib (marketed as Gilotrif) being FDA-approved along with the companion diagnostic, Qiagen's Therascreen® EGFR RGQ PCR Kit. This was an exceedingly clever move as two products are now being marketed. Therein lies another important lesson for drug developers: be on the lookout for companion diagnostic opportunities. The Qiagen Therascreen® EGFR RGQ PCR Kit companion diagnostic test specifically targeted just two mutations on the EGFR gene, the exon 19 deletions and the exon 21 L858R substitution mutation, these being the only mutations that qualify the use of afatinib. There were other EGFR mutations already known which were not the 'official diagnostic' target for afatinib or indeed the Therascreen® EGFR RGQ PCR Kit.

Although second-generation EGFR inhibitors appear able to elicit genuine clinical responses from these particular EGFR mutations that were not achievable by first-generation inhibitors, there was a price to pay. Toxicity issues are a known feature of EGFR inhibitors and the second-generation EGFR inhibitors have *greater* toxicity issues than first-generation inhibitors. A meta-analysis by Ding et al. (2017) of 16 different trials comparing first-and second-generation NSCLC EGFR-TKIs showed that, overall, the risk for rash was higher with afatinib (84.8%) than with erlotinib (62.0%) or gefitinib (62.0%), and the risk for diarrhoea was more than double with afatinib (91.7%) than with erlotinib (42.4%) or gefitinib (44.4%). It appears that the improved drug efficacy over the first-generation resistant mutations is limited by the pharmacokinetics of dosing itself—to achieve the additional clinical responses over the first-generation resistant mutations, the second-generation EGFR-TKIs are prescribed at a stronger and more robust dose, thereby generating the unwanted consequence of greater skin and gastrointestinal toxicity issues. Although this is not ideal, second-generation NSCLC EGFR-TKIs have their place, and it certainly allows physicians another option for treating NSCLC patients. However, with anticipated greater skin and gastrointestinal toxicities, NSCLC patients have to be physically fitter to tolerate this treatment regime.

Whilst second-generation EGFR-TKIs were being developed, more understanding of the biology of EGFR mutation-mediated NSCLC was being uncovered. It transpired that amongst the EGFR mutations that were targeted by the first- and second-generation TKIs, one particular mutation, T790M, was resistant to *every* attempt to overcome its effects. This observation led to T790M being referred to as the 'gatekeeper' mutation as it was seen as the ultimate hurdle to clear. Furthermore, it was discovered that although the frequency of T790M amongst treatment-naïve patients was just under 5%, by the time these NSCLC patients had undergone treatment with first- and second-generation EGFR-TKIs, the frequency of T790M had increased to 50% (Inukai et al. 2006; Kobayashi et al. 2005; Pao et al. 2005). It therefore became very clear that the lack of clinical durability of these first- or second-generation EGFR-TKIs was down to the ability of this single mutation to withstand the pharmacological effects of these TKIs. Two conclusions can be immediately drawn from this observation.

1. First- and second-generation TKIs are able to kill off all NSCLC cells with EGFR mutations except for the T790M mutation. The physical space created by the loss of non-T790M cells allows the resistant T790M cells to very quickly multiply and recolonize the available space. This creates a new and dangerously high concentration of T790M cells as a direct result of the initial treatment with EGFR-TKIs.
2. The T790M mutation is pharmacologically completely resistant to the first- or second-generation EGFR-TKIs such as gefitinib or afatinib. To overcome the effects of this mutation, any third-generation TKI must be pharmacologically different from the earlier-generation TKIs to have any clinical effect.

7.4 Development of Third-Generation EGFR-TKIs for NSCLC

The first- and second-generation EGFR-TKIs for NSCLC perform well in treatment-naïve EGFR-mutant patients, but their clinical efficacy is completely curtailed when the T790M EGFR mutation becomes the dominant mutation form in the disease. It is important to stress that a continuing understanding how the disease evolves as a direct consequence of the previous treatment options is a key factor in developing the next-generation EGFR-TKIs, in this case, the third-generation TKIs (Pao and Chmielecki 2010). Current wisdom and understanding of disease progression inform us that when we attack a drug target that is pivotal to the disease, the disease will invariably counteract by switching to a disease pathway that is completely unaffected by the drug (the non-canonical pathway) or by switching to a specific mutation that is resistant to the drug. Either ways, drug developers know that this phenomenon of 'acquired resistance' to the drug is a very real phenomenon. Inevitably, it often is not a question of if the disease is ever going to evolve to a resistant form but *when* this resistance will happen. Diseases such as cancers are especially efficient and adept at developing resistance mechanisms, and it is the prudent drug developer who looks out for a decrease in drug durability as a clue that the disease may have evolved some resistance to the current treatment options.

The discovery therefore that NSCLC patients who relapse after treatment with first- or second-generation EGFR TKIs have a high percentage of their cancer cells manifesting the T790M mutation is important. T790M mutations are rarely found in treatment-naïve patients (Inukai et al. 2006), and their much higher frequencies in the same patient after treatment with first- and second-generation EGFR TKIs suggest that the physical space vacated by cancer cells killed off by the first- and second-generation TKIs was now being clonally infiltrated by these resistant T790M cells. This is the most plausible explanation for why the T790M frequency is only <5% in treatment-naïve patients but rises to >50% in relapsing patients. This discovery also suggested that these T790M cells are unlikely to be significantly affected by the pharmacology utilized in first- and second-generation EGFR-TKIs.

First- and second-generation EGFR-TKIs utilized the 4-anilinoquinazoline scaffold as its core, inhibiting both EGFR mutants and WT EGFR, resulting not only in significant disease control but also with the predictable side effects of rash and diarrhoea (Dungo and Keating 2013). For the third-generation EGFR-TKIs, the medicinal chemists and pharmacologists departed from the anilinoquinazoline core scaffold and utilized an anilinopyrimidine core instead. This approach generated compounds that showed high potency and selectivity for EGFR L858R/T790M over WT EGFR, therefore serving as *mutant-selective TKIs* targeting EGFR mutants involved in NSCLC. For the first time, therefore, third-generation EGFR-TKIs for NSCLC may now be able to tone down the rash and diarrhoea side effects as a result of the greater mutant selectivity (Zhou et al. 2009; Walter et al. 2013; Gray and Haura 2014). EGFR drug developers were very keen to call this EGFR wild-type sparing, although of course the sparing was only relatively modest. Nevertheless, this was a very important developmental approach driven both by chemistry and by disease biology. Third-generation EGFR TKIs include osimertinib (AstraZeneca), rociletinib (Clovis Oncology) and WZ4002. WZ4002 was the very first third-generation EGFR-TKI to be made, and its story is fascinating (Zhou et al. 2009). The development of WZ4002 came out of Nathanael Gray's laboratory at the Dana-Farber Cancer Institute and its discovery was highly praised in a Nature publication (Zhou et al. 2009). However, what follows next is less clear and probably an important lesson for anyone wishing to develop drugs in a commercial context whilst retaining an academic standing and access to grants. In light of the discovery of WZ4002, a start-up company called Gatekeeper Pharmaceuticals was founded to help develop it further. Clearly, WZ4002 had been discovered within the laboratories and therefore auspices, of Dana-Faber but what was not fully appreciated at that time was that the research work leading to the discovery of WZ4002 had been funded, in part or full, by Novartis. Novartis was notably informed of this development and as a result of protracted legal proceedings about rights and intellectual property relating to WZ4002, Gatekeeper Pharmaceuticals was unable to conduct any meaningful further scientific research. In terms of strategic development, this is an important point. It is critical to understand and appreciate your drug development sponsors and financiers and to be crystal clear about who owns the intellectual rights to these developments. Drug development is, by its nature complex, but it does not need to be unduly complicated. Gatekeeper Pharmaceuticals lost at least 4 years in the legal proceedings, and other companies notably AstraZeneca and Clovis Oncology, then became the main players in developing the third-generation EGFR-TKIs. The revelation that this specific T790M mutation could be the most critical mutation by far for EGFR-mutant NSCLC drug development immediately spawned a huge interest in clinically developing such third-generation EGFR-TKIs. Clearly, both clinical and commercial strategies have to be very quickly implemented by these companies to undertake the clinical development, and the large number of companies that actually did undertake this third-generation clinical development demonstrates that there are huge clinical and commercial reasons driving this. Third-generation EGFR-TKI developers include Hanmi Pharmaceuticals working in collaboration with Boehringer Ingelheim to develop olmutinib (BI 1482694/HM61713), Novartis developing nazartinib (EGF816) and ACEA Biosciences developing avitinib (AC0010).

7.5 The Efficacy of Third-Generation EGFR-TKIs

Currently, the only third-generation EGFR-TKI to be approved by the FDA is osimertinib, developed by AstraZeneca and marketed as Tagrisso. Third-generation EGFR-TKIs require to show improved efficacy in patients who have relapsed following prior treatment with first-or second-generation EGFR-TKIs. As such, this is a huge hurdle for both the drug developer and patient. It is important to remember that a patient’s physical fitness to tolerate new and increasingly toxic regimes decreases dramatically through every successive round or line of treatment. And where additional tumour biopsy specimens are required to confirm the nature of the evolved mutation status, this can be very challenging and limiting, not least for the now-desperately sick and relapsed NSCLC patient who undoubtedly would have had the biopsy procedure previously. This last point is an important strategic consideration and was not lost on the EGFR third-generation drug developers. Both AstraZeneca working on osimertinib and Clovis Oncology working on rociletinib had to ensure that the relapsed patients did indeed have the T790M mutation. Instead of the usual lung tissue biopsy for which the relapsed patient would have already had a previous experience, detection of the T790M mutation was focussed on detecting it in circulating tumour DNA (ctDNA) in blood plasma wherever possible. Fortuitously, there was good concordance between the detection of the T790M mutation in ctDNA and disease lung tissue itself, and this detection method for T790M was adopted especially for the AstraZeneca osimertinib AURA trials.

In the pivotal AURA3 trial, osimertinib had a median duration of progression-free survival of 10.1 months compared with 4.4 months on platinum therapy plus pemetrexed (Mok et al. 2017). On the basis of this improvement in PFS and indeed on the basis of the new liquid blood plasma biopsy, the FDA granted approval to osimertinib.

7.6 Resistance to Third-Generation EGFR-TKIs and Development of Fourth-Generation EGFR-TKIs

As with the first- and second-generation EGFR-TKIs, the third-generation EGFR-TKI osimertinib inevitably fell prey to resistance. Now, this might seem surprising at first as the T790M mutation was seen as the 'last hurdle' of the mutations to overcome. And this is an important point to note. The nature of cancer is that it will evolve ***new*** mechanisms or activate ultra-rare mutations that may hitherto be undetected and given the physical space freed up by killed cells, adopt a clonal cell expansion and grow into that space. This seems to be the most likely reason. DNA sequencing has revealed several new mutations within EGFR, these being C797S and L718Q. In response to this new tranche of resistance, clinical development is currently undertaken to develop the fourth-generation of EGFR-TKIs that have the ability to overcome the effects especially of C797S. At the time of writing, the development of these fourth-generation EGFR-TKIs for NSCLC is still in its infancy, but, already, a compound named EA1045 has been described that appears able to elicit some positive response against C797S (Wang et al. 2016).

7.7 Conclusion

The identification of EGFR as a drug target for NSCLC provides us with one of the most compelling and fascinating lessons in clinical drug development. In the space of 10–15 years, we have progressed from first- to fourth-generation TKIs for NSCLC, all based on a single-target EGFR. It is important that we are able to fully exploit and develop to its fullest extent even a single validated drug target. At times, a good drug target is discovered, and once a drug has been developed and commercialized, we move on to the next target. Although choosing another target in combating the disease is another strategy that is perfectly reasonable, drug developers can be in danger of not extracting all the available clinical potential inherent in a single target. This review sets out some of the features and principles that should guide us as we enter into the generational progression of drugs. The key guiding principle is that disease biology is paramount. Cancer biology will always dictate how we develop our drugs with the available chemistry and pharmacology knowledge we have. Additionally, opportunities to develop companion diagnostic tools based on disease biomarkers should always be explored. When we are able to intelligently integrate the use of computational and bioinformatics tools and databases our understanding of cancer biology with available chemistry tools, we put ourselves in a good position to develop the right drugs to tackle the disease in question.

References

Castellanos EH, Horn L (2015) Generations of epidermal growth factor receptor tyrosine kinase inhibitors: perils and progress. Curr Treat Options in Oncol 16:51

Ding PN, Lord SJ, Gebski V, Links M, Bray V, Gralla RJ, Yang JC, Lee CK (2017) Risk of treatment-related toxicities from EGFR tyrosine kinase inhibitors: a meta-analysis of clinical trials of gefitinib, erlotinib, and afatinib in advanced EGFR-mutated non-small cell lung cancer. J Thorac Oncol 12(4):633–643

Douillard J, Hirsch V, Mok TS et al (2008) Molecular and clinical subgroup analyses from a phase III trial comparing gefitinib with docetaxel in previously treated non-small cell lung cancer (INTEREST). J Clin Oncol 26(suppl):8001

Dungo RT, Keating GM (2013) Afatinib: first global approval. Drugs 73:1503–1515

Fukuoka M, Yano S, Giaccone G et al (2003) Multi-institutional randomized phase II trial of gefitinib for previously treated patients with advanced non-small-cell lung cancer (The IDEAL 1 Trial). J Clin Oncol 21:2237–2246

Fukuoka M, Wu Y-L, Thongprasert S, Sunpaweravong P, Leong S-S, Sriuranpong V et al (2011) Biomarker analyses and final overall survival results from a phase III, randomized, open-label, first-line study of gefitinib versus carboplatin/paclitaxel in clinically selected patients with advanced non–small-cell lung cancer in Asia (IPASS). J Clin Oncol 29:2866–2874

Grandis JR, Sok JC (2004) Signaling through the epidermal growth factor receptor during the development of malignancy. Pharmacol Ther 102:37–46

Gray J, Haura E (2014) Update on third-generation EGFR tyrosine kinase inhibitors. Transl Lung Cancer Res 3:360–362

Hirsch FR, Varella-Garcia M, Cappuzzo F (2009) Predictive value of EGFR and HER2 overexpression in advanced non-small-cell lung cancer. Oncogene 28:S32–S37

Inukai M, Toyooka S, Ito S, Asano H, Ichihara S, Soh J, Suehisa H, Ouchida M, Aoe K, Aoe M, Kiura K, Shimizu N, Date H (2006) Presence of epidermal growth factor receptor gene T790M mutation as a minor clone in non-small cell lung cancer. Cancer Res 66(16):7854–7858

Kazandjian D, Blumenthal GM, Yuan WS, He K, Keegan P, Pazdur R (2016) FDA approval of gefitinib for the treatment of patients with metastatic EGFR mutation–positive non–small cell lung cancer. Clin Cancer Res 22:1307–1312

Kim ES, Hirsh V, Mok T et al (2008) Gefitinib versus docetaxel in previously treated non-small-cell lung cancer (INTEREST): a randomised phase III trial. Lancet 372:1809–1818

Kobayashi S, Boggon TJ, Dayaram T, Jänne PA, Kocher O, Meyerson M, Johnson BE, Eck MJ, Tenen DG, Halmos B (2005) EGFR mutation and resistance of non-small-cell lung cancer to gefitinib. N Engl J Med 352(8):786–792

Kris MG, Natale RB, Herbst RS et al (2003) Efficacy of gefitinib, an inhibitor of the epidermal growth factor receptor tyrosine kinase, in symptomatic patients with non-small cell lung cancer: a randomized trial. JAMA 290:2149–2158

Lemmon MA, Schlessinger J (2010) Cell signalling by receptor tyrosine kinases. Cell 141:1117–1134

Lynch TJ, Bell DW, Sordella R, Gurubhagavatula S, Okimoto RA, Brannigan BW, Harris PL, Haserlat SM, Supko JG, Haluska FG et al (2004) Activating mutations in the epidermal growth factor receptor underlying responsiveness of non-small-cell lung cancer to gefitinib. N Engl J Med 350:2129–2139

Merrick D, Kittelson J, Wintherhalder R, Kotantoulos G, Ingeberg S, Keith RL et al (2006) Analysis of c-ErbB1/epidermal growth factor receptor and c-ErbB2/HER-2 expression in bronchial dysplasia: evaluation of potential targets for chemoprevention of lung cancer. Clin Cancer Res 12:2281–2288

Mok TS et al (2009) Gefitinib or carboplatin–paclitaxel in pulmonary adenocarcinoma. N Engl J Med 361:947–957

Mok TS et al (2017) Osimertinib or platinum–pemetrexed in EGFR T790M–positive lung cancer. N Engl J Med 376:629–640

Paez JG, Janne PA, Lee JC, Tracy S, Greulich H, Gabriel S, Herman P, Kaye FJ, Lindeman N, Boggon TJ et al (2004) EGFR mutations in lung cancer: correlation with clinical response to gefitinib therapy. Science 304:1497–1500

Pao W, Chmielecki J (2010) Rational, biologically based treatment of EGFR-mutant non-small-cell lung cancer. Nat Rev Cancer 10:760–774

Pao W, Miller V, Zakowski M, Doherty J, Politi K, Sarkaria I, Singh B, Heelan R, Rusch V, Fulton L et al (2004) EGF receptor gene mutations are common in lung cancers from “never smokers” and are associated with sensitivity of tumors to gefitinib and erlotinib. Proc Natl Acad Sci U S A 101:13306–13311

Pao W, Miller VA, Politi KA, Riely GJ, Somwar R, Zakowski MF, Kris MG, Varmus H (2005) Acquired resistance of lung adenocarcinomas to gefitinib or erlotinib is associated with a second mutation in the EGFR kinase domain. PLoS Med 2(3):e73

Pérez-Soler R (2004a) Phase II clinical trial data with the epidermal growth factor receptor tyrosine kinase inhibitor erlotinib (OSI-774) in non-small-cell lung cancer. Clin Lung Cancer 6(Suppl 1):S20–S23

Pérez-Soler R (2004b) The role of erlotinib (Tarceva, OSI 774) in the treatment of non-small cell lung cancer. Clin Cancer Res 10:4238s–4240s

Pérez-Soler R, Chachoua A, Hammond LA et al (2004) Determinants of tumor response and survival with erlotinib in patients with non—small-cell lung cancer. J Clin Oncol 22:3238–3247

Posner I, Engel M, Gazit A, Levitzki A (1994) Kinetics of inhibition by tyrphostins of the tyrosine kinase activity of the epidermal growth factor receptor and analysis by a new computer program. Mol Pharmacol 45:673–683

Reck M et al (2010) Erlotinib in advanced non-small cell lung cancer: efficacy and safety findings of the global phase IV Tarceva Lung Cancer Survival Treatment Study. J Thorac Oncol 5:1616–1622

Salomon DS, Brandt R, Ciardiello F, Normanno N (1995) Epidermal growth factor-related peptides and their receptors in human malignancies. Crit Rev Oncol Hematol 19:183–232

Sznol M (2010) Reporting disease control rates or clinical benefit rates in early clinical trials of anticancer agents: useful endpoint or hype? Curr Opin Investig Drugs 11:1340

Walter AO, Sjin RT, Haringsma HJ, Ohashi K, Sun J, Lee K et al (2013) Discovery of a mutant-selective covalent inhibitor of EGFR that overcomes T790M-mediated resistance in NSCLC. Cancer Discov 3:1404–1415

Wang S, Song Y, Yan F, Liu D (2016) Mechanisms of resistance to third-generation EGFR tyrosine kinase inhibitors. Front Med 10(4):383–388

Zhou W, Ercan D, Chen L, Yun CH, Li D, Capelletti M et al (2009) Novel mutant-selective EGFR kinase inhibitors against EGFR T790M. Nature 462:1070–1074

Chapter 8
Dental Stem Cells in Regenerative Medicine: Emerging Trends and Prospects in the Era of Bioinformatics

Saravanan Ramaswamy, Kavitha Odathurai Marusamy, and Gauthaman Kalamegam

Contents

S. Ramaswamy · K. O. Marusamy
Faculty of Dentistry, Ibn Sina National College for Medical Sciences,
Al Mahjar, Jeddah, Saudi Arabia

G. Kalamegam (✉)
Stem Cell Research Unit, Center of Excellence in Genomic Medicine Research,
King Abdulaziz University, Jeddah, Saudi Arabia

Faculty of Medicine, AIMST University, Semeling, Bedong, Kedah, Malaysia
e-mail: kgauthaman@kau.edu.sa

N. A. Shaik et al. (eds.), *Essentials of Bioinformatics, Volume II*,
https://doi.org/10.1007/978-3-030-18375-2_8

Abbreviations

APC	Adenomatous polyposis coli
ATP	Adenosine triphosphate
BLAST	Basic local alignment search tool
CCAP	Cancer Chromosome Aberration Project
CGAP	Cancer Genome Anatomy Project
cGMP	Current Good Manufacturing Practice
COMS	Complementary metal oxide semiconductor
CT	Computed tomography
DAVID	Database for annotation, visualization and integrated discovery
DNA	Deoxyribonucleic acid
dNTPs	Deoxyribonucleotide triphosphates
DPSCs	Dental pulp stem cells
EBI	European Bioinformatics Institute
EMBL	European molecular biology laboratory
ESCs	Embryonic stem cells
G-CSF	Granulocyte colony-stimulating factor
GEO	Gene expression omnibus
GO	Gene ontology
HCS	High-content screening
HGBASE	Human genic biallelic sequences
HGP	Human Genome Project
HTS	High-throughput screening
IKB	Immunome knowledge base
iPSCs	Induced pluripotent stem cells
KEGG	Kyoto Encyclopedia of Genes and Genomes
miRNA	MicroRNA
MRI	Magnetic resonance imaging
MSCs	Mesenchymal stem cells
MSD	Macromolecular structure database
NCBI	National Center for Biotechnology Information
NIH	National Institutes of Health
NM	Nanomaterial
OMIM	Online Mendelian inheritance in man
ORF Finder	Open reading frame finder
PCR	Polymerase chain reaction
PDLSCs	Periodontal ligament stem cells
RefSeq	Reference sequence
RNA	Ribonucleic acid
SAGE	Serial analysis of gene expression
SCAP	Stem cells from apical papilla
SGSCs	Salivary gland stem cells
SMRT	Single-molecule real time
SMS	Single-molecule sequencing

SNP	Single-nucleotide polymorphisms
SOLiD	Sequencing oligonucleotides by ligation and detection
UniProt	Universal Protein resource
UniRef	UniProt Reference
ZMW	Zero-mode waveguides

8.1 Introduction

Oral and dental health is important, and its neglect predisposes to myriad diseases that can not only affect the structures within the oral cavity but also cause systemic illness. Diet, personal habits and tobacco smoking are some of the causes that can affect tooth, soft gingival tissues and underlying deep structures including the bones. Oral cavity consists of diverse bacterial community with nearly more than 700 different strains identified by metagenomic studies (Jenkinson 2011), and generally, the oral microbiota helps to prevent colonization of the pathogenic strains (Marsh 1994). Improper hygiene and compromised health status can lead to excessive multiplication of these bacteria, which then colonize on the teeth and produce a sticky colourless substance commonly known as ‘plaque’. Plaque reacts with sugars in the food that we consume and forms acids which can destroy the outer hard covering of the tooth, namely, the ‘enamel’, and cause tooth decay (dental caries) (Loesche 1986). Apart from being associated with dental caries, the sticky biofilms (plaques) can also lead to infection and inflammation of the gingival tissues resulting in periodontitis and peri-implantitis. Persistence of infection can also be associated with developmental disorders of the tooth, its shape, number and alignment (Luder 2015).

The equilibrium that results following invasion of the cariogenic bacteria depends on many of the cellular and molecular events including the host immune response; cytokine/chemokine signalling; host–pathogen interactions leading to the release of toxic materials; damage of the soft and hard tissue; contribution by odotoblasts during initial stages; and by the pulp fibroblasts and stem cells at later stages (Cooper et al. 2017). Although tissue regeneration and functional restoration is the final process following infection/inflammation and tissue damage, vast insights of the offending pathogen and its pathological sequelae can be readily obtained using bioinformatics. This capability will pave way for detection of the early biomarkers in disease, their management and prevention. In cases of larger structural defects where the inherent in vivo repair/regeneration fails, prosthetic materials are used to aid restoration of both structure and function. Bioinformatics can help identify the right type of biomaterial by providing the surface protein signature which indirectly will influence the cellular properties.

The aim of the present chapter is to highlight (i) the various types of dental stem cells and its role in regenerative medicine and (ii) the importance and necessity for integration of bioinformatics. A brief background information regarding the development of tooth (odontogenesis), common diseases of tooth and adnexa, current management strategies and existing limitations are given in the following section, so that a new reader is exposed to the basics of dentistry. This will help to understand comprehensively that both regenerative medicine and bio-informatics are essential and are poised to change the landscape in future dentistry.

8.2 Development, Structure and Function of Human Tooth

Tooth development is a complex process and is formed from the *embryonic stem cells* (ESCs) at appropriate stages of foetal development. The primary teeth development occurs between the 6th and 8th week of prenatal development and the permanent teeth around the 20th week. Embryologically, the tooth germ cells that eventually form the tooth are developed from two different tissue sources, namely, the *ectodermal epithelium* of the first pharyngeal arch (Fig. 8.1a) and the *ectomesenchyme* of the neural crest (Fig. 8.1b). The enamel of the tooth crown is derived from the ameloblasts (*ectoderm*). The odontoblasts and cementoblasts derived from the *ectomesenchym*e form the pulp, dentin, cementum and the periodontal ligament.

The structure of the tooth can be divided into *crown, neck* and *root*, each of which contains several distinct parts (Fig. 8.2). The crown is the visible portion of the tooth and is made up of 'enamel' the outermost hard layer and the 'dentin', which is the mineralized layer beneath enamel extending from the crown to the root. The enamel provides the strength for chewing and dentin helps protect the teeth from heat and cold. The neck is the intermediate portion between the crown and the root and is formed of 'gums' the pink fleshy gingival tissue and the 'pulp cavity' containing within the 'pulp' blood vessels and nerves. The root extends from below the neck to the tooth sockets in the bone and is made up of 'root canal', the passageway filled with pulp; 'cementum' is the bone-like material that covers the root and is connected to the periodontal ligament. The 'periodontal ligament' is made up of collagen and contains the blood vessels and nerves and the jaw bones containing tooth sockets which hold the teeth in place.

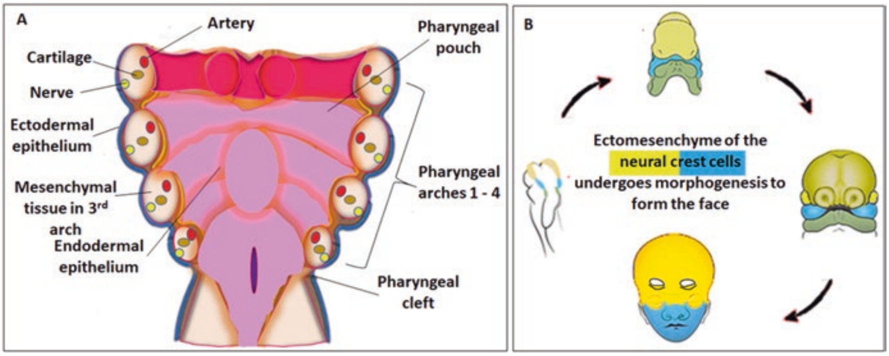


Fig. 8.1 (**a**) The Pharyngeal arches showing the mesenchymal tissue and the epithelium. The first Pharyngeal arch ectoderm invaginates to form oral cavity (pink shaded). (**b**) Neural crest cells and facial development

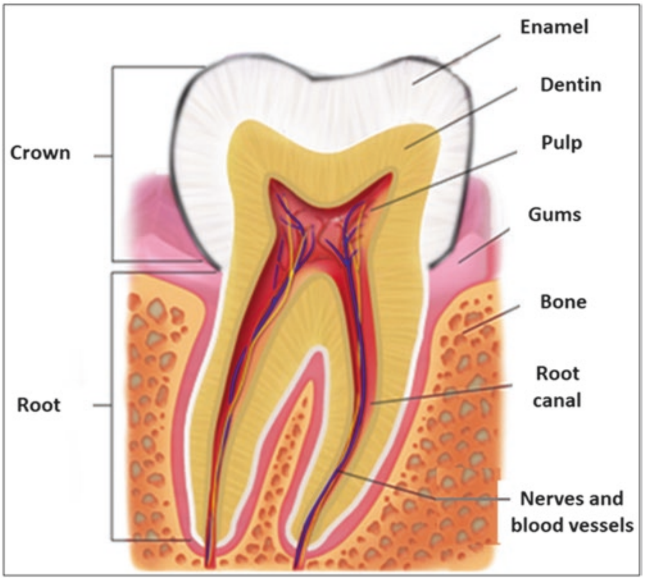


Fig. 8.2 Structure of tooth

8.3 Common Dental Diseases, Current Treatment and Limitations

The most common dental diseases are periodontal diseases and dental caries. Dental disease affects people of all age groups and all races. Patients with poor oral health are more likely to have respiratory and cardiovascular diseases, adverse pregnancy outcomes and diabetes mellitus. Dental diseases are complex diseases with multiple genetic and environmental risk factors. Predictive test for dental caries or for periodontal disease does not currently exist. No gene to date has been identified that has as large an impact on periodontal disease as do environmental influences, such as smoking or diabetes. While genetic testing holds potential for clinical application in the future, clinical measurements remain the best approach to assessment of caries and periodontal disease at present.

8.3.1 Dental Caries

Dental caries refers to enamel, dentine or cementum destruction of bacterial acid produced in dental plaque leading to a cavity in the tooth crown or root (Selwitz et al. 2007). Usually, dental caries progresses as a chronic disease (Fig. 8.3).

Numerous efforts on gene mapping have been made so far to identify specific genetic loci contributing to caries susceptibility (Werneck et al. 2011). Saliva contains components that can directly kill cariogenic bacteria. Saliva is also rich in

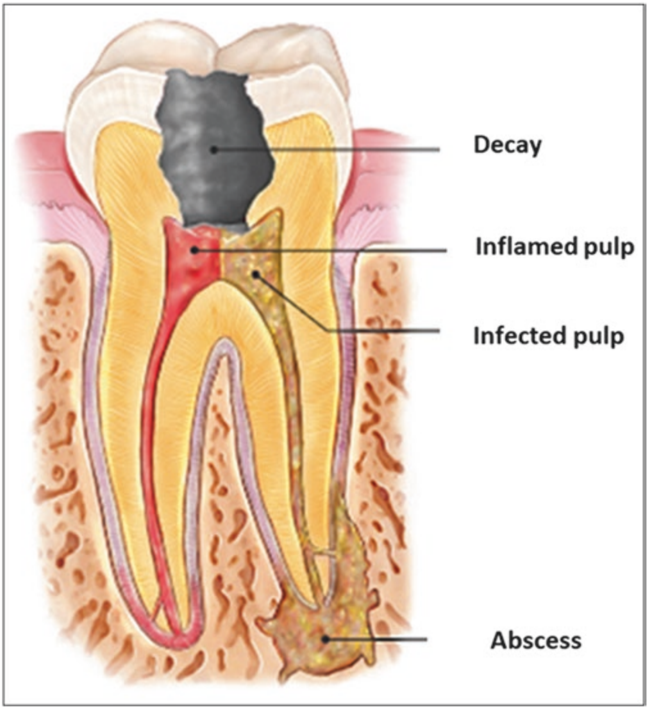


Fig. 8.3 Dental caries progression to pulp and periapical tissues

calcium and phosphates which are actively involved in the enamel remineralization process. The physical flow of saliva helps to dislodge microbial pathogens from teeth and mucosal surfaces. Saliva can also cause microbes to clump together so that they can be swallowed before they become firmly attached. So salivary composition and flow are important factors in caries susceptibility (Stookey 2008). Malposition of the teeth, deep anatomy grooves and areas of retention due to the natural morphology of the tooth structure can cause difficulties in tooth brushing and fluoride penetration and, thus, be considered as caries risk factors (Guzmán-Armstrong 2005). Dietary and taste preferences can influence the amount and type of plaque formation and debris and the presence of relative numbers of cariogenic microorganisms on tooth surfaces. The interactions of the cariogenic potential of foods (e.g., sucrose), the frequency of eating and the physical state (or type) of the diet all can affect individually or jointly the carious process (Wendell et al. 2010).

Future management of dental caries requires early detection and risk assessment. The effects of prevention on caries prevalence and the advantages of improved dental materials have shifted the focus in caries management from restoring tooth structure to development and use of dental materials to prevent disease, remineralization procedures, minimally invasive treatments and materials with which early lesions can be impregnated to prevent further progression.

8.3.2 Periodontal Disease

Periodontal disease typically affects structures which support the teeth. It ranges from a mild gingivitis to a more severe pattern of bone loss. Periodontitis is a chronic multifactorial inflammatory disease, and both environmental and genetic factors play a major role in the progression of the disease with consequent tissue destruction around the dental roots and alveolar bone (Fig. 8.4). The risk of progression of periodontitis is directly associated with the biofilm found in the gingival sulcus, in which both amount and presence of specific species of bacteria represent risk factors. Recently, research has been focussed on the identification of molecular markers such as cytokines, chemokines, membrane surface receptors and antigen recognition proteins capable of determining the risk of disease development (Carinci et al. 2015).

The backbone of periodontal treatment consists of mechanical removal of bacterial deposits and calculus from the subgingival environment either by hand instruments or by ultrasonic devices, performed either surgically or non-surgically, along with a strict regimen of plaque control. In the future, the emerging field of genomics will be identifying individual risk factors, and controlling them will become central to periodontal practice.

8.3.3 Oral Cancer

Oral cancer is the sixth most common malignancy in the world. More than 90% of oral cancers (occurring in the mouth, lip and tongue) are oral squamous cell carcinoma. The incidence rate of oral cancer varies widely throughout the world, with an evident prevalence in South Asian countries. This high incidence occurs in

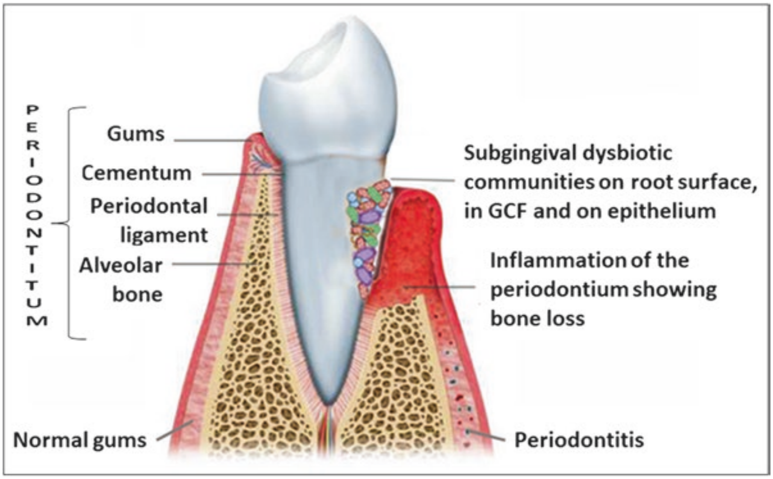

PERIODONTITUM
Gums
Cementum
Periodontal ligament
Alveolar bone
Normal gums
Subgingival dysbiotic communities on root surface, in GCF and on epithelium
Inflammation of the periodontium showing bone loss
Periodontitis

Fig. 8.4 Periodontal disease associated with bone loss. GCF Gingival crevicular fluid

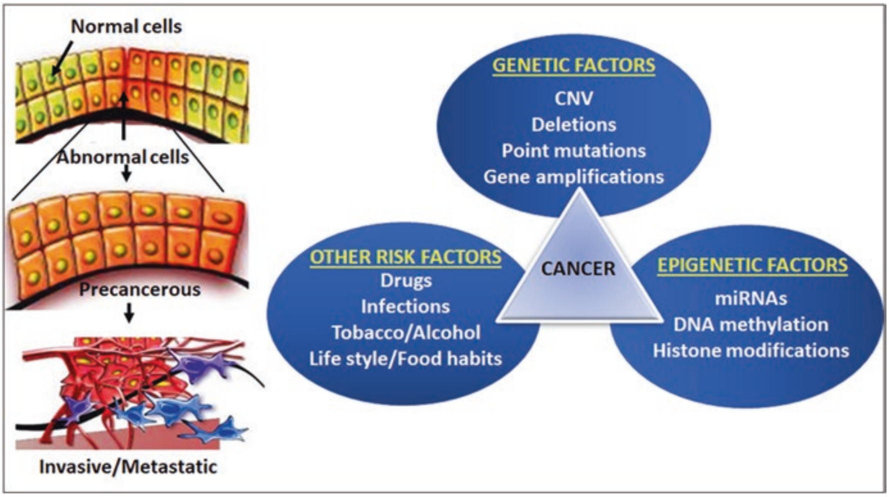


Fig. 8.5 The risk factors in oral carcinoma progression

correlation with oral cancer-associated behaviours such as alcohol and tobacco use (Fig. 8.5). These behaviours lead to genetic variations in tumour suppressor genes (APC, p53), proto-oncogenes (Myc), oncogene (Ras) and genes controlling normal cellular processes (EIF3E, GSTM1). Processes such as segregation of chromosomes, genomic copy number, loss of heterozygosity, telomere stabilities, regulations of cell cycle checkpoints, DNA damage repairs and defects in notch signalling pathways are involved in causing oral cancer (Ali et al. 2017).

The prime objective of oral cancer management is to prevent mortality and to improve the quality of life of the patient. The choice of treatment depends on the site and size of the primary lesion, cell type and degree of differentiation, presence or absence of lymph node metastases and assessment of potential complications of each therapy. Surgery is most commonly accepted in the treatment of oral cancer, followed by radiotherapy. Chemotherapy is an adjunct to the principal curative modalities of surgery and radiation. Understanding the cancer genetics may also permit the development of new cancer therapies.

Given the limitations as with some of the existing management of dental diseases, the use of stem cell-based therapies has largely evolved as an attractive and alternative choice. As such, it will be essential to have some basic understanding about the stem cells and their types as well as their potential use in regenerative medicine.

8.4 Stem Cells and Regenerative Medicine

Regenerative medicine is a branch of medicine that integrates two major disciplines, namely, cell biology and materials engineering, to aid regeneration of functional tissues. It essentially contributes to the repair or replacement of damaged tissues and organs, when the body's natural defence mechanisms for repair and homeostasis become limited or impossible. The field of tissue engineering and regenerative medicine has witnessed tremendous growth in the last two decades mainly due to improved methods in isolation and culture expansion of various stem cells including the oro-dental stem cells.

8.4.1 Classification of Stem Cells

Stem cells are unspecialized cells that have prolonged self-renewal potential and can differentiate into many different cell lineages. Depending on their source from which the stem cells are derived, they can be broadly classified into (i) embryonic stem cells, (ii) adult stem cells and (iii) foetal stem cells (Fig. 8.6). Embryonic stem cells (ESCs) are derived from the inner cell mass of the 4- to 5-day-old blastocyst-stage embryos and are the most versatile stem cell type. They have indefinite self-renewal capacity and the potential to differentiate into almost all the tissue types representing the three germ layers, namely, ectoderm, mesoderm and endoderm (Bongso et al. 1994; Thomson et al. 1998). These cells are therefore commonly referred to as pluripotent stem cells. Adult stem cells are those which are isolated from within the special zones, viz. the 'stem cell niche' of various adult tissues such as the bone marrow (Friedenstein et al. 1966), bone (Owen 1985), limbal region of the cornea (Tseng 1989), epidermis of the skin (Toma et al. 2001), adipose tissue (Zuk et al. 2001), liver (Dabeva and Shafritz 2003), surface of the articular cartilage

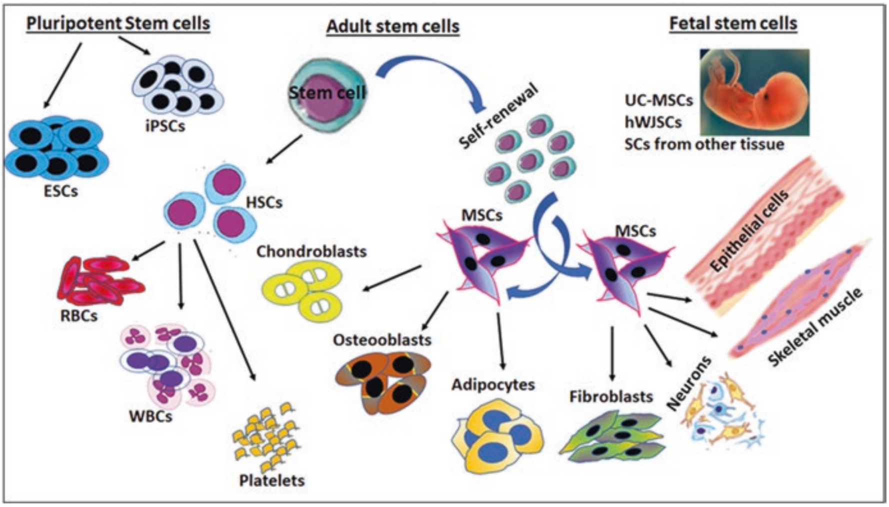


Fig. 8.6 Classification of stem cells and their differentiation potential. ESCs Embryonic stem cells; iPSCs induced pluripotent stem cells; MSCs Mesenchymal stem cells; HSCs Haematopoietic stem cells; RBCs Red blood cells; WBCs White blood cells; UC-MSCs Umbilical cord-mesenchymal stem cells; hWJSCs human Wharton's jelly stem cells

(Dowthwaite et al. 2004), intestine (De Coppi et al. 2006), pancreatic islets (Gallo et al. 2007), endometrium (Gargett et al. 2009), brain (Kang et al. 2010) and heart muscle (Chimenti et al. 2012), and are commonly referred to as postnatal mesenchymal stem cells (MSCs). Foetal stem cells are those derived from the birth-related tissues (umbilical cord and cord blood) and abortuses (Marcus and Woodbury 2008). The stem cells that reside within the special niches in various tissues either contribute to the normal turnover of cells (as in intestinal or skin epithelium) or become activated to differentiate into a specific cell type in response to tissue injury/damage to maintain homeostasis. In addition to these naturally occurring stem cells, currently there are methods to derive pluripotent stem cells from a differentiated cell type using forced expression of pluripotent genes, and these cells are known as induced pluripotent stem cells (iPSCs) (Takahashi and Yamanaka 2006; Yu et al. 2007). Stem cells, therefore, can also be classified according to their differentiation potential into (i) pluripotent stem cells (ESCs, iPSCs) and multipotent stem cells (Adult and foetal MSCs) (Fig. 8.6).

8.4.2 Oro-dental Stem Cells

Literature evidences indicate the presence of MSCs from within the various tissues of the oral cavity. These MSCs are broadly classified into (i) dental and (ii) non-dental MSCs (Fig. 8.7). The dental MSCs include those from the dental pulp (Gronthos et al. 2000), apical papilla (Huang et al. 2008) and the exfoliated deciduous teeth (Miura et al. 2003). The non-dental MSCs include those from the periodontal (Seo et al. 2004), gingival (Zhang et al. 2009), dental follicle (Morsczeck et al. 2005), oral mucosa (Marynka-Kalmani et al. 2010), periosteum (Arnsdorf et al. 2009), oro-facial bone marrow (Akintoye et al. 2006) and the salivary glands (Sato et al. 2007). Additionally, MSCs have also been isolated from the damaged oral tissues such as the inflamed pulp (Alongi et al. 2010) and apical cysts (Marrelli et al. 2013).

Similar to MSCs from other sources, the oro-dental MSCs are also reported to exhibit the stipulated minimal criteria for MSCs by the International Society of Cellular Therapy (Dominici et al. 2006). Accordingly the oro-dental stem cells have the properties of (i) adherence to tissue culture plastic; (ii) differentiation into multiple cell lineages including osteoblasts, chondroblasts and adipocytes; and (iii) positive expression of MSC-related CD makers, namely, CD105, CD73 and CD90, and lack of expression of CD14, CD79A, CD45, CD34 and HLA-DR surface molecules (Dominici et al. 2006) (Fig. 8.8).

8.4.3 Regenerative Medicine Applications of Oro-dental Stem Cells

This section will briefly highlight some of the tissue engineering and regenerative medicine applications in relation to oral-dental disorders.

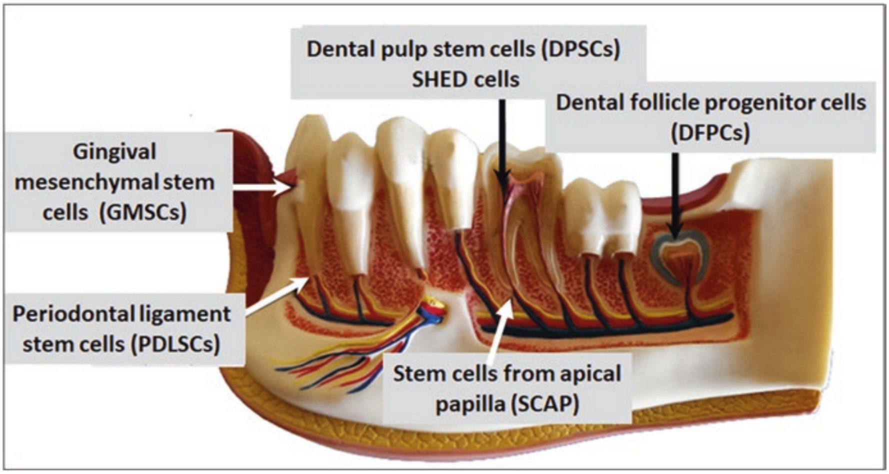

Dental pulp stem cells (DPSCs)
SHED cells
Dental follicle progenitor cells (DFPCs)
Gingival mesenchymal stem cells (GMSCs)
Periodontal ligament stem cells (PDLSCs)
Stem cells from apical papilla (SCAP)

Fig. 8.7 Dental stem cells. SHED cells Stem cells from human exfoliated deciduous teeth

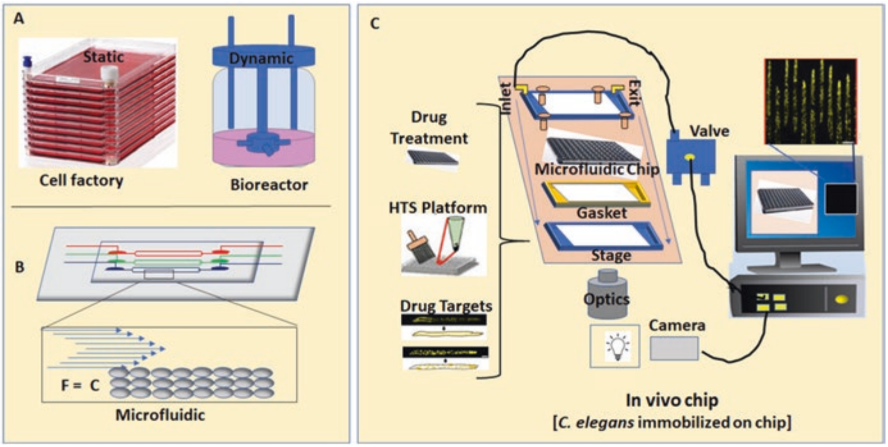


Fig. 8.8 High-throughput systems (HTS). (**a**) Cell factory and Bioreactor that helps to scale up large numbers of cells in 2D (adherent cells) with precise spatiotemporal dynamics and 3D (suspension cells) platforms respectively; (**b**) Microfluidics platform; (**c**) *In vivo* chip

(a) *Dental pulp stem cells (DPSCs)*: Dental caries is a common disorder, and when it becomes deep, pulpectomy is the choice of treatment followed by root canal filling. However, associated complications such as apical periodontal lesions due to microleakage from the tooth crown and vertical fractures eventually results in higher incidences of tooth extraction. A recent pilot clinical study demonstrated that transplantation of autologous dental pulp stem cells (DPSCs) led to complete recovery of the dental pulp after 24 weeks which was similar to that of the untreated normal controls (Nakashima et al. 2017). The authors derived the DPSCs from discarded tooth following stimulation with granulocyte colony-stimulating factor (G-CSF) and expanded them under current good manufacturing practice (cGMP) so as to obtain clinical-grade MSCs. These mobilized DPSCs (1 × 10^6 cells) were then seeded onto a clinical-grade atelocollagen scaffold together with G-CSF and transplanted into pulpectomized teeth in patients with irreversible pulpitis. The transplanted cells were held in place with gently covered gelatin sponge and the cavity was sealed using glass ionomer cement and resin with a bonding agent. The electric pulp test of the pulp was positive at 4 weeks, and both magnetic resonance imaging (MRI) and cone beam computed tomography (CT) at 24 weeks following DPSC transplantation demonstrated function dentin formation (Nakashima et al. 2017).

(b) *Salivary gland stem cells (SGSCs)*: Irreversible salivary gland (SG) damage can occur following disease states such as Sjogren's syndrome, thyroid disorders and metabolic syndromes and after radiotherapy for head and neck cancers (von Bültzingslöwen et al. 2007). Autologous SG stem cell progenitors (SGSCs) isolated from salispheres (in vitro floating spheroidal cultures of cells from SG) have been used to restore glandular function following irradiation or damage (Lombaert et al. 2008; Pringle et al. 2016). Due to the existing limitations in expansion of the SGSC progenitors, MSCs from other sources have been equivocally used (Lim et al. 2013; Ono et al. 2015; Tran et al. 2013). However, it is difficult to achieve the orderly arrangement of cells with correct polarity to enable directional flow of secretions in the duct. Whole salivary gland regeneration encompassing all its cellular components such as acinar, ductal, myoepithelial, endothelial and neuronal cells is therefore essential for efficient functional restoration. Tissue-engineered three-dimensional (3D) scaffolds will be capable of providing the needed tissue architecture, and furthermore, the use of 3D organ bioprinting systems can help achieve functional organ reconstruction similar to that of the normal (Ferreira et al. 2016; Lombaert et al. 2017).

(c) *Periodontal ligament stem cells (PDLSCs)*: Periodontitis is an inflammatory disease that can cause damage both to the tooth and its adnexal tissues, namely, the cementum, periodontal ligament and the alveolar bone (Lu et al. 2013). This could result in tooth loss, and several restorative measures have been attempted to treat periodontitis-associated tissue damage including guided tissue regeneration procedures with use of bone grafts combined with bioactive agents and growth factors (Chen and Jin 2010; Lu et al. 2013); however, these strategies are limited in advanced periodontal defects. The periodontal ligament stem cell progenitors (PDLSCs) have been identified to be committed to some of the developmental cell lineages such as osteoblasts, cementoblasts and fibroblasts and have been used effectively in periodontal tissue regeneration (Catón et al. 2011; Yang et al. 2009). A recent single-centre randomized clinical trial evaluated the feasibility of using PDLSCs derived from impacted third molars (following their removal) under cGMP guidelines to regenerate periodontal intra-bony defects (Chen et al. 2016). The PDLSCs were used together with commercial osteoconductive material (Bio-Oss® to aid tissue regeneration and it was demonstrated that that the alveolar bone height increased with time (3–12 months), and the cell transplantation procedures were clinically safe. However, there were not much differences in the clinical periodontal parameters such as clinical attachment level, probing depth and gingival recession (Chen et al. 2016). Further research in this area using suitable scaffolds, optimization of the cell numbers and multicentre trials are awaited.

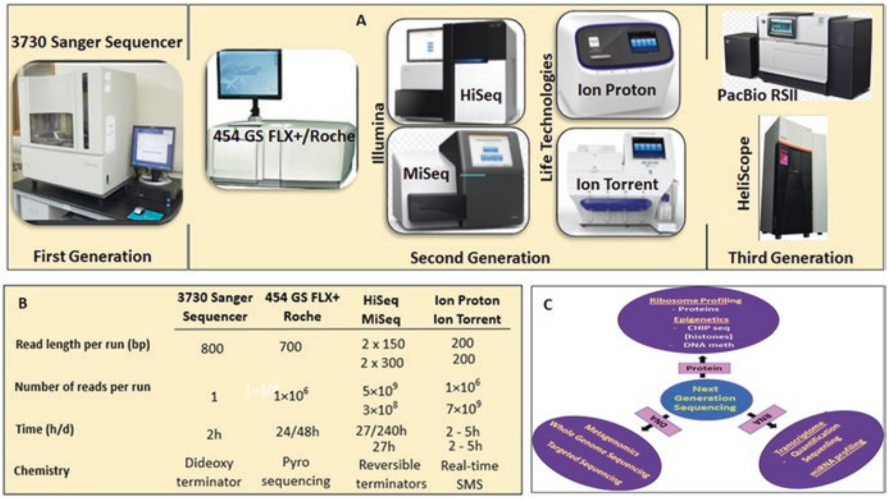
(d) *Stem cells from apical papilla* (SCAP): The sequelae of inflammatory cell invasion and fibrous tissue formation in the peri-apical area following endodontic infections lead to the formation of peri-apical cysts (Nair 2004). A series of interesting clinical observations of periosteal bone formation following removal of the apical cysts (Maeda et al. 2004) led to the hypothesis supporting the presence of stem cell progenitors (Patel et al. 2010) and the isolation and characterization of the human periapical stem cells (PASCs) (Marrelli et al. 2013). Similar to MSCs from most tissue sources, the PASCs demonstrated efficient multilineage differentiation potential including bone and neuronal cell types (Marrelli et al. 2013, 2015). The PASCs are reported to have high proliferative ability and wide differentiation potential, thus making these cells an attractive choice for bone and dental tissue regeneration either alone or in combination with biological scaffolds and growth factors (Tatullo et al. 2017).

8.5 High-Throughput and High-Content Screening

High-throughput screening (HTS) is defined as the use of automated tools to facilitate rapid execution of a large number and a variety of biological assays that may include several substances in each assay (Nel et al. 2012). HTS typically is used to analyse fewer endpoints but repetitively for numerous samples. The advantage is that more information of the endpoint is usually known, and therefore, not much informatics may be necessary and helps with rapid decision-making (Pamies et al. 2018). High-content screening (HCS) helps screening of hundreds to hundreds of thousands of endpoints, capturing large biological information of the model analysed. The advantages of HCS are its holistic and non-targeted nature; however, the generation of huge data needs expertise and time for data analysis (Pamies et al. 2018). The ‘omics’ approaches, namely, transcriptomics, epigenomics, lipidomics, proteomics and metabolomics, as well as imaging technologies will largely come under the purview of HCS (van Vliet 2011; Van Vliet et al. 2014). It is not uncommon to find both HTS and HCS being interchangeably used, but understanding of this difference is necessary, and we shall briefly see their respective applications (Fig. 8.9).

8.5.1 HTS in Drug Screening, Cell Culture and Imaging

In pharmaceutical companies, HTS has been used to facilitate rapid evaluation of potential drugs as early as the 1980s and continues to remain as a cornerstone for small-molecule drug discovery. Using HTS, libraries of compounds can be analysed for their biological activity using robotics (automation); carry out robust biological



B	3730 Sanger Sequencer	454 GS FLX+ Roche	HiSeq MiSeq	Ion Proton Ion Torrent
Read length per run (bp)	800	700	2 x 150 2 x 300	200 200
Number of reads per run	1	1×10^6	5×10^9 3×10^8	1×10^6 7×10^9
Time (h/d)	2h	24/48h	27/240h 27h	2 - 5h 2 - 5h
Chemistry	Dideoxy terminator	Pyro sequencing	Reversible terminators	Real-time SMS

Fig. 8.9 High-content systems (HCS). (**a**) DNA sequencers (Frist, second and third generation); (**b**) Features of DNA sequencers; (**c**) Applications of next-generation sequencing (NGS)

assays, thereby minimizing false positives and increasing the sensitivity of the assay system; as well as interpret results using high-content analytical tools (Hook et al. 2010). Eventually, all the above leads to rapid identification of the lead candidates even in the absence of structure-based design.

Conventionally, the cells used as culture models to screen for drug discovery or development are derived from primary tissues or immortalized cell lines. These cells, however, are not the right choice for use as screening models due to the following reasons: (i) the cells derived from primary tissue have short survival *in vitro*, (ii) presence of aneuploid karyotype as in cells derived from cancer tissue or (iii) having phenotype unrelated to the tissue of interest. The above shortcomings can be readily overcome with use of embryonic stem cells (ESCs), as these cells are genetically stable, can be maintained for sufficient durations in culture without undergoing differentiation and can exert uniform physiological responses (Cho et al. 2013). Modified culture protocol such as use of feeder-free platform enabled culture and maintenance of mouse ESCs in their undifferentiated state for up to 7–8 days in a 96-well plate format is by itself an advancement towards HTS, given the highly sensitive nature of these cells and their stringent culture conditions (Cho et al. 2013). This culture system can be scaled up/automated and used for toxicological screening of known or unknown compounds/drugs to analyse defined endpoints or signalling pathways (Fig. 8.8a). Unlike the in vivo state, stem cell fate can become easily altered upon culture on in vitro mechanical platforms due to imprecise temporal and spatial control of the microenvironments.

Progress in micro-/nano-fabrication and microfluidics has enabled development of culture systems that closely mimic the in vivo conditions. Microplatforms developed using biomicroelectromechanical systems technology have found various cell-/stem cell-based biomedical research applications such as (i) delivering biomolecules to cells in a controllable way, (ii) cell migration or morphogenesis based on gradient-dependent morphogens, (iii) embryoid body generation for three-dimensional studies and (iv) microarray/microwell culture of cell/stem cell to study fate of single stem cells or cell–cell interactions (Park et al. 2015). Similarly, use of microfluidic chip-based (Fig. 8.8b) single-cell isolation and culture has paved way for high-throughput clonogenic assay of heterogenetically different single cells (Lin et al. 2016). An automated microbioreactor system was successful in rapid and consistent large-scale production of antibodies (Velugula-Yellela et al. 2018). All the above studies highlight how automation can help with miniaturization of models with maximization of product.

Success from miniaturization and HTS research in vitro also paved the way for in vivo research applications. Sumiyama et al. (2018) developed a microdevice to generate chimeric blastocysts by aggregation of eight-cell-stage embryo (blastomeres) and mutant mESCs, which upon transfer to pseudopregnant mice resulted in the birth of chimeric pups as indicated by their coat colour (Sumiyama et al. 2018). The microdevice was fabricated using a polystyrene material and consisted of a funnel-like structure to help aggregation of mutant mESCs and blastomeres at the bottom (300 uM in diameter). Although one can argue that direct gene-editing techniques of fertilized mouse embryos are quite efficient for induction of small mutations, the advantages with the generation of individual mice knock-ins/knockouts of a relatively large size cannot be refuted. In addition, the reported cell aggregation technique has its own advantages such as (i) the cells that are held in position by surface tension can be easily transferred to the culture wells with gentle pressure on the liquid from the top and (ii) the expertise essential as with microinjection technique is not required. Moreover, scalability can be achieved as the microdevice can be used in conjunction with a regular 96-well culture plate and hence compatible with use of multichannel pipettes or programmable machines (Sumiyama et al. 2018).

Cellular models such as genetically modified cell lines, spheres, organoids and small/large whole animal models are indispensable for drug screening. Of these, uses of whole animal model as a screening platform would be greatly advantageous as the complete pharmacokinetics of a candidate drug can be analysed and rapidly advance to human clinical trials. Use of large animal models is expensive and time consuming, and therefore, smaller animal models with conserved gene homology to humans serve as an alternative. *Caenorhabditis elegans* has been largely used in drug screening studies. A microfluidics immobilization platform consisting of an exterior surface of a standard 96-well plate format and an interior surface made of microfabricated channels (40 parallel chambers) fitted within a gasket device having a single-input and single-output interfaces for easy flow control was developed to achieve HTS (Ben-Yakar 2019). This microfluidics platform is capable of imaging ~4000 animals in total in less than 3 minutes with an automated image acquisition software by screening pre-determined locations (Fig. 8.8c). Using this HTS platform, ~1000 FDA-approved compounds were recently screened leading to four hits for subsequent validation (Mondal et al. 2016).

Although the above immobilization microfluidics platform (*VivoChip*) is an HTS system, it is limited in that it can test only 30–40 animals per population and therefore may be best suited for secondary screening. This limitation with *VivoChip* is now overcome by using fluorescence imaging with line excitation array detection microscopy technique, which can allow imaging of moving *C. elegans* at a speed of >1 m/s without any blur and therefore can be useful to screen much bigger compound libraries (Martin et al. 2018). Therefore, it is irrefutable that HTS is robust, has several advantages over the conventional methods and is utilized in almost all domains of science and technology.

8.5.2 HCS in Genomics, Proteomics and Metabolomics

Technologies involving miniaturization, automation and analysis have developed rapidly in the last decade and continue to do so gaining far-reaching ramifications in different disciplines. To better understand the role of HCS in biomedical application, let us go through the evolution of high-throughput sequencing platforms from one of the major players in the field of deoxyribonucleic acid (DNA) sequencing. Sequencing fundamentally refers to decoding the nucleic acid sequence or the order of the nucleotides in DNA. Before moving on with high-throughput DNA sequencing platforms, a brief recapitulation of the fundamentals of molecular biology will be pertinent.

Genomics basically refers to genomes and their expression in an organism; the organization of the genes within the genome; and their evolution, conservation and variations or mutational changes. Within the nucleus of each human cell, there are 23 pairs of chromosomes (diploid), except the gametes which are haploid. Continuous stretches of DNA are tightly coiled around the histone proteins and are packaged into a chromosome. Within these vast stretches of the DNA are the genes which are capable of expression and translation into proteins. Genes provide necessary genetic information to the ribonucleic acid (RNA) by a process known as transcription to enable synthesis of either structural or functional proteins which are the building blocks of various tissues, thereby contributing to the development and function of an organism.

Elucidation of the structural details (order of nucleotides) of the DNA stretches will provide far greater information with wide ranging applications in medicine, forensics and agriculture. The quest to completely sequence the human genome led to the initiation of the Human Genome Project (HGP) by the Department of Energy and the National Institutes of Health (United States) in 1990. Scientists from around the world joined this historic project, and the rough draft of the human genome was completed in June 2000. This was further refined and declared completed in 2003, coinciding with the 50th anniversary of the publication reporting the double helical structure of DNA by Francis Crick and James D Watson (Green et al. 2015).

Until the completion of the Human Genome Project, very scant information of the human gene sequences was available. The conventional Sanger sequencing method used in the HGP is not high-throughput and hence is not economical for whole-genome sequencing. It was estimated that the HGP would cost around 2.5–3 billion USD. Genetic sequences are the blueprints of an individual makeup, and it is of great importance to know the DNA sequence of a gene as it has profound biomedical applications in medicine. Advancements in science and technology combined with HTS/HCS can reduce the prohibitive cost of traditional whole genome sequencing, thereby facilitating personalized medicine.

The sequencing platforms can be divided into basic, advanced and next-generation sequencing (Fig. 8.9a). The Sanger method of sequencing was developed by Fredrik Sanger in the year 1977 and involves chain termination method using dideoxynucleotides and DNA polymerase. Sanger sequencing was used in the HGP to determine the sequences of many small fragments (usually less than 900 bp) of human DNA. The fragments were aligned based on overlapping segments to determine the sequences of larger regions of DNA. Although Sanger sequencing method provides high-quality sequence, it is expensive and inefficient for large-scale projects. The other type of basic sequencing method is the chemical termination method developed by Maxam-Gilbert, where, instead of DNA polymerase to generate fragments, radiolabelled DNA is treated with chemicals that break the chain at specific bases into fragments. These first-generation DNA sequencing machines produce reads slightly less than one kilobase (kb) in length. Subsequent newer dideoxy sequencers – such as the ABI PRISM produced by Applied Biosystems – allowed simultaneous sequencing of hundreds of samples, and thus, Sanger sequencing came to be used in the HGP, going on to produce the first draft of human genome much earlier than the anticipated time frame (Lander 2011; Smith et al. 1986).

The second-generation sequencing includes the pyrosequencing method where luminescence was used to measure pyrophosphate synthesis, where ATP sulphurylase is used to convert pyrophosphate into ATP, which acts as the substrate for luciferase to produce light proportional to the amount of pyrophosphate (Nyren and Lundin 1985). Pyrosequencing was later licenced to 454 Life Sciences and subsequently purchased by Roche (Fig. 8.9a), which made a paradigm shift with introduction of techniques supporting massive parallelization (Margulies et al. 2005). This was soon followed by Solexa method of sequencing, which was later acquired by Illumina, wherein modifications led to the use of adapter-bracketed DNA molecules to be passed over a lawn of complementary oligonucleotides bound to a flowcell, instead of parallelizing by performing bead-based emPCR (Bentley et al. 2008). Solid-phase PCR subsequently produces neighbouring clusters of clonal populations from each of the individual original flow-cell binding DNA strands (Fig. 8.9a). This was soon followed with sequencing by oligonucleotide ligation and detection (SOLiD) system from Applied Biosystem, which uses DNA ligase for ligation and not sequencing by synthesis using DNA polymerase (Shendure et al. 2005). The other is the Ion Torrent (Life Technologies) (Fig. 8.9a) where the difference in pH caused by the release of protons (H + ions) during polymerization is used for measurement which was made possible by the use of complementary metal oxide semiconductor (CMOS) technology (Rothberg et al. 2011).

The third-generation sequencing technologies are those capable of single-molecule sequencing (SMS). DNA templates attached to a planar surface were used for sequencing together, where fluorescent reversible terminator dNTPs were washed over one base at a time and imaged (Braslavsky et al. 2003). Although slow and expensive, the use of non-amplified templates helped to avoid associated biases and errors. The most widely used third-generation technology is probably the single-molecule real-time (SMRT) platform (Fig. 8.9a) from Pacific Biosciences (van Dijk et al. 2014), where DNA polymerization occurs in arrays of microfabricated nanostructures called zero-mode waveguides (ZMWs) (Levene et al. 2003). PacBio machines are also capable of producing incredibly long reads, up to and exceeding 10 kb in length, which are useful for de novo genome assemblies (van Dijk et al. 2014). The characteristics of first, second and third generation sequencers such as (i) their read length per run, (ii) number of reads per run, (iii) the time taken, (iv) their underlying principle and the various applications of NGS are given in Fig. 8.9b, c. The field of sequencing is undergoing great revolution and with more recent platforms will certainly yield vast information in the field of molecular biology, which will in turn impact clinical medicine.

8.6 HCS in Dentistry Applications

8.6.1 Oro-dental Disorders

Oro-dental disorders can have genetic, environmental or multifactorial aetiology. Interestingly, of the known genetic syndromes (>5000), nearly 900 are associated with craniofacial/oro-dental disorders, indicating the role of genetics in oro-dental diseases (Crawford et al. 2007). There also exists a wide range of heterogeneity in isolated dental diseases making diagnosis of the genetic basis more challenging. Targeted next-generation sequencing (NGS) is beneficial in the molecular diagnosis of genetically heterogeneous disorders. Moreover, substantial reduction in the cost allows the flexibility to perform whole-exome sequencing (WES) or whole-genome sequencing (WGS). A NGS gene panel targeting 585 known and candidate genes in oro-dental disorders used in screening a cohort of 101 unrelated patients led to the identification of 21 novel pathogenic variants and causative mutations in 39 patients with an overall diagnostic rate of 39% (Prasad et al. 2016). Furthermore, among 50 unrelated patients with amelogenesis imperfecta (AI) and 21 patients with syndromic selective tooth agenesis (STHAG), a definitive diagnosis was established in 14 (27%) and 15 (71%) cases, respectively (Prasad et al. 2016).

Periodontitis is a chronic inflammatory disease of the periodontium characterized by extensive destruction of the tooth and its adnexa. The complex interaction between the microbial biofilms and the host immune response is understood underlying reason for bone and connective tissue disorders. Transcriptome analysis by microarrays is a valuable tool to study changes in gene expression and is a useful technique to study changes in gene expression patterns from tissue samples in patients with periodontitis (Beikler et al. 2008). RNA sequencing (RNA-Seq) has several advantages over microarray hybridization technique such as (i) unbiased approach due to direct sequencing and (ii) highly accurate in detecting gene expression with dynamic detection range. RNA-Seq of pooled gingival tissue samples of periodontitis patients and healthy controls identified 400 upregulated genes and 62 downregulated genes by differential expression analysis, in periodontitis tissues that mainly interact in the immune-related signalling molecules and pathways (Kim et al. 2016). Differential alternative splicing analysis revealed unique transcription variants in periodontitis tissues, thus highlighting the usefulness of RNA-Seq and the high-content screening for differential gene expression and alternative splicing in elucidating the mechanisms of pathogenesis in periodontitis (Kim et al. 2016).

RNA-Seq and microarray have also helped largely in understanding the molecular signature in molar morphogenesis. In a recent study, differential transcript expression and functional network during morphogenesis of additional molars at three key developmental stages were profiled in miniature pigs using the RNA isolated from additional molar germs. Coding and non-coding transcripts were identified using Coding–Non-Coding Index (CNCI) and annotated transcripts through mapping to the porcine, Wuzhishan miniature pig, mice, cow and human genomes. Many new unannotated genes plus 450 putative long intergenic non-coding RNAs (lincRNAs) were identified (Wang et al. 2017). Regulatory network analyses revealed that WNT and TGF-β pathways play a determining role in regulating sequential morphogenesis of additional molars (Wang et al. 2017).

8.6.2 Biofilms

Oral microbiome, which is referred to as the oral microflora or oral microbiota, is defined as all the microorganisms residing in the human oral cavity and their collective genome. Oral microbiome harbours on teeth, gingival sulcus, tongue, cheeks, hard and soft palates, and tonsils, and it is a critical component of oral health and disease (Fig. 8.10).

These biofilm communities are not only heterogeneous with respect to the species they contain but also can be architecturally diverse; for instance, they can range from a few cells thick to visually conspicuous biofilms (Bernimoulin 2003; Marsh 2006). The more diverse the community and the greater the biofilm biomass, the more likely it is that pathogenic species such as *Porphyromonas gingivalis* and *Treponema denticola* will integrate and promote periodontal disease (Kolenbrander 2000).

The ability of bacteria to aggregate via autoaggregative interactions (self-aggregation) and coaggregation (the specific recognition and adhesion of different species of bacteria to one another) is proposed to be integral to biofilm development (Short et al. 1982; Kolenbrander 2000).

Biofilm bacteria are up to 1000-fold less susceptible to antimicrobials than planktonic cells (Gilbert et al. 2002; Mah and O'Toole 2001; Roberts and Mullany 2010). The reasons behind this reduced susceptibility are multifactorial and include

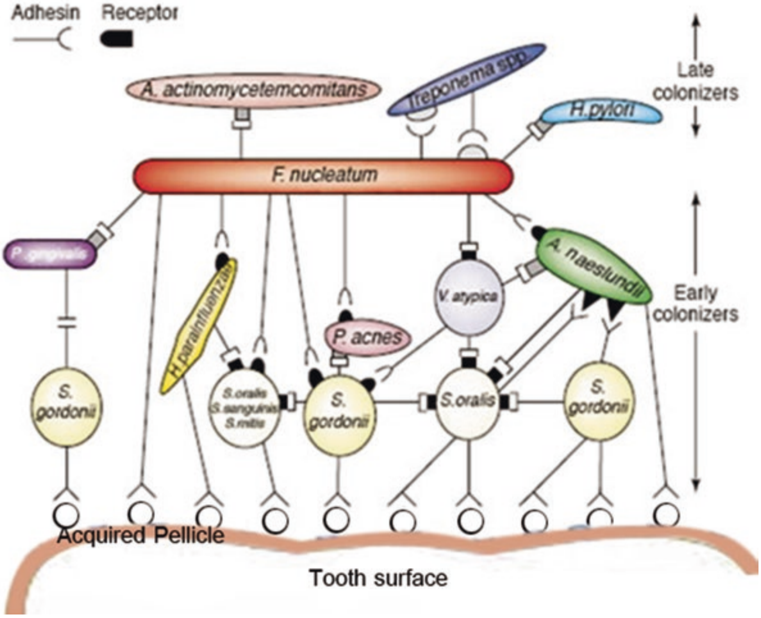


Fig. 8.10 Dental biofilm colonisers

retarded antimicrobial penetration of biofilm due to reaction diffusion limitation (Stewart 1996), altered growth rates, intraspecies and interspecies metabolite and/or cell–cell signalling interactions resulting in altered biofilm-specific phenotypes (Peters et al. 2012) and cross-species protection afforded by removal or inactivation of a given antimicrobial by a biofilm species (Gilbert et al. 2002).

Approaches to controlling the species composition and overall density of dental plaque biofilm communities encompass abrasive regimens (e.g. tooth-brushing and flossing) and chemical treatments (e.g. mouthwash).

Application of high-throughput sequencing greatly helps in understanding human oral microbiome. Numerous model biofilm systems exist as a representative of the oral cavity to examine biofilm development and/or the impact of antimicrobial compound conditions. The development of such representative *in vitro* model biofilms is important to accurately predict the *in vivo* efficacy of current or newer antimicrobials that may be used in oral hygiene products. These can be large-scale systems suitable for long-term studies, such as newly modified Robbins devices, Sorbarod-based biofilm systems and constant-depth film fermenters or simple devices such as flow cells. A critical drawback to the operation of such model systems is their physical footprint (resulting in limited capabilities for performing parallel replicate studies) and the often limiting requirement for large amounts of media in which to develop biofilms. This latter point is of great importance if the medium is expensive or time-consuming to obtain, especially if it is from natural sources (e.g. saliva or wound exudate). For example, when conducting flow cell studies, an overnight experiment can require 500 mL (Foster and Kolenbrander 2004).

A microfluidic system, either custom-made or available commercially, removes such a limitation and also allows, by virtue of its small footprint, multiple biofilm experiments to be run in parallel (Fig. 8.11). The potential for linking such a system to 3D imaging systems is only now just being realized and an opportunity to create high-throughput screens of antimicrobial or biofilm-structure-altering compounds can be explored.

Many biofilm systems use either medium or artificial saliva as the nutrient source. This is primarily due to the inherent difficulties in collecting large enough quantities of human saliva. These types of artificial media can have significant effects on biofilm composition and also the responsiveness of the species to environmental changes or chemical challenges. As a result, the use of pooled human saliva as an inoculum and as a medium source is gaining popularity in model oral biofilm systems (Foster and Kolenbrander 2004; Ledder et al. 2006; McBain et al. 2005), although saliva quantity is an issue. High-throughput approach has the potential to reproducibly grow oral multispecies biofilms that contain species that are indigenous to dental plaque.

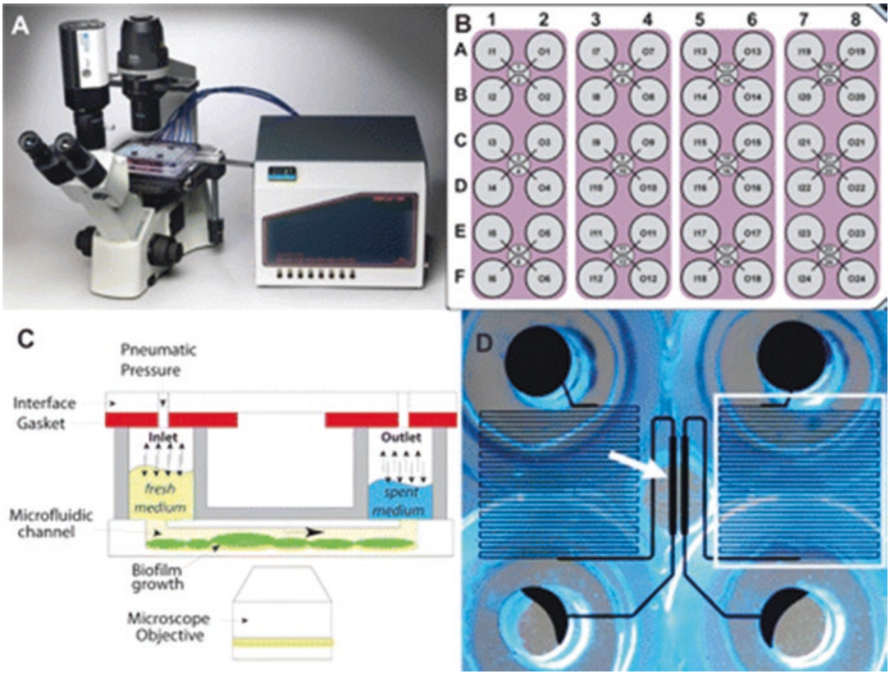


Fig. 8.11 BioFlux high-throughput system for screening of flow biofilm viability

8.6.3 Nanomaterials

Micro-organisms within a biofilm are able to protect themselves from the immune system and antibiotics (Smith 2005). Bacteria are estimated to be 10–1000 times more tolerant to host defences and antibiotics than in their planktonic state (Costerton et al. 1999; Smith 2005). The approaches employed to resist biofilm formation are either the production of cytotoxic materials designed to kill bacteria upon contact (Kuroda and Caputo 2013) or anti-adhesion strategies whereby the materials circumvent bacterial attachment, biofilm formation. Compared to antibiotic-containing materials, surfaces that resist bacterial attachment do not induce the evolutionary pressure, which would lead to bacterial resistance. This characteristic means that this class of material is of particular interest in an age of growing antibiotic resistance. The mechanisms that have been employed to prevent attachment include electrostatic repulsion, steric repulsion, topography and hydration (Magennis et al. 2016). In order to optimize the rate at which new biomaterials could be discovered and their biological properties assessed, the microarray format has now become routine. In this way, hundreds of unique polymers are generated on-slide and assayed on a single substrate in a single experiment. Surface analysis techniques such as time of flight secondary ion mass spectroscopy (ToF-SIMS), atomic force microscopy (AFM), surface wettability measured through water contact angles (WCA), surface plasmon resonance (SPR) and X-ray photoelectron spectroscopy (XPS) allow for rapid characterization of polymer microarrays (Fig. 8.12). Together with the microarray format, these techniques are known as high-throughput surface characterization (HTSC) (Davies et al. 2010).

Research is already underway into bioinspired devices where ligands and proteins direct cell behaviours such as colonization and proliferation, so-called 'third-generation' biomaterials (Hench and Polak 2002). High-throughput strategies are leading to novel material discovery when large number scan be screened and 'hits' identified retrospectively rather than planning those to yield positive results. Manufactured nanomaterials (NMs, materials with at least one dimension <100 nm) and nanoparticles (NPs, NMs with all three dimensions <100 nm) are considered as distinct from normal chemical compounds on account of their size, chemical composition, shape, surface structure, surface charge, aggregation and solubility (Donaldson and Poland 2013).

At present, the very limited and often conflicting data derived from published literature—and the fact that different NMs are physicochemically so heterogeneous—make it difficult to generalize about health risks associated with exposure to NMs. The adoption of high-throughput screening (HTS) and high-content analysis (HCA) for nanomaterial (NM) toxicity testing allows the testing of numerous materials at different concentrations and, on different types of cells, reduces the effect of inter-experimental variation, and makes substantial savings in time and cost. HCA and HTS approaches should deliver information on key biological indicators of NM–cell interactions, such as cell proliferation, cellular morphology, membrane permeability, lysosomal mass/pH, DNA and chromosome damage, activation of

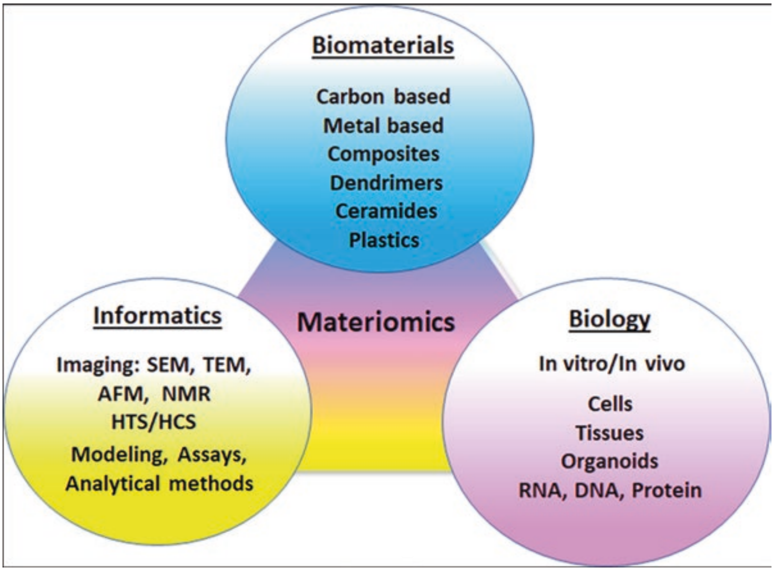


Fig. 8.12 Interface of material science, informatics, and biology

transcription factors, mitochondrial membrane potential changes, oxidative stress monitoring and post-translational modification (Prina-Mello et al. 2013).

8.7 Scientific Databases: The Goldmines of Research

Cell and tissue culture are essential pre-requisites for many for biotechnological research. Advances in research on human and other cells have led to vast knowledge expansion in fields such as cancer research, genetics and public health. This in turn is associated with a corresponding increase in related scientific literature. It is practically a daunting task to identify specific information pertaining to individual research needs. Availability of a practical, user-friendly database containing cell lines, plasmids, vectors, selection agents, concentrations and media would be a great advantage. A database consisting of over 3900 cell lines and 1900 plasmids/vectors collected from 2700 pieces of published literature was established and is being expanded (Amirkia and Qiubao 2012). The electronic web-based version of the database can be accessed at http://celllines.toku-e.com/. With continual addition of data, the database can greatly aid future research.

The European Bioinformatics Institute (EBI) of the European Molecular Biology Laboratory (EMBL) is involved in building and providing biological databases to support both data submission and utilization. A number of free databases are operated and include EMBL Nucleotide Sequence Database (EMBL-Bank), the Protein Databases (SWISS-PROT and TrEMBL), the Macromolecular Structure Database

(MSD) and ArrayExpress for gene expression (Stoesser et al. 2002). As a result of genome sequencing effects, the EMBL Nucleotide Sequence Database is growing rapidly, and necessary scientific information can be accessed at http://www.ebi.ac.ukembl/. New nucleotide sequences or biological information can be submitted by individual scientists or sequencing groups through submission portals such as *Webin* or Sequin, and prior vector contamination screening using interactive web-based services can be utilized (Stoesser et al. 2002). The list of EMBL-Bank web-based resources including detailed information on submissions, data access, genome data as well as database searching and analysis tools is available in the literature (Stoesser et al. 2002).

Gene polymorphisms play a determining role in defining the basis of phenotypic references between individual that has intricate relationships in disease predisposition and drug responses. Human Genic Bi-Allelic Sequences (HGBASE) is a resource of human gene-linked polymorphisms (Brookes et al. 2000). Information gathered from other public resources are systematically screened to avoid redundancy, and these polymorphism records are provided in a standardized user-friendly database in conjunction with other available public resources. The records are categorized as (i) single base differences, (ii) insertion–deletion variants, (iii) simple tandem repeat polymorphisms and (iv) 'generic' (or complex) changes involving alterations not described by the preceding three alternatives (Brookes et al. 2000). Data collection and submission can be done using standard formats and guidelines provided in the website and can be accessed at http://hgbase.interactiva.de.

The information system for molecular biology by the National Institutes of Health (NIH) is the National Center for Biotechnology Information (NCBI). Apart from the GenBank nucleic acid resource that supports data analysis and retrieval, resource links to other biological data are available in the NCBI website and can be accessed at http://www.ncbi.nlm.nih.gov. The available resources under the NCBI website include the Database retrieval tools such as Entrez, 'PubMed', LocusLink and The Taxonomy Browser. The data analysis resources include BLAST, Electronic PCR, OrfFinder, RefSeq, UniGene, database for SNPs (dbSNP), Cancer Chromosome Aberration Project (CCAP), Cancer Genome Anatomy Project (CGAP), SAGEmap, Gene Expression Omnibus (GEO), Online Mendelian Inheritance in Man (OMIM) and many more (Wheeler et al. 2006).

Immune system requires coordinated expression of many genes and proteins to mediate their function. In tandem with the explosion of genomic and proteomic data, the molecular data related to complex human immune system are readily available covering cellular, structural or organ levels for both normal and diseased states. The Immunome Knowledge Base (IKB) is a dedicated resource for immunological information and is formed by integration of three earlier databases, namely, 'Immunome', 'ImmTree' and 'ImmunomeBase' (Ortutay and Vihinen 2009). IKB is freely available for academic research at http://bioinf.uta.fi/IKB/.

Changing patterns in DNA methylation are early even in cancer development. Hypomethylation of the gene promoter regions (CpG islands) is associated with increased gene activity as seen in various cancers, while hypermethylation is associated with gene repression or silencing. Therefore DNA methylation analysis will help in better understanding the process of tumour development and progression and serve in prognostic assessment (Jones and Baylin 2002; Laird 2003). A sequence similarity search program based on the original BLAST algorithm but querying in silico bisulphite modified genome sequences to screen oligonucleotide sequence similarities was developed and is known as methBLAST. In addition, methPrimerDB as database for storage and retrieval of validated PCR-based methylation assays was also developed (Pattyn et al. 2006). Free public access to perform methBLAST searches or submit user-based information is possible. The methBLAST and methPrimerDB can be accessed at http://medgen.ugent.be/methblast and http://medgen.ugent.be/methprimerdb.

Other additional useful databases are (i) ZINC a free public resource for ligand discovery and can be accessed at http://zinc.docking.org (Irwin et al. 2012); (ii) pathway analysis-related databases, KEGG PATHWAY database (Kanehisa and Goto 2000), BioCArta (Nishimura 2001), DAVID (Dennis et al. 2003), GenMAPP (Dennis et al. 2003), GeneOntology (Ashburner et al. 2000) and PathAct (Mogushi and Tanaka 2013); (iii) protein-related databases, UniProt (Apweiler et al. 2004), UniRef (Suzek et al. 2007), neXtProt (Lane et al. 2011); (iv) metabolome-related databases, human metabolome database (Wishart et al. 2016) and metabolomics workbench (Sud et al. 2015); and (v) microRNA-related databases, miRbase (Griffiths-Jones et al. 2006), miRWalk (Dweep et al. 2011) and miRTarBase (Chou et al. 2015). Numerous other databases are available, and listing or detailing them is beyond the scope of this book chapter. Some of the mentioned databases in this chapter (Table 8.1) are intended to create awareness and serve as a guide for the beginners, especially for students undertaking scientific research.

Table 8.1 Useful web resources

Title	URL
BLAST	https://blast.ncbi.nlm.nih.gov/Blast.cgi
dbSNP	https://www.ncbi.nlm.nih.gov/snp
EMBL nucleotide sequence database	www.ebi.ac.uk/embl/
EMBL-EBI home page	www.ebi.ac.uk/
Entrez	https://www.ncbi.nlm.nih.gov/Web/Search/entrezfs.html
Expressed Sequence Tag (EST) resources	www.ebi.ac.uk/embl/Access/est.html
LocusLink	https://www.ncbi.nlm.nih.gov/Web/Newsltr/Summer99/locus.html
PAH gene database	http://www.mcgill.ca/pahdb/
RefSeq	https://www.ncbi.nlm.nih.gov/refseq/
Sequence Retrieval Service (SRS)	http://srs.ebi.ac.uk/
SEQUIN	https://www.ebi.ac.uk/Services/Sequin
Taxonomy Browser	https://www.ncbi.nlm.nih.gov/Taxonomy/Browser/wwwtax.cgi
UniGene	http://www.bioinfo.org.cn/relative/NCBI-UniGene.htm
WEBIN	www.ebi.ac.uk/embl/Submission/

8.8 Conclusions

Scientific advancements are continuously changing the landscape of clinical fields including medicine, dentistry, pharmacy and nursing. These advancements have led us to better understand our genome, proteome and metabolome which significantly impact most if not all aspects of life and hence clinical practice. Like the numerous benefits witnessed with regenerative medicine and tissue engineering in other disciplines, they are set to revolutionize the field of dentistry too. Some of the applications will include (i) use of engineered cells to promote faster growth and filling of the cavities, (ii) restoration of tooth with normal formation of dentin and enamel, (iii) selection of materials/implants with surfaces that naturally inhibit microbial interference, (iv) customization of disease resistance dental tissues and (v) personalized orthodontics. The availability of vast scientific information and technological resources if rightly exploited will have tremendous benefits in both medicine and dentistry.

Acknowledgements The authors acknowledge the grant (13-MED2437-03) funding provided by the National Plan for Science, Technology and Innovation (MAARIFAH), King Abdulaziz City for Science and Technology, The Kingdom of Saudi Arabia. They also acknowledge with thanks the Science and Technology Unit (STU), King Abdulaziz University for their excellent technical support.

Statement of Conflicts of Interest All authors have no conflicts of interests.

References

Akintoye SO, Lam T, Shi S, Brahim J, Collins MT, Robey PG (2006) Skeletal site-specific characterization of orofacial and iliac crest human bone marrow stromal cells in same individuals. Bone 38:758–768

Ali J, Sabiha B, Jan HU, Haider SA, Khan AA, Ali SS (2017) Genetic etiology of oral cancer. Oral Oncol 70:23–28

Alongi DJ et al (2010) Stem/progenitor cells from inflamed human dental pulp retain tissue regeneration potential. Regen Med 5:617–631

Amirkia V, Qiubao P (2012) Cell-culture database: literature-based reference tool for human and mammalian experimentallybased cell culture applications. Bioinformation 8:237

Apweiler R et al (2004) UniProt: the universal protein knowledgebase. Nucleic Acids Res 32:D115–D119

Arnsdorf EJ, Jones LM, Carter DR, Jacobs CR (2009) The periosteum as a cellular source for functional tissue engineering. Tissue Eng A 15:2637–2642

Ashburner M et al (2000) Gene ontology: tool for the unification of biology. Nat Genet 25:25

Beikler T, Peters U, Prior K, Eisenacher M, Flemmig TF (2008) Gene expression in periodontal tissues following treatment. BMC Med Genet 1:30. https://doi.org/10.1186/1755-8794-1-30

Bentley DR et al (2008) Accurate whole human genome sequencing using reversible terminator chemistry. Nature 456:53–59. https://doi.org/10.1038/nature07517

Ben-Yakar A (2019) High-content and high-throughput in vivo drug screening platforms using microfluidics. Assay Drug Dev Technol 17:8–13

Bernimoulin JP (2003) Recent concepts in plaque formation. J Clin Periodontol 30:7–9

Bongso A, Fong C-Y, Ng S-C, Ratnam S (1994) Fertilization and early embryology: isolation and culture of inner cell mass cells from human blastocysts. Hum Reprod 9:2110–2117

Braslavsky I, Hebert B, Kartalov E, Quake SR (2003) Sequence information can be obtained from single DNA molecules. Proc Natl Acad Sci U S A 100:3960–3964. https://doi.org/10.1073/pnas.0230489100

Brookes AJ et al (2000) HGBASE: a database of SNPs and other variations in and around human genes. Nucleic Acids Res 28:356–360

Carinci F, Palmieri A, Girardi A, Cura F, Scapoli L, Lauritano D (2015) Genetic risk assessment of periodontal disease in healthy patients. J Forensics Res 6:260

Catón J, Bostanci N, Remboutsika E, De Bari C, Mitsiadis TA (2011) Future dentistry: cell therapy meets tooth and periodontal repair and regeneration. J Cell Mol Med 15:1054–1065

Chen F-M, Jin Y (2010) Periodontal tissue engineering and regeneration: current approaches and expanding opportunities. Tissue Eng Part B Rev 16:219–255

Chen F-M et al (2016) Treatment of periodontal intrabony defects using autologous periodontal ligament stem cells: a randomized clinical trial. Stem Cell Res Ther 7:33

Chimenti I et al (2012) Isolation and expansion of adult cardiac stem/progenitor cells in the form of cardiospheres from human cardiac biopsies and murine hearts. In: Somatic stem cells: Methods and Protocols. Humana Press, Totowa, NJ, pp 327–338

Cho M, Cho T-J, Lim JM, Lee G, Cho J (2013) The establishment of mouse embryonic stem cell cultures on 96-well plates for high-throughput screening. Mol Cells 35:456–461

Chou C-H et al (2015) miRTarBase 2016: updates to the experimentally validated miRNA-target interactions database. Nucleic Acids Res 44:D239–D247

Cooper PR, Chicca IJ, Holder MJ, Milward MR (2017) Inflammation and regeneration in the dentin-pulp complex: net gain or net loss? J Endod 43:S87–S94

Costerton JW, Stewart PS, Greenberg EP (1999) Bacterial biofilms: a common cause of persistent infections. Science 284:1318–1322

Crawford PJ, Aldred M, Bloch-Zupan A (2007) Amelogenesis imperfecta. Orphanet J Rare Dis 2:17

Dabeva MD, Shafritz DA (2003) Hepatic stem cells and liver repopulation. In: Seminars in liver disease, 2003. Vol 04. Copyright© 2003 by Thieme Medical Publishers, Inc., 333 Seventh Avenue, New York, NY 10001, USA. Tel.:+ 1 (212) 584-4662, pp 349–362

Davies MC et al (2010) High throughput surface characterization: a review of a new tool for screening prospective biomedical material arrays. J Drug Targeting 18:741–751

De Coppi P et al (2006) Isolation of mesenchymal stem cells from human vermiform appendix. J Surg Res 135:85–91

Dennis G Jr, Sherman BT, Hosack DA, Yang J, Gao W, Lane HC, Lempicki RA (2003) DAVID: database for annotation, visualization, and integrated discovery. Genome Biol 4:P3

Dominici M et al (2006) Minimal criteria for defining multipotent mesenchymal stromal cells. The International Society for Cellular Therapy position statement. Cytotherapy 8:315–317

Donaldson K, Poland CA (2013) Nanotoxicity: challenging the myth of nano-specific toxicity. Current Opinion in Biotechnology 24:724–734

Dowthwaite GP et al (2004) The surface of articular cartilage contains a progenitor cell population. J Cell Sci 117:889–897

Dweep H, Sticht C, Pandey P, Gretz N (2011) miRWalk–database: prediction of possible miRNA binding sites by "walking" the genes of three genomes. J Biomed Inform 44:839–847

Ferreira JN, Rungarunlert S, Urkasemsin G, Adine C, Souza GR (2016) Three-dimensional bioprinting nanotechnologies towards clinical application of stem cells and their secretome in salivary gland regeneration. Stem Cells Int 2016:7564689

Foster JS, Kolenbrander PE (2004) Development of a multispecies oral bacterial community in a saliva-conditioned flow cell. Appl Environ Microbiol 70:4340–4348. https://doi.org/10.1128/aem.70.7.4340-4348.2004

Friedenstein A, Piatetzky-Shapiro I, Petrakova K (1966) Osteogenesis in transplants of bone marrow cells. J Embryol Exp Morphol 16:381–390

Gallo R et al (2007) Generation and expansion of multipotent mesenchymal progenitor cells from cultured human pancreatic islets. Cell Death Differ 14:1860

Gargett CE, Schwab KE, Zillwood RM, Nguyen HP, Wu D (2009) Isolation and culture of epithelial progenitors and mesenchymal stem cells from human endometrium. Biol Reprod 80:1136–1145

Gilbert P, Maira-Litran T, McBain AJ, Rickard AH, Whyte FW (2002) The physiology and collective recalcitrance of microbial biofilm communities. Adv Microbial Physiology 46:202–256

Green ED, Watson JD, Collins FS (2015) Human Genome Project: twenty-five years of big biology. Nature 526:29

Griffiths-Jones S, Grocock RJ, van Dongen S, Bateman A, Enright AJ (2006) miRBase: microRNA sequences, targets and gene nomenclature. Nucleic Acids Res 34:D140–D144. https://doi.org/10.1093/nar/gkj112

Gronthos S, Mankani M, Brahim J, Robey PG, Shi S (2000) Postnatal human dental pulp stem cells (DPSCs) in vitro and in vivo. Proc Natl Acad Sci 97:13625–13630

Guzmán-Armstrong S (2005) Rampant caries. J Sch Nurs 21:272–278

Hench LL, Polak JM (2002) Third-generation biomedical materials. Science 295:1014–1017

Hook AL, Anderson DG, Langer R, Williams P, Davies MC, Alexander MR (2010) High throughput methods applied in biomaterial development and discovery. Biomaterials 31:187–198

Huang GT-J, Sonoyama W, Liu Y, Liu H, Wang S, Shi S (2008) The hidden treasure in apical papilla: the potential role in pulp/dentin regeneration and bioroot engineering. J Endod 34:645–651

Irwin JJ, Sterling T, Mysinger MM, Bolstad ES, Coleman RG (2012) ZINC: a free tool to discover chemistry for biology. J Chem Inf Model 52:1757–1768

Jenkinson HF (2011) Beyond the oral microbiome. Environ Microbiol 13:3077–3087

Jones PA, Baylin SB (2002) The fundamental role of epigenetic events in cancer. Nat Rev Genet 3:415

Kanehisa M, Goto S (2000) KEGG: Kyoto encyclopedia of genes and genomes. Nucleic Acids Res 28:27–30

Kang S-G et al (2010) Isolation and perivascular localization of mesenchymal stem cells from mouse brain. Neurosurgery 67:711–720

Kim YG et al (2016) Transcriptome sequencing of gingival biopsies from chronic periodontitis patients reveals novel gene expression and splicing patterns. Hum Genomics 10:28. https://doi.org/10.1186/s40246-016-0084-0

Kolenbrander PE (2000) Oral microbial communities: biofilms, interactions, and genetic systems. Annu Rev Microbiol 54:413–437

Kuroda K, Caputo GA (2013) Antimicrobial polymers as synthetic mimics of host-defense peptides. Wiley Interdiscip Rev Nanomed Nanobiotechnol 5:49–66

Laird PW (2003) Early detection: the power and the promise of DNA methylation markers. Nat Rev Cancer 3:253

Lander ES (2011) Initial impact of the sequencing of the human genome. Nature 470:187–197. https://doi.org/10.1038/nature09792

Lane L et al (2011) neXtProt: a knowledge platform for human proteins. Nucleic Acids Res 40:D76–D83

Ledder RG, Gilbert P, Pluen A, Sreenivasan PK, De Vizio W, McBain AJ (2006) Individual microflora beget unique oral microcosms. J Appl Microbiol 100:1123–1131. https://doi.org/10.1111/j.1365-2672.2006.02847.x

Levene MJ, Korlach J, Turner SW, Foquet M, Craighead HG, Webb WW (2003) Zero-mode waveguides for single-molecule analysis at high concentrations. Science 299:682–686. https://doi.org/10.1126/science.1079700

Lim J-Y et al (2013) Systemic transplantation of human adipose tissue-derived mesenchymal stem cells for the regeneration of irradiation-induced salivary gland damage. PLoS One 8:e71167

Lin CH, Chang HC, Hsu CH (2016) A microfluidic platform for high-throughput single-cell isolation and culture. J Vis Exp. https://doi.org/10.3791/54105

Loesche WJ (1986) Role of Streptococcus mutans in human dental decay. Microbiol Rev 50:353

Lombaert IM et al (2008) Rescue of salivary gland function after stem cell transplantation in irradiated glands. PLoS One 3:e2063

Lombaert I, Movahednia MM, Adine C, Ferreira JN (2017) Concise review: salivary gland regeneration: therapeutic approaches from stem cells to tissue organoids. Stem Cells 35:97–105

Lu H, Xie C, Zhao Y-M, Chen F-M (2013) Translational research and therapeutic applications of stem cell transplantation in periodontal regenerative medicine. Cell Transplant 22:205–229

Luder HU (2015) Malformations of the tooth root in humans. Front Physiol 6:307

Maeda H, Wada N, Nakamuta H, Akamine A (2004) Human periapical granulation tissue contains osteogenic cells. Cell Tissue Res 315:203–208

Magennis E, Hook A, Davies M, Alexander C, Williams P, Alexander MR (2016) Engineering serendipity: high-throughput discovery of materials that resist bacterial attachment. Acta Biomaterialia 34:84–92

Mah TF, O'Toole GA (2001) Mechanisms of biofilm resistance to antimicrobial agents. Trends Microbiol 9:34–39

Marcus AJ, Woodbury D (2008) Fetal stem cells from extra-embryonic tissues: do not discard. J Cell Mol Med 12:730–742

Margulies M et al (2005) Genome sequencing in microfabricated high-density picolitre reactors. Nature 437:376–380. https://doi.org/10.1038/nature03959

Marrelli M, Paduano F, Tatullo M (2013) Cells isolated from human periapical cysts express mesenchymal stem cell-like properties. Int J Biol Sci 9:1070

Marrelli M, Paduano F, Tatullo M (2015) Human periapical cyst–mesenchymal stem cells differentiate into neuronal cells. J Dent Res 94:843–852

Marsh PD (1994) Microbial ecology of dental plaque and its significance in health and disease. Adv Dent Res 8:263–271

Marsh PD (2006) Dental plaque as a biofilm and a microbial community–implications for health and disease. BMC Oral health BioMed Central 1:S14

Martin C, Li T, Hegarty E, Zhao P, Mondal S, Ben-Yakar A (2018) Line excitation array detection fluorescence microscopy at 0.8 million frames per second. Nat Commun 9:4499

Marynka-Kalmani K, Treves S, Yafee M, Rachima H, Gafni Y, Cohen MA, Pitaru S (2010) The lamina propria of adult human oral mucosa harbors a novel stem cell population. Stem Cells 28:984–995

McBain AJ, Sissons C, Ledder RG, Sreenivasan PK, De Vizio W, Gilbert P (2005) Development and characterization of a simple perfused oral microcosm. J Appl Microbiol 98:624–634. https://doi.org/10.1111/j.1365-2672.2004.02483.x

Miura M, Gronthos S, Zhao M, Lu B, Fisher LW, Robey PG, Shi S (2003) SHED: stem cells from human exfoliated deciduous teeth. Proc Natl Acad Sci 100:5807–5812

Mogushi K, Tanaka H (2013) PathAct: a novel method for pathway analysis using gene expression profiles. Bioinformation 9:394

Mondal S, Hegarty E, Martin C, Gökçe SK, Ghorashian N, Ben-Yakar A (2016) Large-scale microfluidics providing high-resolution and high-throughput screening of Caenorhabditis elegans poly-glutamine aggregation model. Nat Commun 7:13023

Morsczeck C et al (2005) Isolation of precursor cells (PCs) from human dental follicle of wisdom teeth. Matrix Biol 24:155–165

Nair P (2004) Pathogenesis of apical periodontitis and the causes of endodontic failures. Crit Rev Oral Biol Med 15:348–381

Nakashima M, Iohara K, Murakami M, Nakamura H, Sato Y, Ariji Y, Matsushita K (2017) Pulp regeneration by transplantation of dental pulp stem cells in pulpitis: a pilot clinical study. Stem Cell Res Ther 8:61

Nel A, Xia T, Meng H, Wang X, Lin S, Ji Z, Zhang H (2012) Nanomaterial toxicity testing in the 21st century: use of a predictive toxicological approach and high-throughput screening. Acc Chem Res 46:607–621

Nishimura D (2001) BioCarta Biotech Software & Internet Report: The Computer Software Journal for Scient 2:117–120

Nyren P, Lundin A (1985) Enzymatic method for continuous monitoring of inorganic pyrophosphate synthesis. Anal Biochem 151:504–509

Ono H, Obana A, Usami Y, Sakai M, Nohara K, Egusa H, Sakai T (2015) Regenerating salivary glands in the microenvironment of induced pluripotent stem cells. Biomed Res Int 2015:1

Ortutay C, Vihinen M (2009) Immunome knowledge base (IKB): an integrated service for immunome research. BMC Immunol 10:3

Owen M (1985) Lineage of osteogenic cells and their relationship to the stromal system. In: Peck WA, editor. Bone and mineral research, Vol. 3. Amsterdam: Elsevier Science Publishers; pp 1–25

Pamies D et al (2018) Advanced Good Cell Culture Practice for human primary, stem cell-derived and organoid models as well as microphysiological systems. ALTEX 35:353–378. https://doi.org/10.14573/altex.1710081

Park D, Lim J, Park JY, Lee S-H (2015) Concise review: stem cell microenvironment on a chip: current technologies for tissue engineering and stem cell biology. Stem Cells Transl Med 4:1352–1368

Patel J, Gudehithlu KP, Dunea G, Arruda JA, Singh AK (2010) Foreign body-induced granulation tissue is a source of adult stem cells. Transl Res 155:191–199

Pattyn F et al (2006) methBLAST and methPrimerDB: web-tools for PCR based methylation analysis. BMC Bioinformatics 7:496

Peters BM, Jabra-Rizk MA, O'May GA, Costerton JW, Shirtliff ME (2012) Polymicrobial interactions: impact on pathogenesis and human disease. Clin Microbiol Rev 25:193–213. https://doi.org/10.1128/cmr.00013-11

Prasad MK et al (2016) A targeted next-generation sequencing assay for the molecular diagnosis of genetic disorders with orodental involvement. J Med Genet 53:98–110

Prina-Mello A, Crosbie-Staunton K, Salas G, del Puerto Morales M, Volkov Y (2013) Multiparametric toxicity evaluation of SPIONs by high content screening technique: identification of biocompatible multifunctional nanoparticles for nanomedicine Ieee. Trans Magn 49:377–382

Pringle S et al (2016) Human salivary gland stem cells functionally restore radiation damaged salivary glands. Stem Cells 34:640–652

Roberts AP, Mullany P (2010) Oral biofilms: a reservoir of transferable, bacterial, antimicrobial resistance. Expert Rev Anti Infect Ther 8:1441–1450. https://doi.org/10.1586/eri.10.106

Rothberg JM et al (2011) An integrated semiconductor device enabling non-optical genome sequencing. Nature 475:348–352. https://doi.org/10.1038/nature10242

Sato A, Okumura K, Matsumoto S, Hattori K, Hattori S, Shinohara M, Endo F (2007) Isolation, tissue localization, and cellular characterization of progenitors derived from adult human salivary glands. Cloning Stem Cells 9:191–205

Selwitz RH, Ismail AI, Pitts NB (2007) Dental caries. Lancet 369:51–59

Seo B-M et al (2004) Investigation of multipotent postnatal stem cells from human periodontal ligament. Lancet 364:149–155

Shendure J et al (2005) Accurate multiplex polony sequencing of an evolved bacterial genome. Science 309:1728–1732. https://doi.org/10.1126/science.1117389

Short HB, Clark VL, Kellogg DS Jr, Young FE (1982) Anaerobic survival of clinical isolates and laboratory strains of Neisseria gonorrhoea: use in transfer and storage. J Clin Microbiol 15:915–919

Smith AW (2005) Biofilms and antibiotic therapy: is there a role for combating bacterial resistance by the use of novel drug delivery systems? Adv Drug Deliv Rev 57:1539–1550

Smith LM et al (1986) Fluorescence detection in automated DNA sequence analysis. Nature 321:674–679. https://doi.org/10.1038/321674a0

Stewart PS (1996) Theoretical aspects of antibiotic diffusion into microbial biofilms. Antimicrob Agents Chemother 40:2517–2522

Stoesser G et al (2002) The EMBL nucleotide sequence database. Nucleic Acids Res 30:21–26

Stookey GK (2008) The effect of saliva on dental caries. J Am Dent Assoc 139(Suppl):11s–17s

Sud M et al (2015) Metabolomics Workbench: an international repository for metabolomics data and metadata, metabolite standards, protocols, tutorials and training, and analysis tools. Nucleic Acids Res 44:D463–D470

Sumiyama K, Matsumoto N, Garcon-Yoshida J, Ukai H, Ueda HR, Tanaka Y (2018) Easy and efficient production of completely embryonic-stem-cell-derived mice using a micro-aggregation device. PLoS One 13:e0203056. https://doi.org/10.1371/journal.pone.0203056

Suzek BE, Huang H, McGarvey P, Mazumder R, Wu CH (2007) UniRef: comprehensive and non-redundant UniProt reference clusters. Bioinformatics 23:1282–1288

Takahashi K, Yamanaka S (2006) Induction of pluripotent stem cells from mouse embryonic and adult fibroblast cultures by defined factors. Cell 126:663–676

Tatullo M, Codispoti B, Pacifici A, Palmieri F, Marrelli M, Pacifici L, Paduano F (2017) Potential use of human periapical cyst-mesenchymal stem cells (hPCy-MSCs) as a novel stem cell source for regenerative medicine applications. Front Cell Dev Biol 5:103

Thomson JA, Itskovitz-Eldor J, Shapiro SS, Waknitz MA, Swiergiel JJ, Marshall VS, Jones JM (1998) Embryonic stem cell lines derived from human blastocysts science, vol 282, pp 1145–1147

Toma JG, Akhavan M, Fernandes KJ, Barnabé-Heider F, Sadikot A, Kaplan DR, Miller FD (2001) Isolation of multipotent adult stem cells from the dermis of mammalian skin. Nat Cell Biol 3:778

Tran SD et al (2013) Paracrine effects of bone marrow soup restore organ function, regeneration, and repair in salivary glands damaged by irradiation. PLoS One 8:e61632

Tseng SC (1989) Concept and application of limbal stem cells. Eye 3:141

van Dijk EL, Auger H, Jaszczyszyn Y, Thermes C (2014) Ten years of next-generation sequencing technology. Trends Genet 30:418–426. https://doi.org/10.1016/j.tig.2014.07.001

van Vliet E (2011) Current standing and future prospects for the technologies proposed to transform toxicity testing in the 21st century. ALTEX 28:17–44

Van Vliet E et al (2014) Current approaches and future role of high content imaging in safety sciences and drug discovery Alternatives to Animal Experimentation. ALTEX 31:479–493

Velugula-Yellela SR et al (2018) Use of high-throughput automated microbioreactor system for production of model IgG1 in CHO Cells. J Vis Exp JoVE. https://doi.org/10.3791/58231

von Bültzingslöwen I et al (2007) Salivary dysfunction associated with systemic diseases: systematic review and clinical management recommendations. Oral Surg Oral Med Oral Pathol Oral Radiol Endod 103:S57. e51–S57. e15

Wang F et al (2017) Transcriptome analysis of coding and long non-coding RNAs highlights the regulatory network of cascade initiation of permanent molars in miniature pigs. BMC Genomics 18:148. https://doi.org/10.1186/s12864-017-3546-4

Wendell S et al (2010) Taste genes associated with dental caries. J Dent Res 89:1198–1202

Werneck R et al (2011) A major gene effect controls resistance to caries. J Dent Res 90:735–739

Wheeler DL et al (2006) Database resources of the national center for biotechnology information. Nucleic Acids Res 35:D5–D12

Wishart DS, Mandal R, Stanislaus A, Ramirez-Gaona M (2016) Cancer metabolomics and the human metabolome database. Meta 6:10

Yang ZH et al (2009) Apical tooth germ cell-conditioned medium enhances the differentiation of periodontal ligament stem cells into cementum/periodontal ligament-like tissues. J Periodontal Res 44:199–210

Yu J et al (2007) Induced pluripotent stem cell lines derived from human somatic cells. Science 318:1917–1920

Zhang Q, Shi S, Liu Y, Uyanne J, Shi Y, Shi S, Le AD (2009) Mesenchymal stem cells derived from human gingiva are capable of immunomodulatory functions and ameliorate inflammation-related tissue destruction in experimental colitis. J Immunol 183:7787–7798

Zuk PA et al (2001) Multilineage cells from human adipose tissue: implications for cell-based therapies. Tissue Eng 7:211–228

Chapter 9
Multiple Analyte Profiling (xMAP) Technology Coupled with Functional Bioinformatics Strategies: Potential Applications in Protein Biomarker Profiling in Autoimmune Inflammatory Diseases

Peter Natesan Pushparaj

Contents

Abbreviations

CS	Capture Sandwich
DAVID	Database for Annotation, Visualization, and Integrated Discovery
ELISA	Enzyme-Linked Immunosorbent Assays
IPA	Ingenuity Pathway Analysis
ISA	Indirect Serological Assay
KEGG	Kyoto Encyclopedia of Genes and Genomes
MAP	Multi-Analyte Profiling
NBMIs	Microsphere-Based Multiplex Immunoassays

P. N. Pushparaj (✉)
Center of Excellence in Genomic Medicine Research, Faculty of Applied Medical Sciences, King Abdulaziz University, Jeddah, Saudi Arabia

Department of Medical Laboratory Technology, Faculty of Applied Medical Sciences, King Abdulaziz University, Jeddah, Saudi Arabia

N. A. Shaik et al. (eds.), *Essentials of Bioinformatics, Volume II*,
https://doi.org/10.1007/978-3-030-18375-2_9

OA	Osteoarthritis
RA	Rheumatoid Arthritis
x	Analyte or Biomarker

9.1 Introduction

The xMAP (x = analyte or biomarker, MAP = multi-analyte profiling) technology was invented in the 1990s by the scientists at the Luminex Corporation in the United States of America (USA) for the multiple simultaneous detection of analytes in biological samples. It is a major advancement in the high-throughput bioassays using solid-phase isolation method combined with cutting-edge fluidics, optics, and digital signal processing with patented "microsphere" (bead)-based technology. xMAP technology enables rapid, cost-effective, and simultaneous analysis of multiple analytes within a single biological sample. Importantly, it is an open architecture technology and can be configures to formulate an array of assays rapidly, precisely, and cost-effectively. The xMAP technology gives many benefits for the end user, and therefore it is utilized in pharmaceutical, clinical, and research laboratories (Kellar and Iannone 2002; Kellar et al. 2006; Graham et al. 2019). Now xMAP technology is the most commonly used bead-based multiplexing platform with over 15,500 instruments installed, 35,000 peer-reviewed publications, and more than 70 Luminex Partners providing xMAP customers over 1300 research kits as well as custom assay solutions (Graham et al. 2019). The xMAP instruments currently available in the market such as Luminex 200, FLEXMAP 3D, and MAGPIX are shown in Fig. 9.1 (Angeloni et al. 2014). The main aim of this chapter is to discuss the latest findings and applications of xMAP immunoassays coupled with functional bioinformatics strategies to unravel protein biomarkers in autoimmune inflammatory diseases such as rheumatoid arthritis (RA).

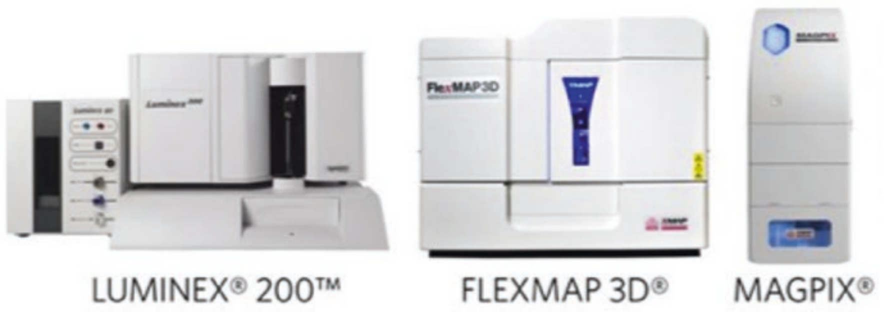


Fig. 9.1 Luminex xMAP instruments currently available in the market such as Luminex 200, FLEXMAP 3D, and MAGPIX (Angeloni et al. 2014). (Courtesy: Luminex Corporation, USA)

9.2 Principle of xMAP Technology

The xMAP technology is based on the principle of microspheres in a liquid suspension as determiners of analyte specificity. Microsphere sets are either polypropylene or magnetic in nature that are impregnated with two spectrally distinct fluorophores. The spectral signature of the microsphere are determined by the different concentrations of internal fluorescent dyes, yielding up to 100 spectrally unique bead sets (Fig. 9.2). Using third internal fluorescent dye, the microspheres can be expanded up to 500 distinct bead sets (Graham et al. 2019). The specific reagents for bioassays like antigens, antibodies, or oligonucleotides can be coupled with each distinct bead type and used in a single assay for the multiplex detection of up to 500 analytes in a single sample. The bead mixture is incubated with the sample, and a fluorescent reporter such as Cy-3, Cy-5, Alexa 532, Streptavidin-R-Phycoerythrin, etc., is coupled to a target molecule that allows the detection of analytes captured on the microsphere surface using a Luminex instrument (Fig. 9.1).

This bead-based suspension array system for measuring analytes provides both medium to high-throughput and high-content data, and researchers may easily scale the number of analytes studied and customize the assays and applications (Lin et al. 2015; Manglani et al. 2019). xMAP technology can be used for antibody array stud-

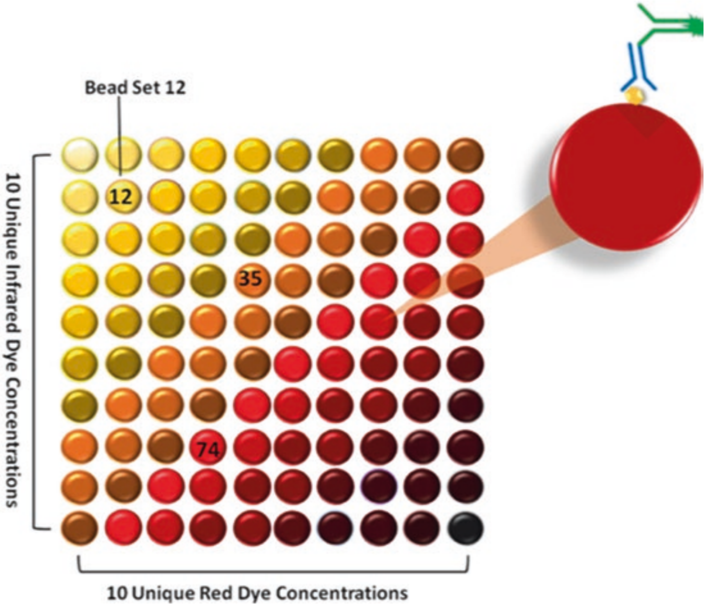

Bead Set 12
12
35
74
10 Unique Infrared Dye Concentrations
10 Unique Red Dye Concentrations

Fig. 9.2 xMAP technology uses internally dyed polypropylene or magnetic microspheres. Luminex color-codes microspheres (beads) internally with specific concentrations of different florescent dyes, providing up to 500 distinctly color-coded microsphere sets. (Adapted and modified from Reslova et al. (2017), https://doi.org/10.3389/fmicb.2017.00055, and this work is licensed under a Creative Commons Attribution 4.0 Generic License)

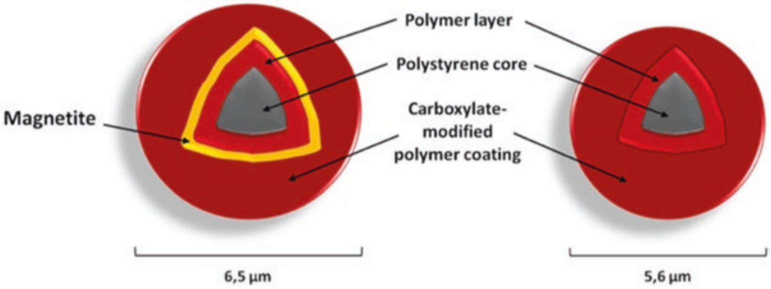


Fig. 9.3 The structure of microsphere. The polystyrene divinylbenzene core is surrounded by a polymer layer, which is formed by polystyrene methacrylic acid (infusion of dyes). The surface of each microsphere is irregular, porous, and carboxylated. Magnetic microspheres have an additional layer of magnetite within the polymer layer and so differ also in size. (Adapted from Reslova et al. (2017), https://doi.org/10.3389/fmicb.2017.00055, and this work is licensed under a Creative Commons Attribution 4.0 Generic License)

Table 9.1 Different types of commercially available microspheres used in xMAP assays[a]

Type of microsphere	Size of microsphere (µm)	Nature of microsphere	Maximum number of sets	Compatible xMAP instrument	Type of analyte
MicroPlex	5.6	Nonmagnetic	100	Flow cytometry-based	All
MagPlex	6.5	Magnetic	500	All xMAP	All
MagPlex-TAG	6.5	Magnetic	150	All xMAP	Nucleic acid
LumAvidin	5.6	Nonmagnetic	100	Flow cytometry-based	Proteins
SeroMAP	5.6	Nonmagnetic	100	Flow cytometry-based	Proteins

[a] Adapted and modified from Reslova et al. (2017), https://doi.org/10.3389/fmicb.2017.00055, and this work is licensed under a Creative Commons Attribution 4.0 Generic License

ies, as the workflow is simple and does not need purification, and picomolar detection levels and dynamic ranges of more than three orders of magnitude have been achieved. As a result, xMAP suspension microsphere arrays have been utilized in an array of biomarker studies. Importantly, based on the type of Luminex instrument used, up to 500 bead sets can be used in each well of a 96- or 384-well plate, generating a high-throughput measurement of protein or oligonucleotide targets (Fig. 9.3).

Many types of microspheres are commercially available for the xMAP assays (Table 9.1), and their selection is determined by the type of instrumentation used, detection mode, and the number of analytes or biomarkers of interest (Table 9.2) (Dunbar and Li 2010; Houser 2012; Reslova et al. 2017). Normal xMAP microspheres are 5.6 µm polystyrene beads with approximately 100 million carboxyl groups (COOH) on the surface for covalent coupling of capture reagents (Tang and Stratton 2006; Angeloni et al. 2014). On the other hand, the magnetic microspheres

Table 9.2 The list of Luminex instruments used for xMAP assays[a]

Type of xMAP instrument	Analytes per reaction	Detection method	Compatible microspheres	Type of microplate
Luminex MAGPIX®	50	Immobilization of microspheres in magnetic field	Magnetic microspheres	96-well plate
Luminex100®/200™	100 (80 with MagPlex)	Flow cytometry-based	All types of microspheres	96-well plate
FlexMAP 3D®	500	Flow cytometry-based	All types of microspheres	96 and 384-well plate

[a] Adapted and modified from Reslova et al. (2017), https://doi.org/10.3389/fmicb.2017.00055, and this work is licensed under a Creative Commons Attribution 4.0 Generic License

(Fig. 9.2) vary in size and structure by the addition of a magnetite layer (Dunbar and Li 2010; Houser 2012; Reslova et al. 2017). The efficiency of washing is greatly increased in the xMAP assays using magnetic beads as the magnetic separation step augments the elimination of unwanted or unbound constituents of the sample. Importantly, the magnetic MagPlex-TAG beads are covalently linked with specific 24 base pair-(bp)-long anti-TAG oligonucleotides that bind with the target sequences with the complementary TAG sequence. It is termed as the xTAG technology and optimized to have least cross-reactivity with other non-specific oligonucleotide sequences in the sample (Babady et al. 2012; Angeloni et al. 2014).

9.3 Mechanism of Signal Detection in xMAP Instruments

In the Luminex xMAP instruments, the beads are analyzed mostly with two different lasers. The inner fluorescent dyes of the beads are excited by red classification laser/LED (635 nm) for the specific identification and classification of microsphere set based on its spectral signature. The green reporter laser/LED (525–532 nm) recognizes the fluorescent reporter bound to the captured analyte on the bead surface. The emission spectra of both red and green lasers are simultaneously read in purpose-designed xMAP readers (Table 9.2). The xMAP instruments differ by their mechanisms of fluorescence detection and by the maximum number of samples analyzed in a single sample (Angeloni et al. 2014).

The basic MAGPIX xMAP instrument is only compatible with magnetic beads such as MagPlex and MagPlex-TAG. The principle of xMAP assay in the MAGPIX instrument is based on the immobilization of magnetic beads in the monolayer on the magnetic surface (Fig. 9.4). Contrary to the flow-based xMAP instruments, the fluorescent imager of the MAGPIX system reads all the microspheres at once. The reading of a 96-well-plate in the MAGPIX system takes about 60 min, and the maximum reading capacity is currently limited to 50 bead sets (Angeloni et al. 2014).

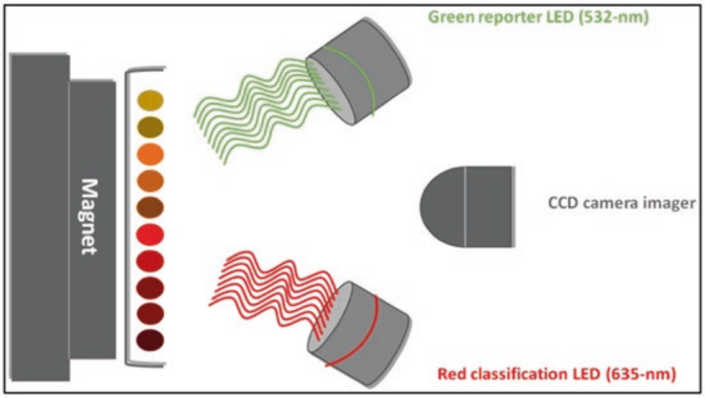


Fig. 9.4 Principle of MAGPIX fluorescent imager. The immobilized MagPlex microspheres on the magnet are recognized by LEDs and recorded as a picture by a CCD camera (LED light-emitting diode, CCD charge-coupled device). (Adapted from Reslova et al. (2017), https://doi.org/10.3389/fmicb.2017.00055, and this work is licensed under a Creative Commons Attribution 4.0 Generic License)

9.4 Microsphere-Based Multiplex Immunoassay (MBMI)

The concentration or the detection of a particular analyte (protein) in a biological sample or solution is done by microsphere-based multiplex immunoassays (MBMIs) using an antibody or immunoglobulin (Angeloni et al. 2014). In conventional enzyme-linked immunosorbent assays (ELISA), a single analyte is measured in a sample. However, multiplex detection of more than one analyte of interest in a sample simultaneously is not possible using conventional ELISA (Bokken et al. 2012) and requires relatively large volume of sample, negligible non-specific binding or increased background. MBMIs are alternative to conventional ELISA, and conventional ELISA assays can easily be converted to the MBMI format using an uncomplicated, efficient, and cost-saving method with a superior range and sensitivity (Angeloni et al. 2014). The commonly used methods in MBMI are capture sandwich (CS), indirect serological assay (ISA), and competitive ELISA. The competitive ELISA (Type I) enables detection of an analyte with a single capture antibody linked to the surface of a microsphere and a competitive, labelled antigen reversibly linked to the antibody, whereas in the competitive ELISA (Type 2), the assay format is reversed with the antigen attached to the microsphere and the antibody labelled (Fig. 9.5) (Bjerre et al. 2009).

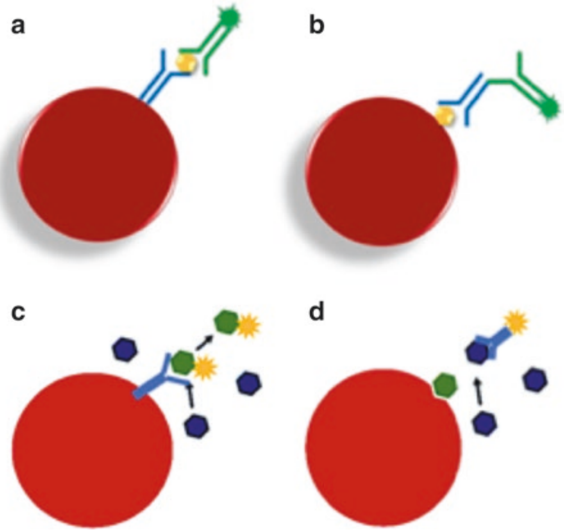


Fig. 9.5 Principle of microsphere-based multiplex immunoassays. (**a**) Capture sandwich (CS; yellow hexagon = target; blue Y = capture antibodies; green Y = detection antibody; green star = fluorescent reporter); (**b**) indirect serological assay (ISA; yellow hexagon = capture antigen; blue Y = specific target antibody; green Y = detection anti-antibody; green star = fluorescent reporter). (**c**) Competitive ELISA (Type I) enables detection of an analyte with a single capture antibody linked to the surface of a microsphere and a competitive, labelled antigen reversibly linked to the antibody. (**d**) In the competitive ELISA (Type 2), the assay format is reversed with the antigen attached to the microsphere and the antibody labelled. (Adapted and modified from both Angeloni et al. (2014) and Reslova et al. (2017), and this work is licensed under a Creative Commons Attribution 4.0 Generic License)

9.5 xMAP Technology in Biomarker Profiling in Rheumatoid Arthritis

xMAP technology is an open architecture system offered by Luminex to customers and commercial partners to develop multiplex assays in an array of formats for a variety of applications (Graham et al. 2019). The xMAP technology is used in many different applications such as the identification of disease-specific target proteins present in the biotinylated samples using antibody suspension bead arrays (Darmanis et al. 2013) (Fig. 9.6). Some of the key applications are the biomarker discovery and profiling, vaccine development, mapping signaling networks, transplant medicine and HLA testing, pathogen detection, etc. (Dunbar and Hoffmeyer 2013; Reslova et al. 2017; Graham et al. 2019).

Here, we describe the use of xMAP technology for the multiplex detection of an array of cytokines, chemokines, and growth factors in the serum of patients suffering from autoimmune diseases such as rheumatoid arthritis (RA) using microsphere-based

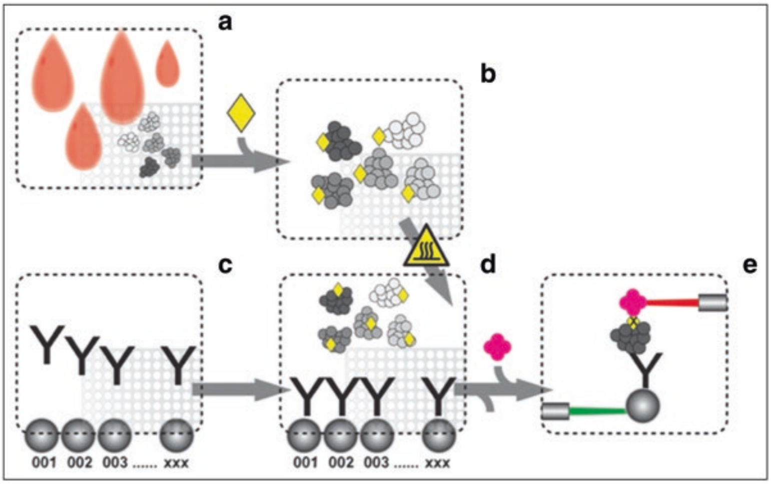


Fig. 9.6 The experimental steps involved in the antibody suspension bead arrays using biotinylated samples. The samples were distributed into the microtiter plates in a defined and randomized manner. (**a**) The proteins in the samples are labelled with biotin, (**b**) and beads with distinct color codes are coupled with antibodies to create a suspension bead array. (**c**) Beads and samples are mixed for incubation after the samples have been heat treated in assay buffer to expose the epitopes. (**d**) The unbound proteins and antibodies are removed, and fluorescent (phycoerythrin) streptavidin is added for detection. (**d**) Each bead type is then identified via a red laser, and the emitted reporter fluorescence of each bead of the same type is determined using a green laser. The mean fluorescence intensity (MFI) for each bead type is a measure of the presence and amount of a specific protein present in the sample that has reacted with its corresponding antibody, attached to the beads (**e**) using the Luminex instrument. (Adapted and modified from Darmanis et al. (2013), https://doi.org/10.1371/journal.pone.0081712.g003, and this work is licensed under a Creative Commons Attribution 2.0 Generic License)

multiplex immunoassay formats (MBMI) described above (Bahlas et al. 2019). Rheumatoid arthritis (RA) is an autoimmune inflammatory disease demonstrated by synovitis and joint destruction associated with comorbidities affecting the bone, brain, lungs, and underlying vasculature (Smolen et al. 2016; McInnes and Schett 2017). RA could gradually lead to permanent disability and severely affects the socioeconomic status of these patients (Siebert et al. 2015; McInnes and Schett 2017; Firestein and McInnes 2017). An array of genetic and environmental factors are responsible for the enteropathogenesis of RA, mainly by increasing the biosynthesis of proinflammatory cytokines compared to anti-inflammatory cytokines both systemically in the blood and the synovial membranes of the joints (Siebert et al. 2015; McInnes et al. 2016; McInnes and Schett 2017; Firestein and McInnes 2017). Studies have shown that the levels of proinflammatory mediators are significantly higher than the anti-inflammatory mediators in the RA synovial membrane and potentiate the damage of adjacent cartilages and bone erosion (Siebert et al. 2015; McInnes and Schett 2017; Firestein and McInnes 2017).

Table 9.3 The list of analytes present in the LHC6003M xMAP kit[a]

Cytokines	Chemokines	Growth factors
G-CSF	Eotaxin	EGF
GM-CSF	IP-10	FGF-basic
IFN-α	MCP-1	HGF
IFN-γ	MIG	VEGF
IL-1β	MIP-1α	
IL-1RA	MIP-1β	
IL-2	RANTES	
IL-2R		
IL-4		
IL-5		
IL-6		
IL-7		
IL-8		
IL-10		
IL-12 (p40/p70)		
IL-13		

[a] Adapted from Thermo Fisher Scientific, USA

We have recently measured an array of cytokines, chemokines, and growth factors by the Human Cytokine Magnetic 30-Plex Panel (LHC6003M) according to the manufacturer's instructions (Thermo Fisher Scientific, USA). The plasma samples of healthy volunteers ($n = 10$), osteoarthritis ($n = 10$), and RA patients ($n = 25$) who met the diagnostic criteria of 2010 ACR/EULAR (5) (Bahlas et al. 2019) were used for the xMAP assay using the MAGPIX instrument (Luminex Corporation, USA). The Human Cytokine Magnetic 30-Plex Panel consists of an array of cytokines, chemokines, and growth factors as listed in Table 9.3.

The raw data obtained for all the 30 different analytes was analyzed by the Luminex xPONENT® multiplex assay analysis software (Luminex Corporation, USA) to calculate the absolute concentration. Additionally, the concentration of each analyte determined was further analyzed using GraphPad Prism (Version 7) software to compute the statistical significance using student's unpaired t-Test (two-tailed) (Figs. 9.7 and 9.8). The P values ≤0.05 were considered to be statistically significant (Bahlas et al. 2019). Besides, there are other software packages such as R's drLumi package to read and analyze xPONENT®-derived multiplexed data (Breen 2017).

9.6 Functional Bioinformatics Analysis of xMAP Data

One of the major challenges encountered by the research and development sectors of pharmaceutical companies is the construction of cellular and molecular signaling networks and the identification of disease and drug-specific signatures for the

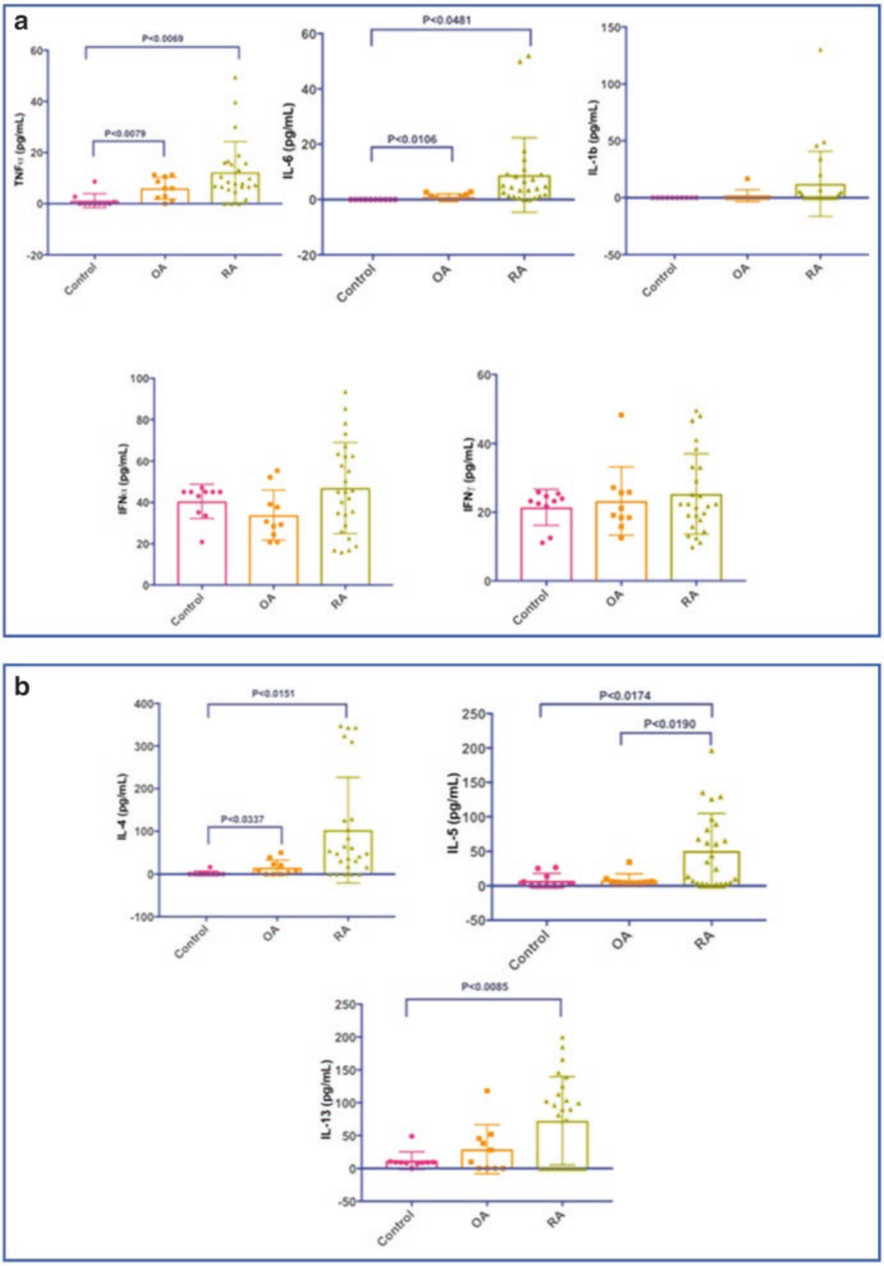


Fig. 9.7 The levels of (**a**) Th1cytokines, (**b**) Th2 cytokines, and (**c**) chemokines in the plasma of RA patients with active disease, OA patients, and normal controls. The plasma concentrations (pg/mL) of all the analytes are expressed as mean ± SD. $P < 0.05$ was considered to be statistically significant (Bahlas et al. 2019)

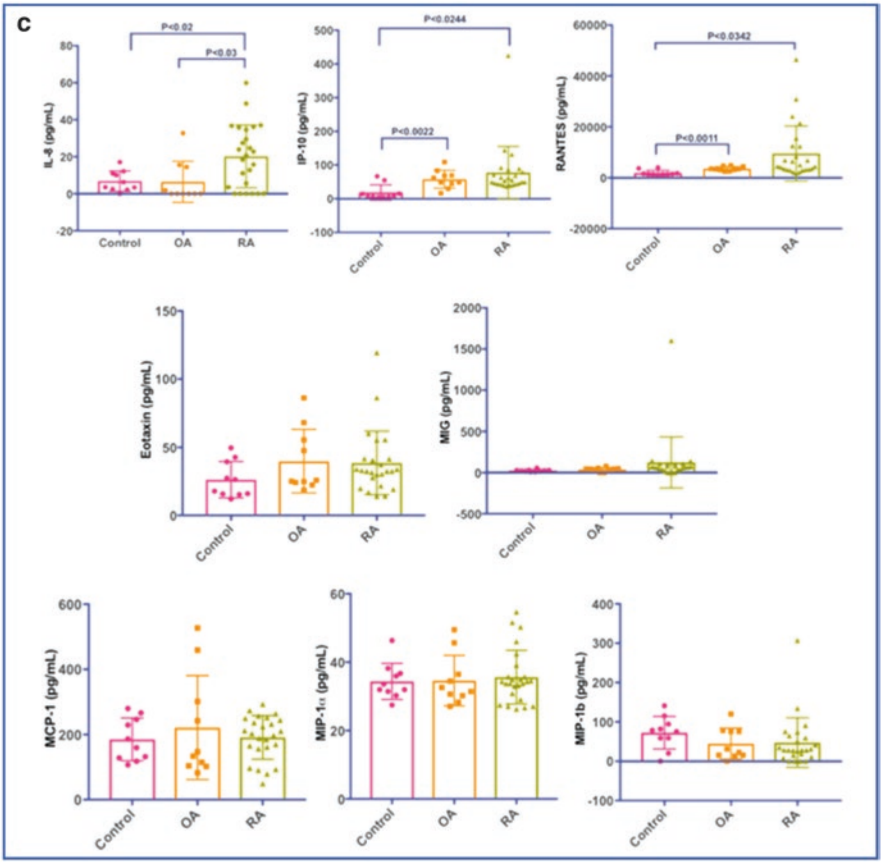


Fig. 9.7 (continued)

development of personalized therapies. The differentially regulated pathways or signaling maps are usually obtained from manual literature search, automated text mining algorithms, or canonical pathway databases (Alexopoulos et al. 2010; Wang et al. 2015) and could be used in combination with gene or miRNA expression or mass spectrometry data to deduce pathways specific to cell types or diseases (Alexopoulos et al. 2010). The gene or pathway enrichment analyses are mostly done by the Database for Annotation, Visualization, and Integrated Discovery (DAVID), Ingenuity Pathway Analysis (IPA), Pathway Studio, Reactome, Kyoto Encyclopedia of Genes and Genomes (KEGG), STRING, Path Visio, etc.(Pushparaj 2019). Therefore, the differentially regulated cytokines, chemokines, and growth factors identified through xMAP immunoassays can be analyzed using free online databases such as DAVID for functional annotation and pathway enrichment analysis or using commercially available softwares such as IPA and Pathway Studio to get more insights on the role(s) of these soluble mediators in health and disease.

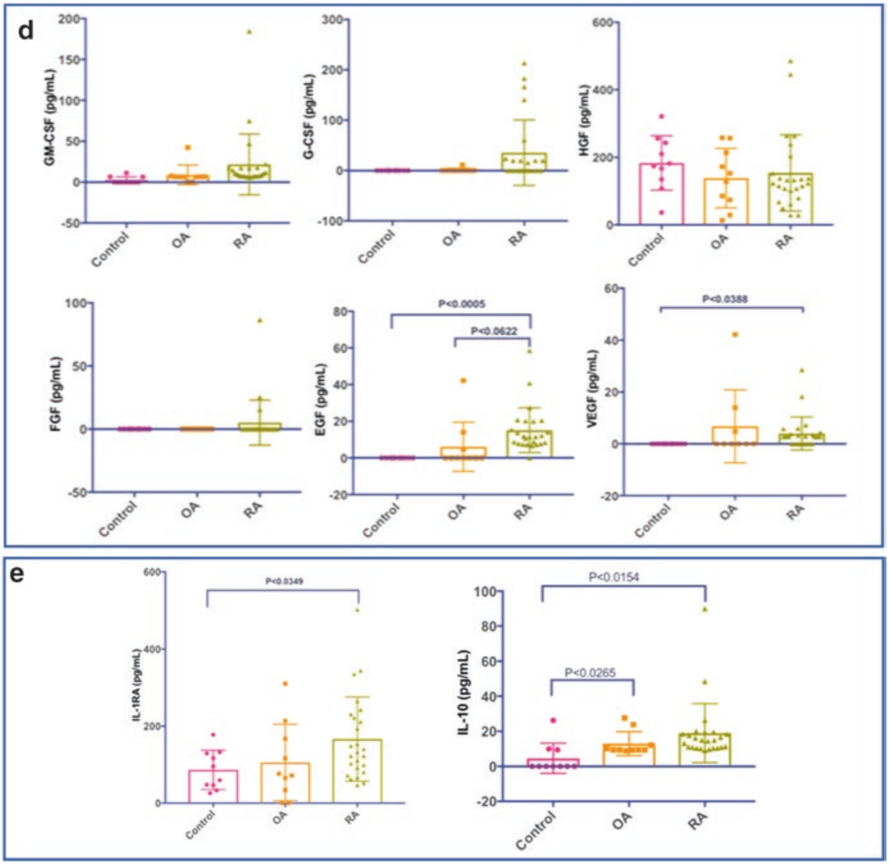


Fig. 9.8 The levels of (**d**) growth factors, (**e**) anti-inflammatory cytokines, and (**f**) Th17 and other cytokines in the plasma of RA patients with active disease, OA patients, and normal controls. The plasma concentrations (pg/mL) of all the analytes are expressed as mean ± SD. $P < 0.05$ was considered to be statistically significant (Bahlas et al. 2019)

Besides, heatmap and hierarchical cluster analysis of differentially regulated cytokines, chemokines, and growth factors derived from xMAP immunoassays can be performed using Genesis Software (Fig. 9.9) (Quackenbush 2002; Pushparaj 2019).

9.7 Conclusions

xMAP technology is a flexible and open multiplexing platform used in academia and industry to develop assays for both gene and protein expression. Contrary to conventional technologies, xMAP technology can easily be scaled up or down the number of analytes or biomarkers studied and to customize wide variety of

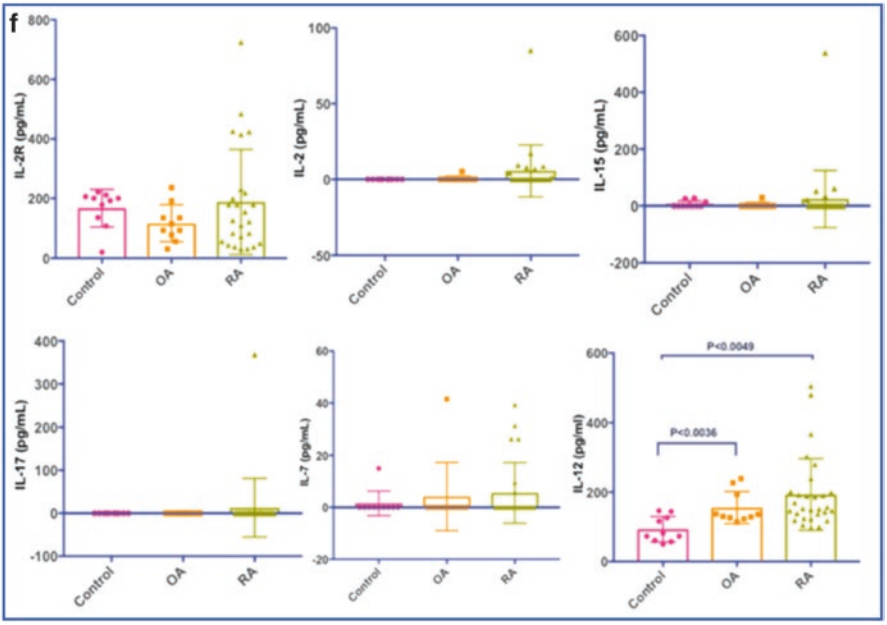


Fig. 9.8 (continued)

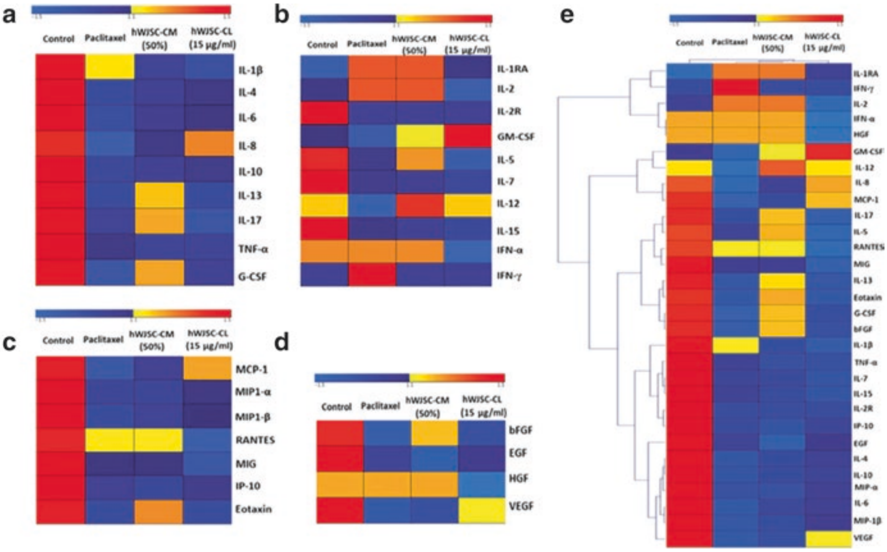


Fig. 9.9 Heatmap and hierarchical cluster analysis using Genesis Software of the cytokines, chemokines, and growth factors in the cell culture supernatant of OVCAR3 cells following treatment with human Wharton's jelly stem cell (hWJSC) extracts such as the conditioned medium (hWJSC-CM) (50%), cell lysate (hWJSC-CL) (10 μg/mL), and paclitaxel (5 nM) for 48 h and analyzed by the 30plex xMAP assay using MAGPIX. Heatmap of the differentially regulated (**a**) proinflammatory cytokines, (**b**) anti-inflammatory cytokines, (**c**) chemokines, and (**d**) growth factors. (**e**) Hierarchical clustering of the differentially regulated cytokines, chemokines, and growth factors in the treatment groups compared to the control (Kalamegam et al. 2019), https://doi.org/10.3892/ol.2019.10094)

cost-effective bioassays. xMAP technology uses cutting-edge fluidics, optics, and digital signal processing combined with patented microsphere technology. The multiplexing of 1 to 500 analytes can be performed rapidly with precision in a single sample with less sample volume which is suitable for wide variety of applications such as biomarker discovery and validation, vaccine development, mapping signaling networks, transplant medicine and HLA testing, pathogen detection, etc. Besides, an ever-increasing menu of xMAP assays for other applications is available from the Luminex Corporation, USA, and its commercial partners (Angeloni et al. 2014; Graham et al. 2019). More importantly, the differentially regulated analytes evaluated by xMAP assays can be further subjected to functional bioinformatics analysis using both open source and commercially available software to decipher cellular and molecular signaling networks and the identification disease and drug-specific signatures for the development of personalized medicine.

Acknowledgments This work is funded by the National Plan for Science, Technology, and Innovation (MAARIFAH) – King Abdulaziz City for Science and Technology – the Kingdom of Saudi Arabia-award number 12-BIO2719-03. The authors also acknowledge with thanks the Science and Technology Unit (STU), King Abdulaziz University for their excellent technical support.

References

Alexopoulos LG, Melas IN, Chairakaki AD, Saez-Rodriguez J, Mitsos A (2010) Construction of signaling pathways and identification of drug effects on the liver cancer cell HepG2. Conf Proc IEEE Eng Med Biol Soc 2010:6717–6720

Angeloni S, Cordes R, Dunbar S, Garcia C, Gibson G, Martin C et al (2014) xMAP cookbook: a collection of methods and protocols for developing multiplex assays with xMAP technology, 2nd edn. Luminex, Austin

Babady NE, Mead P, Stiles J, Brennan C, Li H, Shuptar S, Stratton CW, Tang YW, Kamboj M (2012) Comparison of the Luminex xTAG RVP fast assay and the Idaho technology FilmArray RP assay for detection of respiratory viruses in pediatric patients at a cancer hospital. J Clin Microbiol 50(7):2282–2288

Bahlas S, Damiati L, Dandachi N, Sait H, Alsefri M, Pushparaj PN (2019) Rapid immunoprofiling of cytokines, chemokines and growth factors in patients with active rheumatoid arthritis using Luminex multiple Analyte profiling technology for precision medicine. Clin Exp Rheumatol 37(1):112–119

Bjerre M, Hansen TK, Flyvbjerg A, Tønnesen E (2009) Simultaneous detection of porcine cytokines by multiplex analysis: development of magnetic bioplex assay. Vet Immunol Immunopathol 130:53–58

Bokken GC, Bergwerff AA, van Knapen F (2012) A novel bead-based assay to detect specific antibody responses against toxoplasma gondii and Trichinella spiralis simultaneously in sera of experimentally infected swine. BMC Vet Res 8:36

Breen EJ (2017) Protein multiplexed immunoassay analysis with R. Methods Mol Biol 1619:495–537

Darmanis S, Cui T, Drobin K, Li SC, Öberg K, Nilsson P, Schwenk JM, Giandomenico V (2013) Identification of candidate serum proteins for classifying well-differentiated small intestinal neuroendocrine tumors. PLoS One 8(11):e81712

Dunbar SA, Hoffmeyer MR (2013) Microsphere-based multiplex immunoassays: development and applications using Luminex® xMAP® technology. In: Wild D (ed) The immunoassay handbook. Elsevier Science & Technology, Oxford, UK, pp 157–174

Dunbar S, Li D (2010) Introduction to Luminex® xMAP® technology and applications for biological analysis in China. Asia Pac Biotech 14:26–30

Firestein GS, McInnes IB (2017) Immunopathogenesis of rheumatoid arthritis. Immunity 46(2):183–196

Graham H, Chandler DJ, Dunbar SA (2019) The genesis and evolution of bead-based multiplexing. Methods 158:2–11

Houser B (2012) Bio-Rad's Bio-Plex (R) suspension array system, xMAP technology overview. Arch Physiol Biochem 118:192–196

Kalamegam G, Sait KHW, Anfinan N, Kadam R, Ahmed F, Rasool M, Naseer MI, Pushparaj PN, Al-Qahtani M (2019) Cytokines secreted by human Wharton's jelly stem cells inhibit the proliferation of ovarian cancer (OVCAR3) cells in vitro. Oncol Lett 17(5):4521–4531

Kellar KL, Iannone MA (2002) Multiplexed microsphere-based flow cytometric assays. Exp Hematol 30(11):1227–1237

Kellar KL, Mahmutovic AJ, Bandyopadhyay K (2006) Multiplexed microsphere-based flow cytometric immunoassays. Curr Protoc Cytom; Chapter 13:Unit13.1.

Lin A, Salvador A, Carter JM (2015) Multiplexed microsphere suspension array-based immunoassays. Methods Mol Biol 1318:107–118

Manglani M, Rua R, Hendricksen A, Braunschweig D, Gao Q, Tan W, Houser B, McGavern DB, Oh K (2019) Method to quantify cytokines and chemokines in mouse brain tissue using Bio-Plex multiplex immunoassays. Methods 158:22–26

McInnes IB, Schett G (2017) Pathogenetic insights from the treatment of rheumatoid arthritis. Lancet 389(10086):2328–2337

McInnes IB, Buckley CD, Isaacs JD (2016) Cytokines in rheumatoid arthritis – shaping the immunological landscape. Nat Rev Rheumatol 12(1):63–68

Pushparaj PN (2019) Introduction to functional bioinformatics. In: Shaik NA, Hakeem KR, Banaganapalli B, Elango R (eds) Essentials of bioinformatics volume I. understanding bioinformatics: genes to proteins. Springer International Publishing, Switzerland, pp 235–254

Quackenbush J (2002) Microarray data normalization and transformation. Nat Genet 32(Suppl):496–501

Reslova N, Michna V, Kasny M, Mikel P, Kralik P (2017) xMAP technology: applications in detection of pathogens. Front Microbiol 8:55

Siebert S, Tsoukas A, Robertson J, McInnes I (2015) Cytokines as therapeutic targets in rheumatoid arthritis and other inflammatory diseases. Pharmacol Rev 67(2):280–309

Smolen JS, Aletaha D, McInnes IB (2016) Rheumatoid arthritis. Lancet 388(10055):2023–2038

Tang Y, Stratton C (2006) Advanced techniques in diagnostic microbiology. Springer, Berlin

Wang J, Zuo Y, Man YG, Avital I, Stojadinovic A, Liu M, Yang X, Varghese RS, Tadesse MG, Ressom HW (2015) Pathway and network approaches for identification of cancer signature markers from omics data. J Cancer 6(1):54–65

Chapter 10
Design and Development of Small Molecules from Somatic, Stem Cell Reprogramming, and Therapy

Praveen Kumar Guttula and Mukesh Kumar Gupta

Contents

Abbreviations

Germ line stem cells	GSCs
Glial cell line-derived neurotrophic factor	GDNF
Induced pluripotent stem cells	(iPS) cells
Multipotent germ line stem cells	mGSs
Spermatogonial stem cells	SSCs

P. K. Guttula · M. K. Gupta (✉)
Gene Manipulation Laboratory, Department of Biotechnology and Medical Engineering,
National Institute of Technology, Rourkela, India
e-mail: guptam@nitrkl.ac.in

N. A. Shaik et al. (eds.), *Essentials of Bioinformatics, Volume II*,
https://doi.org/10.1007/978-3-030-18375-2_10

10.1 Introduction

Nuclear reprogramming offers unique opportunity to modify the fate of the somatic cells and stem cells to obtain the pluripotent stem cells. These reprogrammed cells can be directed to specific cell lineage, by targeted differentiation, for their application in cell-based therapy and tissue engineering (Yamanaka and Blau 2010). Since the first report of somatic cell reprogramming by somatic cell nuclear transfer (SCNT) in the year 1997 (Schnieke et al. 1997), the technologies for nuclear reprogramming have grown exponentially and have reached a stage wherein somatic cells can be directly reprogrammed by introduction of pluripotent Oct4, Sox2, Nanog, and Lin28 (OSNL) or Oct4, Sox2, Klf4, and c-Myc (OSKM) factors to produce induced pluripotent stem (iPS) cells (Takahashi and Yamanaka 2006). More recently, small molecules such as epigenetic modifiers (e.g., valproic acid, TSA, 5-Aza-C, BIX, RG108, etc.) and modulators of cell signaling (e.g., pluripotin, reversine, PD0325901, kenpaullone, BIM, BayK, etc.) have been identified, which can either alter the fate of the somatic cells or, at least, enhance the efficiency of nuclear reprogramming (Li and Ding 2010). The later approach has drawn significant industrial attention as it paved the way to chemically synthesize newer molecules for generating pluripotent stem cells by nuclear reprogramming. However, a clear molecular mechanism of nuclear reprogramming remains poorly elusive. Furthermore, pluripotent cells, including iPS cells, generated through pluripotent factors and/or small molecular reprogramming are known to show epigenetic errors that deter their clinical application. Here, mainly we are focusing on spermatogonial stem cells because very less amount of work has been done and reprogramming is still elusive. Spermatogonial stem cells (SSCs) can divide themselves and result in an immense number of promising progenitors which were intended to differentiate into spermatozoa throughout the life span (Kubota et al. 2004; Hess et al. 2006) and be used for the treatment of male infertility. Due to the capacity of unlimited self-renewal, these cells are studied in long-term maintenance of SSC in culture condition in the laboratory for various applications like tissue engineering and transgenesis (Honaramooz et al. 2008; Hamra et al. 2005). SSCs were cultured with unknown media composition with addition of growth factors, such as glial cell line-derived neurotrophic factor (GDNF) (de Rooij 2006; Kanatsu-Shinohara et al. 2004). SSCs were also called as germ line stem cells (GSCs) which can be reprogrammed into multipotent germ line stem cells (mGSs). In spite of their spermatogonial origin, mGS cells proliferate without GDNF and produce teratomas in seminiferous tubules (Kanatsu-Shinohara et al. 2004). The absence of GDNF affects the growth characteristics of mGS cells (Zechner et al. 2009), so directly or indirectly GDNFs play a vital role in the formation mGS cells. Glial cell line-derived neurotrophic factor (GDNF) is a key player in restoration and regeneration of the damaged neurons. It has the ability to improve the terminals, the sprouting ends of dopamine neurons, where the dopamine brain cells are those pivotal cells lost in people with Parkinson's disease leading to the stiffness, slowness, and tremor (Lin et al. 1993; Hoffer et al. 1994; Beck et al. 1995; Bowenkamp et al. 1995; Hudson et al. 1995; Sauer et al. 1995;

Gash et al. 1996; Hebert et al. 1996). Most of the community of preclinical laboratory-based scientists recognized that growth factors probably are the most likely candidates to be used as first-line therapy to slow the progression of Parkinson's disease. GDNF is considered to be an important therapeutic target for various neurological disorders like Parkinson's disease (Kordower et al. 2000; Bensadoun et al. 2000), and it is also known to be essential for proliferation and self-renewal of spermatogonial stem cells (SSCs) in testes. Sertoli cells, its receptors, and brain cells secrete the GDNF; this binds to GFR alpha 1 expressed in undifferentiated spermatogonia in testes (Meng et al. 2000; Jung et al. 2010). Apart from these, an agonist *N*^{4}-{7-chloro-2-[(E)-2-(2-chloro-phenyl)-vinyl]-quinolin-4-yl}-*N*^{1},*N*^{1}-diethyl-pentane-1,4-diamine (XIB4035) which mimics the effect of GDNF in neuro-2A cells was found from reported literature, and it was used as a model to study the related functions mediated through GFR alpha 1 protein (Tokugawa et al. 2003). In this current study, GFR alpha 1 from a position 145–425 of the mouse (*Mus musculus*) was retrieved and modeled using a template in SwissModel server. Then the model was docked with non-peptidyl small molecule XIB4035. The generated model was subjected to structure-based pharmacophore modeling. The pharmacophore features were identified and saved as query file for virtual screening. After screening, it gives some hits which have similar pharmacophoric features. From generated data, the hits were docked with the GFR alpha 1 protein to identify the novel agonist molecule.

10.2 Methods

10.2.1 Sequence Retrieval, Homology Modeling, and Structure Analysis

The binding domain (145–425) of GFR alpha 1 protein sequence of mouse (*Mus musculus*) with accession number P97785 was retrieved from UniProtKB (Arnold et al. 2006) (http://www.uniprot.org/). The protein sequence was submitted to automated model building server SwissModel (Cochrane and Galperin 2009) (https://swissmodel.expasy.org/interactive). The various physicochemical properties of the protein were studied by using ProtParam (Gasteiger et al. 2005) (http://web.expasy.org/protparam/).

10.2.2 Model Validation

Functional analysis and validation of the generated model was predicted using ProFunc (Laskowski et al. 2005) web server (http://www.ebi.ac.uk/thornton-srv/databases/profunc/). Ramachandran plot using RAMPAGE was studied (http://mordred.bioc.cam.ac.uk/~rapper/rampage.php) online server (Lovell et al. 2003).

10.2.3 Molecular Docking

The two-dimensional (2D) structure of *N*^{4}-{7-chloro-2-[(E)-2-(2-chloro--phenyl)-vinyl]-quinolin-4-yl}-*N*^{1}, *N*^{1}-diethyl-pentane-1, 4-diamine (XIB4035) (Fig. 10.1) was retrieved from PubChem (Kim et al. 2015) database (https://pubchem.ncbi.nlm.nih.gov/). Then, the retrieved ligand was docked with the receptor of three-dimensional (3D) structure of GFR alpha 1 protein using PATCHDOCK (Duhovny et al. 2002; Schneidman-Duhovny et al. 2005) (http://bioinfo3d.cs.tau.ac.il/PatchDock/) online web server which accesses the surface flexibility. The best ten results will be submitted to the FireDock (Andrusier et al. 2007; Mashiach et al. 2008) (http://bioinfo3d.cs.tau.ac.il/FireDock/) for the refinement. Interactions of the best complex with less global energy were analyzed using LIGPLOT$^+$ (Laskowski and Swindells 2011; Wallace et al. 1995).

10.2.4 Pharmacophore Modeling and Structure-Based Pharmacophore Modeling

Pharmacophore modeling is a type of modeling in which the necessary features of a molecule are identified; this is also crucial for the molecular ligand recognition by a biological macromolecule. In pharmacophore modeling, training set molecule consider pharmacophore features such as the hydrogen-bond acceptor (HBA), the hydrogen-bond donor (HBD), ring aromatic (RA), hydrophobic (HY), positive ionizable (PI), negative ionizable (NI). In structure-based pharmacophore modeling, the pharmacophore models were generated from the receptor binding site. The pharmacophore features of GFR alpha 1 protein were identified using biophore feature analysis in BioPredicta module of VLifeMDS (VLife 2008).

10.2.5 Database Preparation and Virtual Screening and Docking Analysis

The molecules dataset in sdf format were retrieved from DrugBank. After retrieving the dataset by using VLife engine module, all the SDF molecules dataset were imported and converted from 2D to 3D mol2 format. The first 100 molecules were taken as a database for virtual screening. MolSign module was used for the virtual screening of database to screen the novel lead molecules having same pharmacophoric features present in reference molecule. Batch grip-based docking analysis of identified novel lead molecules was done using BioPredicta module in VLifeMDS (VLife 2008) (Fig. 10.2).

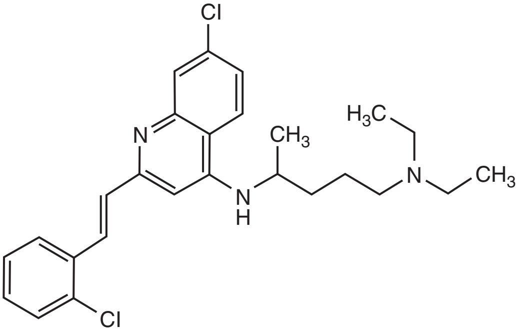


Fig. 10.1 Two-dimensional (2d) structure of XIB4035

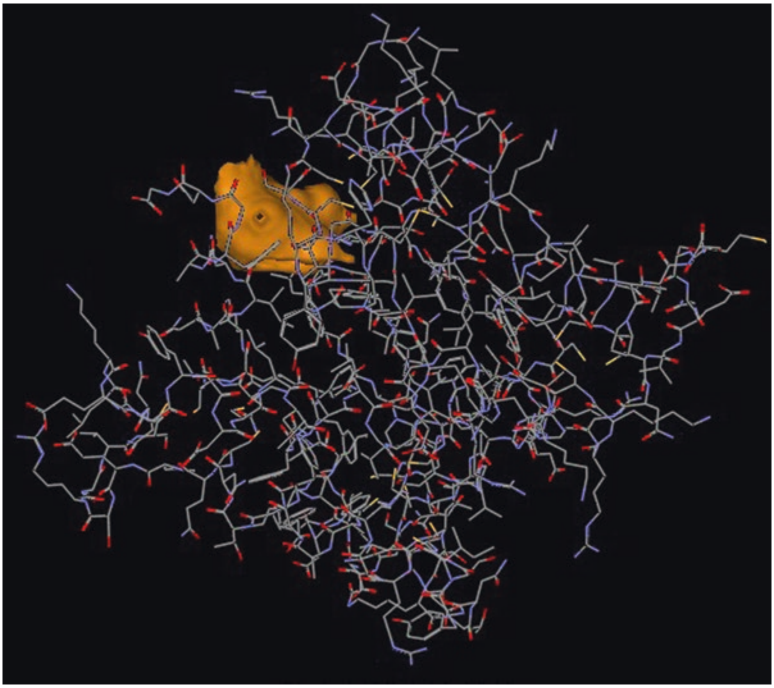


Fig. 10.2 Predetermined cavity of pharmacophore model of protein GFR alpha 1

10.3 Results and Discussion

10.3.1 Modeling and Structure Analysis

The retrieved protein sequence was modeled by SwissModel with template 3fub.2.A to build a model (Fig. 10.3). The model was built by the SwissModel (Cochrane and Galperin 2009) server homology modeling pipeline for the top-ranking templates

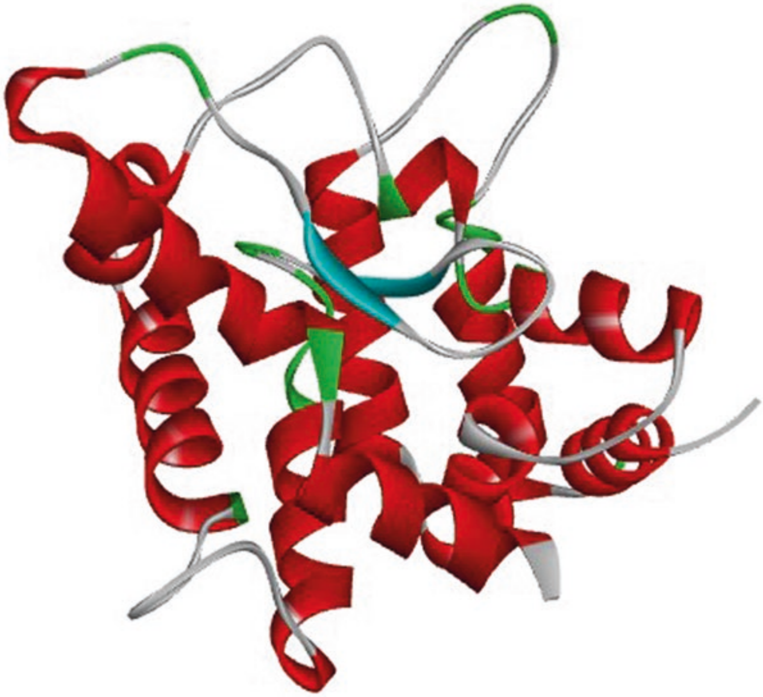


Fig. 10.3 Homology modeling of GFR alpha1 domain region by SwissModel, viewed in discovery studio 3.5v

using ProMod3. The GFR alpha 1 protein was may be unstable due to high instability index (II) (53.30) which is greater than 40. The aliphatic index (AI) was 64.23, which indicates the increase of thermostability of the proteins. The grand average of hydrophobicity (GRAVY) index of GFR alpha 1 was found to be −0.426, which indicated the interaction of water molecule (Fig. 10.4).

10.3.2 Model Validation

Ramachandran Plot analysis using RAMPAGE (Lovell et al. 2003) showed that the GFR alpha 1 model had 95.4% residues in most favored region {phi (ϕ) and psi (φ)} angles, which helped to know that the generated model was a good model (Fig. 10.5) (Table 10.1).

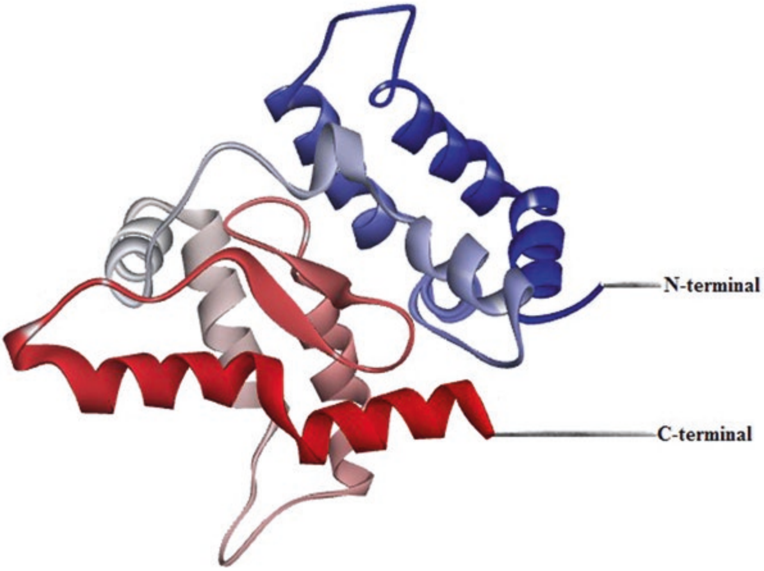


Fig. 10.4 Homology modeling of GFR alpha 1 domain region shown in solid ribbon, n-terminal (in blue), and c-terminal (in red), viewed in discovery studio 3.5v

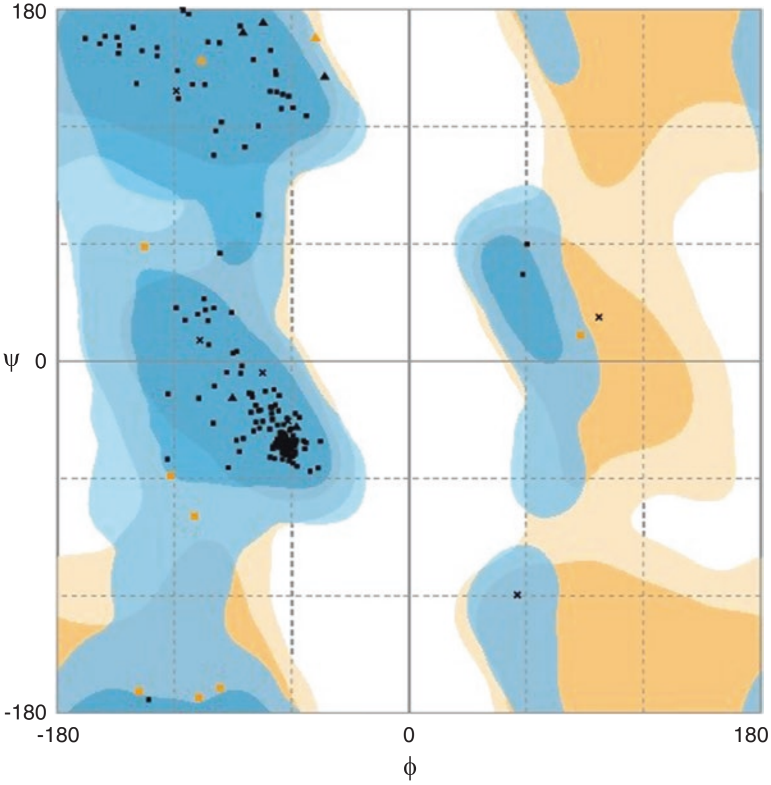


Fig. 10.5 Ramachandran plot of modeled GFR alpha 1 using RAMPAGE

Table 10.1 Ramachandran plot analysis with parameters

Ramachandran plot analysis parameters	No of residues (in %)
No of residues and percentage in most favored regions	188 (95.4%)
No of residues and percentage in additionally allowed regions	9 (4.6%)
No of residues and percentage in disallowed region	0 (0.0%)

10.3.3 Three-Dimensional Architecture of GFR Alpha 1

The GFR alpha 1 comprises 113 α-helix (56.8% amino acid), 4 β-sheets (2.0%), 6 strands (3.0%), and 76 (38.2%) other secondary elements. Further, the structure also contains 1 β-sheet, 1 β-hairpin, 2 strands, 13 helices, 23 helix-helix interaction, 7 β turns, 1 ϒ turn, and 10 disulfides.

10.3.4 Molecular Docking

Docking studies are very crucial for visualizing the interaction between the ligand and receptor. Docking studies were done primarily using PATCHDOCK (Duhovny et al. 2002; Schneidman-Duhovny et al. 2005), which accesses the surface flexibility addresses by intermolecular penetration. Docking between the ligand and the generated protein model, which obtained the ligand bound to the specific binding site of the protein to show as an agonist on GDNF receptor, and induce signal transduction mechanism through GFR alpha 1 in mouse cells. The best dock model was retrieved from FireDock (Andrusier et al. 2007; Mashiach et al. 2008) with low global energy (−40.73 Kcal/mol) shown in Fig. 10.6. To know the interaction, recognition site selected the GFR alpha 1-XIB4035 complex in LIGPLOT^{+} (Laskowski and Swindells 2011; Wallace et al. 1995), which shows all the hydrogen bonds and hydrophobic interactions between receptor and ligand. It shows one hydrogen bond between receptor (Arg 27) and ligand in the distance (2.83) (Figs. 10.7, 10.8, and 10.9).

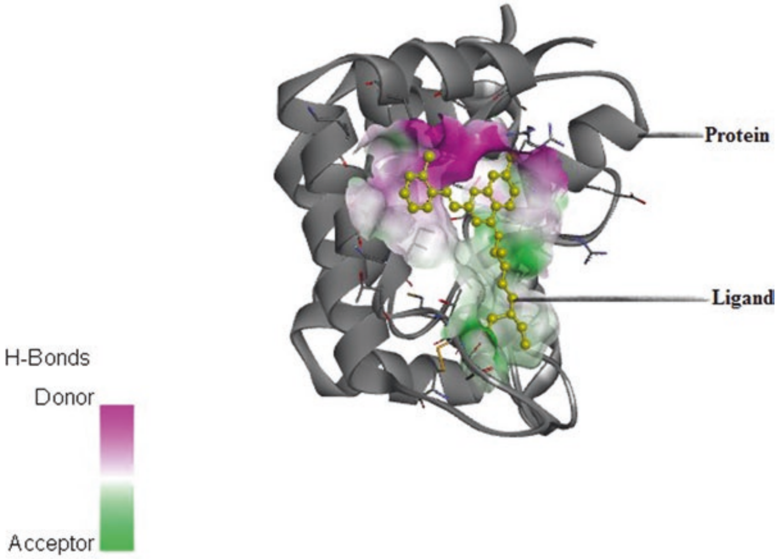


Fig. 10.6 A 3d model of GFR alpha1 binding with XIB4035

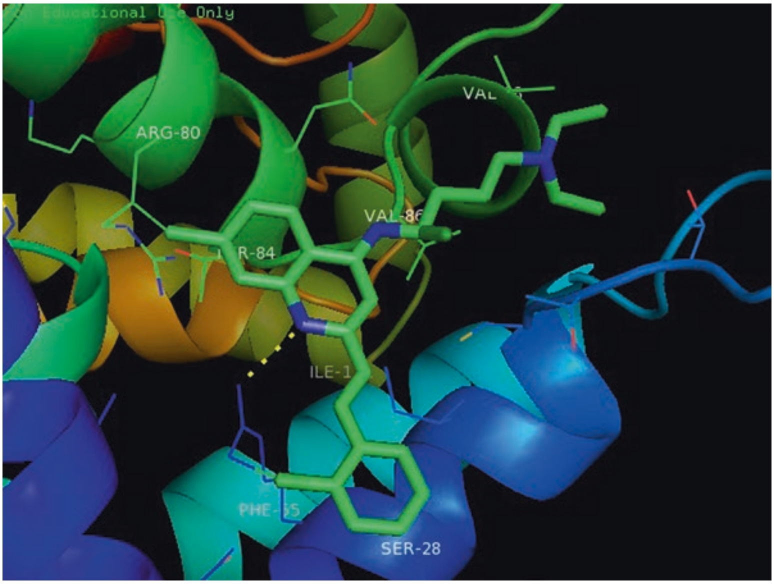


Fig. 10.7 A 3d model of GFR alpha1 binding with XIB4035 in PyMol

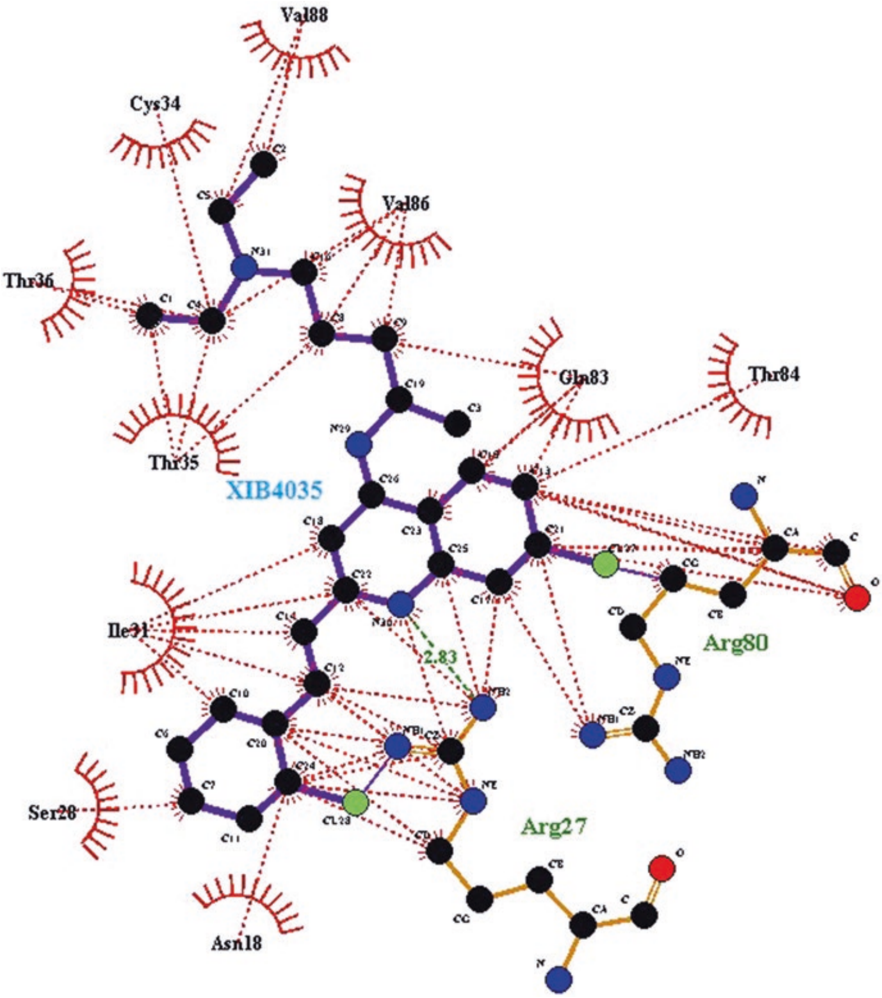


Fig. 10.8 A 2d model of GFR alpha1 interacting with XIB4035 showing hydrogen bond and hydrophobic interactions

10.3.5 Structure-Based Pharmacophore (SBP) Modeling

The modeled protein GFR alpha 1 was subjected to SBP to identify different pharmacophore features; four query pharmacophoric features (three hydrogen-bond donors, one aliphatic group) were generated. Six common amino acid residues (CYS 70A, SER 71A, CYS 72A, GLN 199A, GLY 202A, ASN 203A) were present near to the pharmacophore features. The abovementioned four pharmacophoric features were saved as a query file for virtual screening, as shown in Fig. 10.10.

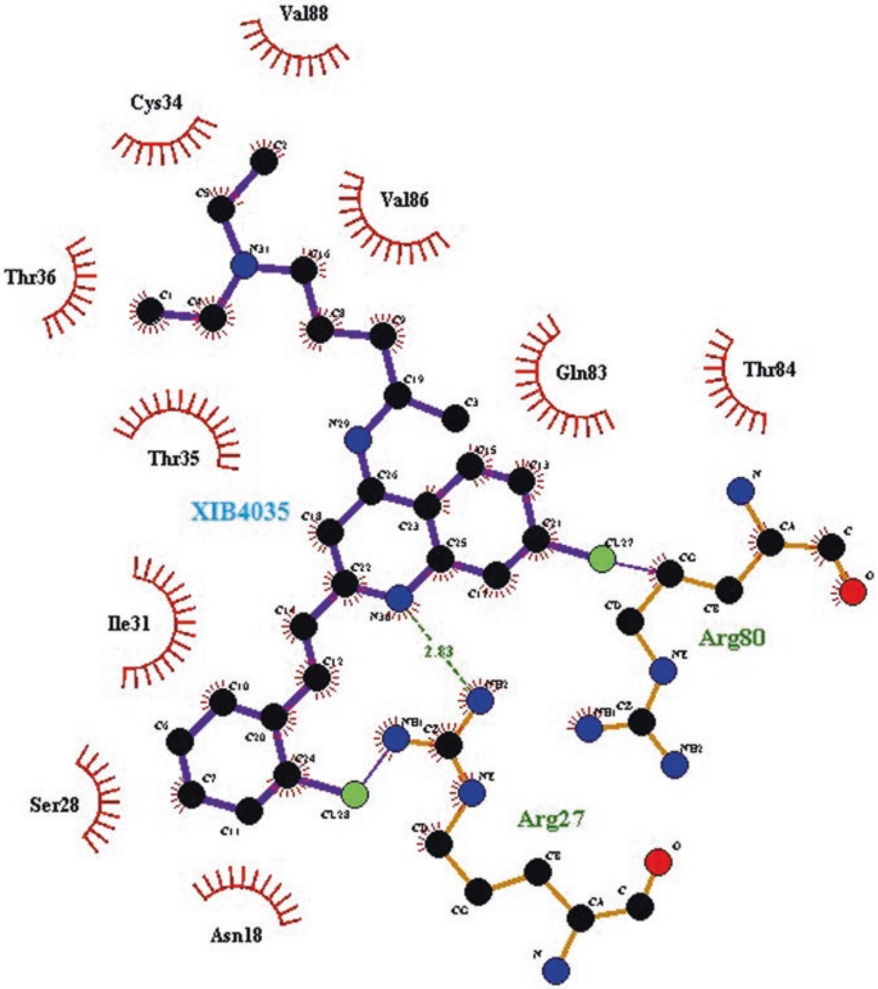


Fig. 10.9 A 2d model of GFR alpha1 bind with XIB4035, showing covalent bond interactions

10.3.6 *Virtual Screening and Docking Analysis*

The query file which was obtained from SBP was screened against the prepared database to identify the novel molecules which may have agonist activity. Thirty-nine hits were identified having three common pharmacophore features which include three hydrogen bond donors shown in Table 10.2. The distance between the common identified pharmacophoric features in 39 hit molecules is shown in Fig. 10.11. GRIP-based batch docking analysis of 39 novel molecules using BioPredicta module in VLifeMDS (VLife 2008) reveals that structure eight p10 molecule out of 39 molecules shows the best conformations with dock score (−151.883564 kcal/mol) and shows two hydrogen bonds between receptor and ligand (Figs. 10.12 and 10.13).

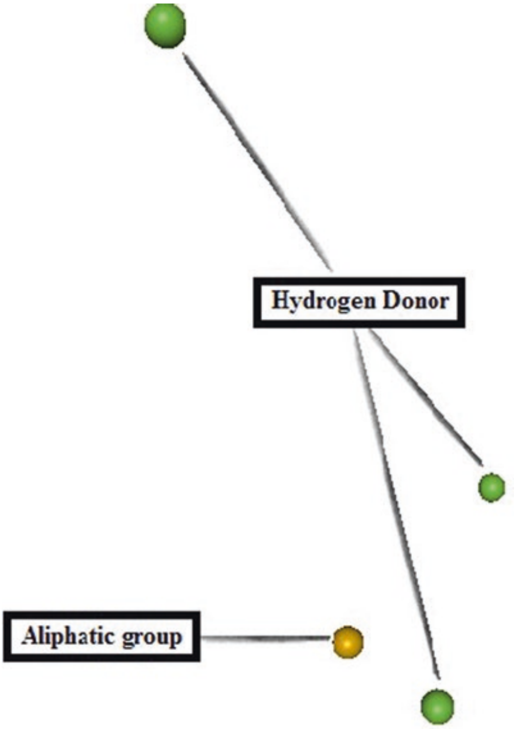


Fig. 10.10 Pharmacophore features of GFR alpha 1

Table 10.2 Showing different pharmacophore features

Biophore features	Number of features	Aromatic AroC	Aliphatic	Positive	Negative	HAc	HDr	Mols
HDr, HDr, HDr	3	0	0	0	0	0	3	39
HAc, HDr, HDr	3	0	0	0	0	1	2	38
HAc, HDr, HDr	3	0	0	0	0	1	2	13
HDr, HDr, AroC	3	1	0	0	0	0	2	12
HAC, HDr, HDr	3	0	0	0	0	1	2	10
HAC, HDr, HDr, HDr	4	0	0	0	0	1	3	3
HDr, HDr, AroC	3	1	0	0	0	0	2	1

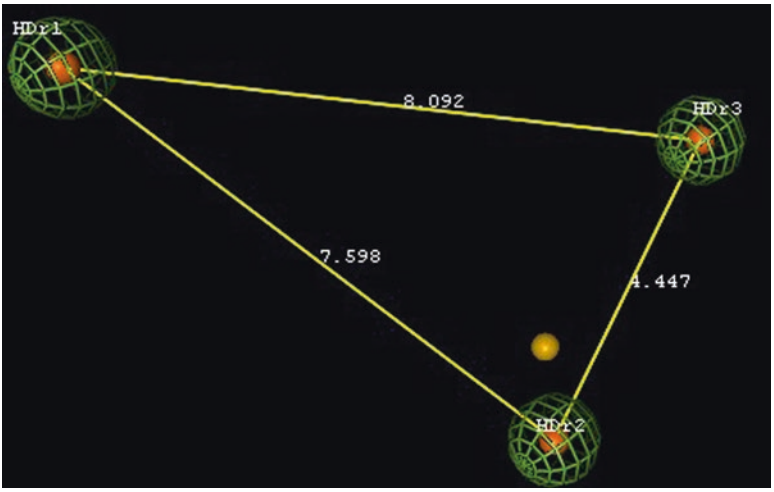


Fig. 10.11 Pharmacophore features of GFR alpha 1 with distances

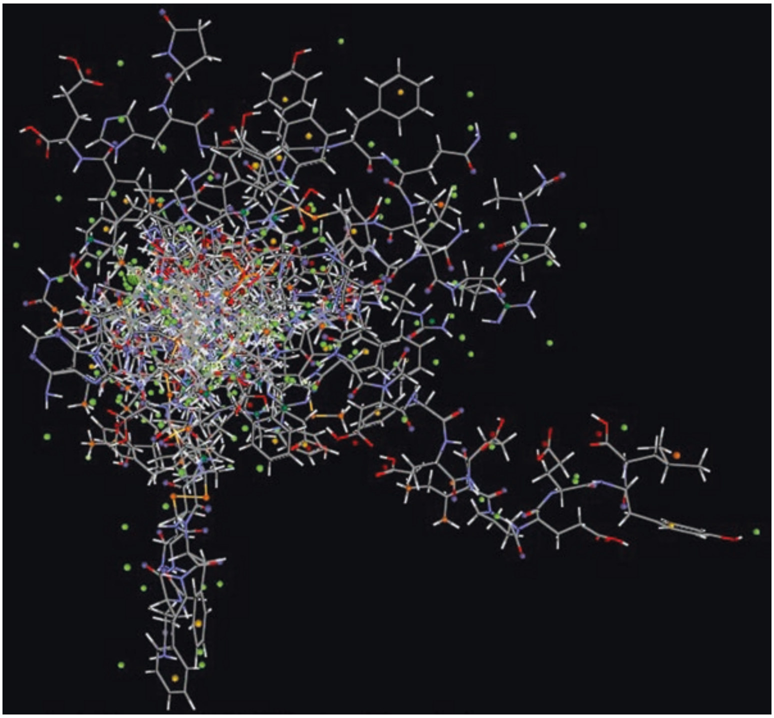


Fig. 10.12 Alignment of 39 molecules which have same pharmacophore features shown in VLife 4.6$_v$

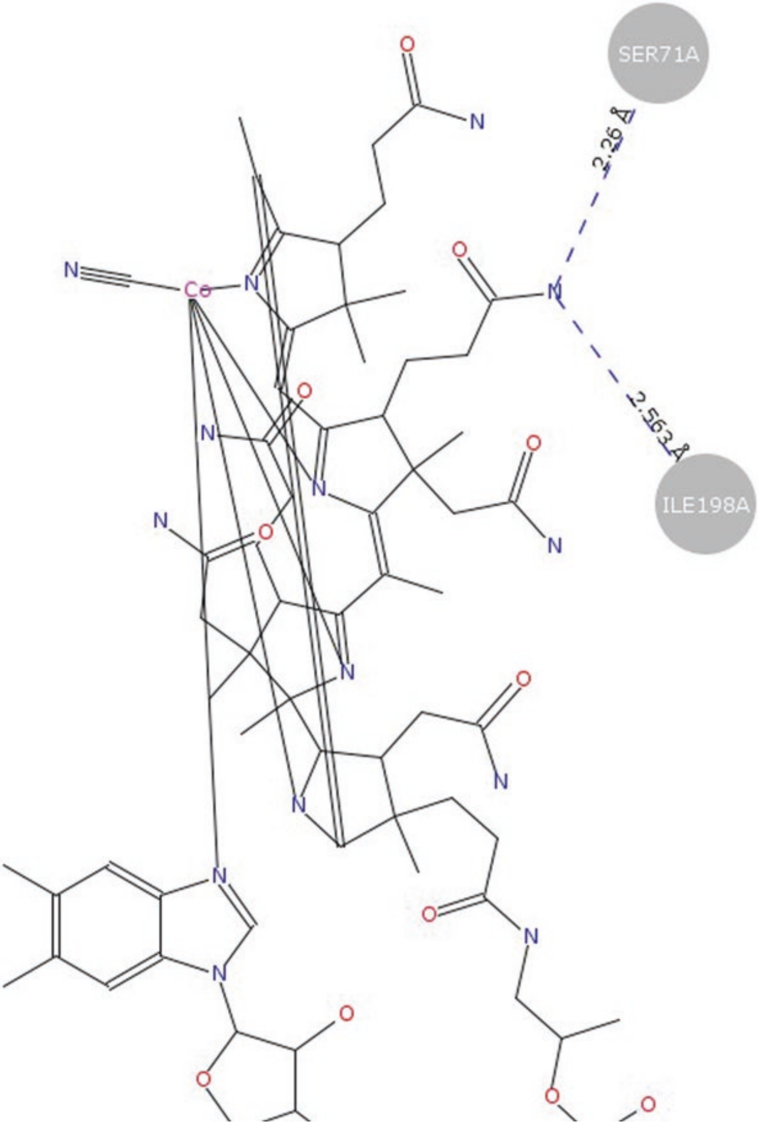


Fig. 10.13 Interaction of structure 8_p 10 with GFR alpha 1

10.4 Conclusion

GDNF and its receptors GFR alpha 1 are expressed in undifferentiated spermatogonia in testis. An agonist was found from reported literature XIB4035 (N^4-{7-chloro-2-[(E)-2-(2-chloro-phenyl)-vinyl]-quinolin-4-yl}-N^1,N^1-diethyl-pentane-1,4-diamine). By using structural biology tools, the gfralpha1 protein was modeled by using SwissModel. The docking analysis of GFR alpha 1with XIB4035 revealed that it has the strong binding affinity as there is one hydrogen bond between ligand XIB4035 with the residues ARG27 of the GFR alpha 1 protein. The structural biology tools make it easier for the determination of pharmacophore modeling of protein GFR alpha 1. Structure-based pharmacophore modeling identified features like three hydrogen bonds and one aliphatic group. The pharmacophore model was screened against DrugBank database for virtual screening. From the virtual screening, 39 hit molecules were identified and were again docked by VLifeMDS. One novel molecule having docking score of −151.883564 Kcal/mol was identified. From the earlier study, we can tell that identified novel molecule may have a similar effect like GDNF in reprogramming of spermatogonial stem cells and may also be used as therapeutic target for Parkinson's disease.

Acknowledgments The authors are thankful to Bioinformatics Infrastructure facility at National Institute of Technology, Rourkela funded by DBT for providing the facilities to carry out the research.

References

Andrusier N, Nussinov R, Wolfson HJ (2007) FireDock: fast interaction refinement in molecular docking. Proteins: Structure, Function, and Bioinformatics 69(1):139–159

Arnold K, Bordoli L, Kopp J, Schwede T (2006) The SWISS-MODEL workspace: a web-based environment for protein structure homology modelling. Bioinformatics 22(2):195–201

Beck KD, Valverde J, Alexi T, Poulsen K, Moffat B, Vandlen RA, Rosenthal A, Hefti F (1995) Mesencephalic dopaminergic neurons protected by GDNF from axotomy-induced degeneration in the adult brain. Nature 373(6512):339

Bensadoun JC, Déglon N, Tseng JL, Ridet JL, Zurn AD, Aebischer P (2000) Lentiviral vectors as a gene delivery system in the mouse midbrain: cellular and behavioral improvements in a 6-OHDA model of Parkinson's disease using GDNF. Exp Neurol 164(1):15–24

Bowenkamp KE, Hoffman AF, Gerhardt GA, Henry MA, Biddle PT, Hoffer BJ, Granholm AC (1995) Glial cell line-derived neurotrophic factor supports survival of injured midbrain dopaminergic neurons. J Comp Neurol 355(4):479–489

Cochrane GR, Galperin MY (2009) The 2010 nucleic acids research database issue and online database collection: a community of data resources. Nucleic Acids Res 38(suppl_1):D1–D4

de Rooij DG (2006) Rapid expansion of the spermatogonial stem cell tool box. Proc Natl Acad Sci 103(21):7939–7940

Duhovny D, Nussinov R, Wolfson HJ (2002) Efficient unbound docking of rigid molecules. In: InInternational workshop on algorithms in bioinformatics. Springer, Berlin, Heidelberg, pp 185–200

Gash DM, Zhang Z, Ovadia A, Cass WA, Yi A, Simmerman L, Russell D, Martin D, Lapchak PA, Collins F, Hoffer BJ (1996) Functional recovery in parkinsonian monkeys treated with GDNF. Nature 380(6571):252

Gasteiger E, Hoogland C, Gattiker A, Wilkins MR, Appel RD, Bairoch A (2005) Protein identification and analysis tools on the ExPASy server. In: The proteomics protocols handbook. Humana press, pp 571–607. New York, USA

Hamra FK, Chapman KM, Nguyen DM, Williams-Stephens AA, Hammer RE, Garbers DL (2005) Self renewal, expansion, and transfection of rat spermatogonial stem cells in culture. Proc Natl Acad Sci U S A 102(48):17430–17435

Hebert MA, Van Horne CG, Hoffer BJ, Gerhardt GA (1996) Functional effects of GDNF in normal rat striatum: presynaptic studies using in vivo electrochemistry and microdialysis. J Pharmacol Exp Ther 279(3):1181–1190

Hess RA, Cooke PS, Hofmann MC, Murphy KM (2006) Mechanistic insights into the regulation of the spermatogonial stem cell niche. Cell Cycle 5(11):1164–1170

Hoffer BJ, Hoffman A, Bowenkamp K, Huettl P, Hudson J, Martin D, Lin LF, Gerhardt GA (1994) Glial cell line-derived neurotrophic factor reverses toxin-induced injury to midbrain dopaminergic neurons in vivo. Neurosci Lett 182(1):107–111

Honaramooz A, Megee S, Zeng W, Destrempes MM, Overton SA, Luo J, Galantino-Homer H, Modelski M, Chen F, Blash S, Melican DT (2008) Adeno-associated virus (AAV)-mediated transduction of male germ line stem cells results in transgene transmission after germ cell transplantation. FASEB J 22(2):374–382

Hudson J, Granholm AC, Gerhardt GA, Henry MA, Hoffman A, Biddle P, Leela NS, Mackerlova L, Lile JD, Collins F, Hoffer BJ (1995) Glial cell line-derived neurotrophic factor augments midbrain dopaminergic circuits in vivo. Brain Res Bull 36(5):425–432

Jung YH, Gupta MK, Oh SH, Uhm SJ, Lee HT (2010) Glial cell line-derived neurotrophic factor alters the growth characteristics and genomic imprinting of mouse multipotent adult germline stem cells. Exp Cell Res 316(5):747–761

Kanatsu-Shinohara M, Inoue K, Lee J, Yoshimoto M, Ogonuki N, Miki H, Baba S, Kato T, Kazuki Y, Toyokuni S, Toyoshima M (2004) Generation of pluripotent stem cells from neonatal mouse testis. Cell 119(7):1001–1012

Kim S, Thiessen PA, Bolton EE, Chen J, Fu G, Gindulyte A, Han L, He J, He S, Shoemaker BA, Wang J (2015) PubChem substance and compound databases. Nucleic Acids Res 44(D1):D1202–D1213

Kordower JH, Emborg ME, Bloch J, Ma SY, Chu Y, Leventhal L, McBride J, Chen EY, Palfi S, Roitberg BZ, Brown WD (2000) Neurodegeneration prevented by lentiviral vector delivery of GDNF in primate models of Parkinson's disease. Science 290(5492):767–773

Kubota H, Avarbock MR, Brinster RL (2004) Growth factors essential for self-renewal and expansion of mouse spermatogonial stem cells. Proc Natl Acad Sci U S A 101(47):16489–16494

Laskowski RA, Swindells MB (2011) LigPlot+: multiple ligand–protein interaction diagrams for drug discovery. J Chem Inf Model 51:2778

Laskowski RA, Watson JD, Thornton JM (2005) ProFunc: a server for predicting protein function from 3D structure. Nucleic Acids Res 33(suppl_2):W89–W93

Li W, Ding S (2010) Small molecules that modulate embryonic stem cell fate and somatic cell reprogramming. Trends Pharmacol Sci 31(1):36–45

Lin LF, Doherty DH, Lile JD, Bektesh S, Collins F (1993) GDNF: a glial cell line-derived neurotrophic factor for midbrain dopaminergic neurons. Science 260(5111):1130–1132

Lovell SC, Davis IW, Arendall WB, De Bakker PI, Word JM, Prisant MG, Richardson JS, Richardson DC (2003) Structure validation by Cα geometry: ϕ, ψ and Cβ deviation. Proteins: Structure, Function, and Bioinformatics. 50(3):437–450

Mashiach E, Schneidman-Duhovny D, Andrusier N, Nussinov R, Wolfson HJ (2008) FireDock: a web server for fast interaction refinement in molecular docking. Nucleic Acids Res 36(suppl_2):W229–W232

Meng X, Lindahl M, Hyvönen ME, Parvinen M, de Rooij DG, Hess MW, Raatikainen-Ahokas A, Sainio K, Rauvala H, Lakso M, Pichel JG (2000) Regulation of cell fate decision of undifferentiated spermatogonia by GDNF. Science 287(5457):1489–1493

Sauer H, Rosenblad C, Björklund A (1995) Glial cell line-derived neurotrophic factor but not transforming growth factor beta 3 prevents delayed degeneration of nigral dopaminergic neurons following striatal 6-hydroxydopamine lesion. Proc Natl Acad Sci 92(19):8935–8939

Schneidman-Duhovny D, Inbar Y, Nussinov R, Wolfson HJ (2005) PatchDock and SymmDock: servers for rigid and symmetric docking. Nucleic Acids Res 33(suppl_2):W363–W367

Schnieke AE, Kind AJ, Ritchie WA, Mycock K, Scott AR, Ritchie M, Wilmut I, Colman A, Campbell KH (1997) Human factor IX transgenic sheep produced by transfer of nuclei from transfected fetal fibroblasts. Science 278(5346):2130–2133

Takahashi K, Yamanaka S (2006) Induction of pluripotent stem cells from mouse embryonic and adult fibroblast cultures by defined factors. Cell 126(4):663–676

Tokugawa K, Yamamoto K, Nishiguchi M, Sekine T, Sakai M, Ueki T, Chaki S, Okuyama S (2003) XIB4035, a novel nonpeptidyl small molecule agonist for GFRα-1. Neurochem Int 42(1):81–86

VLife MD (2008) 3.5 Molecular design suite. VLife Sciences Technologies, Pune

Wallace AC, Laskowski RA, Thornton JM (1995) LIGPLOT: a program to generate schematic diagrams of protein-ligand interactions. Protein Eng Des Sel 8(2):127–134

Yamanaka S, Blau HM (2010) Nuclear reprogramming to a pluripotent state by three approaches. Nature 465(7299):704

Zechner U, Nolte J, Wolf M, Shirneshan K, Hajj NE, Weise D, Kaltwasser B, Zovoilis A, Haaf T, Engel W (2009) Comparative methylation profiles and telomerase biology of mouse multipotent adult germline stem cells and embryonic stem cells. Mol Hum Reprod 15(6):345–353

Chapter 11
Molecular Modeling and Drug Design Techniques in Microbial Drug Discovery

Chandrabose Selvaraj

Contents

C. Selvaraj (✉)
School of Basic Sciences, Indian Institute of Technology, Mandi,
Kamand, Himachal Pradesh, India

N. A. Shaik et al. (eds.), *Essentials of Bioinformatics, Volume II*,
https://doi.org/10.1007/978-3-030-18375-2_11

Abbreviations

AMR	Antimicrobial resistance
BLAST	Basic Local Alignment Search Tool
CADD	Computer aided drug designing
CG	Coarse-grained
CG-MD	Coarse-grained molecular dynamics
CHARMM	Chemistry at Harvard Macromolecular Mechanics
CoMFA	Comparative molecular field analysis
CoMSIA	Comparative molecular similarity indices analysis
DADA	D-alanyl-D-alanine
DFT	Density functional theory
DPD	Dissipative particle dynamics
ESBLs	Extended spectrum β-lactamases
FDA	Food and Drug Administration
FEP	Free-energy perturbation method
GA	Genetic algorithms
GISA	Glycopeptides-intermediately-resistant *S. aureus*
GPU	Graphical processor unit
HTS	High throughput screening
IUPAC	International Union of Pure and Applied Chemistry
LB	Ligand-based
LBDD	Ligand-based drug design
LBVS	Ligand-based virtual screening
MD	Molecular dynamics
MDR	Multi-drug resistance
MRSA	*Staphylococcus aureus* resistant to methicillin
NCBI	National Center for Biotechnology Information
NMR	Nuclear magnetic resonance
PCA	Principle component analysis
PCR	Polymerase chain reaction
PK	Pharmacokinetic
PLS	Partial least squares
QM/MM	Quantum mechanics/molecular mechanics
QSAR	Quantitative structure-activity relationship
SB	Structure-based
SBDD	Structure-based drug design
SBVS	Structure-based virtual screening
SI	Sequence identity
TEIC	Teicoplanin
VANC	Vancomycin

11.1 Introduction

A potential drug molecule is one that effectively binds and modulates a molecular target in such a manner that is less toxic, safe and effective in the disease context for which it is doled out. The drug discovery development is a complex process, which can take 12–15 years and entail costs of more than $1 billion. In the modern era of drug discovery, development involves the cooperation of many disciplines such as chemistry, biology, mathematics and computer science (Herrling 2005). A chemical moiety with significant therapeutic value is extensively analyzed for its safety and efficacy before it is marketed. The multistep process, termed 'drug discovery,' includes identification and validation of the drug target and of the lead molecule. The drug development process is categorized, basically, into the two major phases of drug discovery and drug development. The drug discovery process involves two important approches; identification and validation of a potential disease-oriented target molecule and another approach is phenotypic screening to identify and refine the potential small molecules that can interact with target (Ernst and Obrecht 2008). This molecular interaction can be to block, promote or modify the activity of the target. In recent years, the drug discovery process has undergone radical changes due to the entry of various novel techniques in genomics; proteomics have been developed in drug target identification and validation has become more specific (Umashankar and Gurunathan 2015). In the past decade, emergence of microbial resistance (Amini and Tavazoie 2011) and complicated new diseases and unexpected adverse side effects have accelerated the identification of potent lead molecules (Ashrafuzzaman 2014). Infectious diseases, particularly Gram-positive bacterial infection, are among the major serious threats to public health worldwide: they are difficult to treat and are associated with high morbidity and mortality rates. Gram-negative bacteria are highly adaptive pathogens that produce resistance to antibiotics through several mechanisms. The production of β-lactamases and hydrolyzation of the β-lactam ring represents the most common resistant mechanism in Gram-negative bacteria against β-lactam antibiotics. Most bacteria can develop and adapt themselves according to their surroundings and subsequently develop several protective mechanisms to reduce their susceptibility to antibiotics. In some cases, bacteria allow horizontal gene transfer within and between species to become more resistant to antibiotics (Palumbi 2001; Thomas and Nielsen 2005). This horizontal gene transfer provides the most important mechanism to accelerate the spectrum of β-lactamases (ESBLs), causing severe problems in drug resistant in the health care world (Giske et al. 2008; Hawkey and Jones 2009). Bacterial strains capable of producing ESBLs are resistant to several antibiotics, including penicillins and cephalosporins, and they are resistant to other antibacterials such as quinolones and aminoglycosides. This antibiotic resistance shows a strong correlation between the segment of the population that uses antibiotics and the prevalence of antibiotic-resistant bacteria in the same population; the correlation has been found on both national and regional levels (Bronzwaer et al. 2002; Albrich et al. 2004).

11.2 Global Battle Against Infectious Diseases

In the middle of the seventeenth century, smallpox infection was the most fatal and feared of diseases. The discovery of penicillin developed a new generation of antibiotics that cured a wide range of infectious diseases. Several researches focused on understanding what mechanisms the microbes used to survive antibiotics, and several pharmaceutical and biotech companies nearly stampeded to identify a significant bacterial target and to create novel methodologies against the bacteria. Recent evidence suggests that mutation with humans is not the only way bacteria develop antibiotic resistance; they can also transfer genetic instructions for avoiding an antibiotic to other bacterial species. In the late 1800s, pathogen-specific medical diagnosis lent a hand to the identification of microbes that caused specific diseases. Molecular genetics technique, polymerase chain reaction (PCR) and, more recently, sophisticated, high throughput rapid sequencing of the genome of the pathogen are all used to observe the individual genetic variants ,facilitating identification of the familial base of drug immunity. Other factor-based, diagnostic tools including microchip and serological techniques and enzyme-linked immunosorbent assay can be more sensitive than traditional techniques in finding and measuring antibodies to pathogens (Pallen et al. 2010). Current data suggest that Gram-positive bacteria cause 45–70% of infectious diseases and are behind the increase in rates of drug resistance in many infections. The pace of drug resistance among bacterial pathogens is increasing; virtually no new antibiotics are being developed (Spellberg et al. 2004). Gram-positive organisms such as the bacteria of the genera *Staphylococcus*, *Streptococcus* and *Enterococcus* are the predominant bacterial spp causing clinical infection, hence, recent attention has focused on the multi-drug resistance (MDR) and antimicrobial resistance (AMR) (Menichetti 2005; Doernberg et al. 2017).

Sulfonamide synthetic antimetabolites were first used clinically in 1932 for a wide range of both Gram-negative and Gram-positive bacteria. These synthetic metabolites inhibit dihydropteroate synthetase leads to repressed DNA replication. Until 1938, β-lactam was another widely used antibiotic. The 28 members that include antibiotics/β-lactamase inhibitor combinations are broadly classified into three subclasses: penicillins, cephalosporins and carbapenems, which are critically used in very broad-spectrum activity against most aerobic and anaerobic Gram-positive and Gram-negative bacteria (Walsh 2003; Collignon et al. 2009; Lewis 2013). Recently, glycopeptides like vancomycin (VANC) and teicoplanin (TEIC) have been widely used against Gram-positive bacteria; these share a mechanism of natural process similar to that of β-lactams, except their interruption on cell wall synthesis via an interaction with the D-alanyl-D-alanine (DADA) moiety of peptidoglycan precursors inhibits the cross-linking stabilization step in bacterial cell wall formation (Malabarba and Goldstein 2005). The cyclic lipopeptide daptomycin has an extensive range of activity on Gram-positive bacterial infection and also on MRSA. Structurally, daptomycin comprises a 13-member hydrophobic polypeptide with a lipophilic side chain having a unique mechanism of natural process, which is leads insertion of the lipophilic region into the bacterial cell wall, oligomerizing into pore-like constructions, through which a significant efflux of potassium ions results in rapid bacterial cell death (Silverman et al. 2003; Steenbergen et al. 2005).

11.3 Methods in Drug Design

Drug development commences with the identification of a molecular target and lead molecules followed by lead optimization and preclinical in vitro and in vivo studies to recognize potent compounds that fulfill the primary criteria for the drug development (Bleicher et al. 2003). But, the development of lead molecules through in vitro and in vivo methods takes a long time and is very expensive (DiMasi et al. 2003); hence, in recent years in silico drug designing has been widely used to predict active lead molecules. Here, we look at discovery. Traditional drug discovery (in vitro and in vivo) requires about 12–14 years and costs up to $1.2–$1.4 billion dollars to get a drug from discovery to market (Hileman 2006). About 90% of the drugs entering clinical trials fail to obtain FDA approval and reach the consumer market (Tollman 2001). Lately, high throughput screening (HTS) experiments are used to sort thousands of molecules with robotic automation; however, HTS is still expensive and requires a great amount of resources. Therefore, computer-aided drug designing (CADD) can cut cost- and time-associated drawbacks and ensure the best possible lead compounds are used in animal studies. CADD tools have not merely been applied to distinguish potential lead molecules; they can also predict effectiveness and possible side effects and aid in improving bioavailability of the possible drug molecules (Yang et al. 2016). CADD plays a crucial role in the identification of many pharmaceutically available drugs, ones that have obtained FDA approval and reached the consumer market (Kitchen et al. 2004; Clark 2006; Talele et al. 2010). CADD methods are broadly classified into two categories: structure-based (SB) drug discovery and ligand-based (LB) drug discovery.

11.3.1 Structure-Based Drug Design

Structure-based drug design (SBDD) methods are prominent tools in modern medicinal chemistry that utilize three-dimensional structural information from biological targets (Salum et al. 2008). Understanding the mechanism of small molecule reorganization and interaction with biological macromolecules is of great importance in pharmaceutical research and development. In recent years, due to wide range of application such as molecular docking, molecular dynamic simulation, and structure-based virtual screening (SBVS), SBDD has played a crucial role in the identification of potential drug molecules against various drug target (Kalyaanamoorthy and Chen 2011). In SBDD, binding site topology (including clefts, cavities and subpockets) and the electrostatic properties of the target molecule were carefully examined (Wilson and Lill 2011).

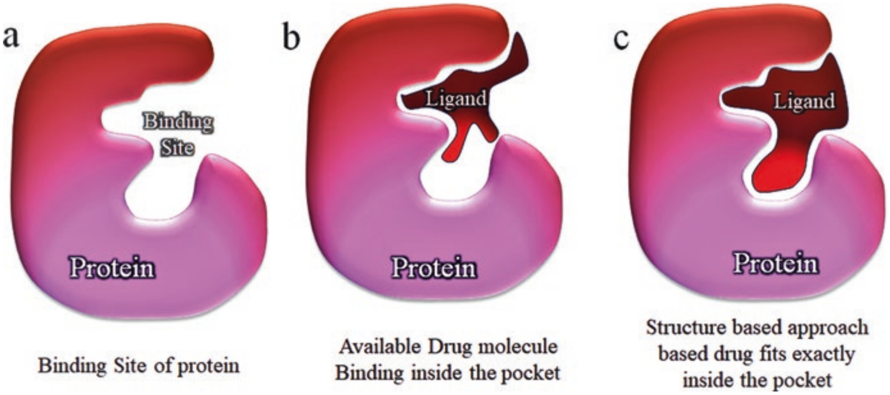

a
Binding Site
Protein
Binding Site of protein
b
Ligand
Protein
Available Drug molecule Binding inside the pocket
c
Ligand
Protein
Structure based approach based drug fits exactly inside the pocket

Fig. 11.1 Mechanism of SBDD showing the design of a new molecule as per the binding site feature of a protein

SBDD is an iterative method involving multiple steps for finding a lead. The first step of SBDD includes the cloning, purification and structure elucidation of the target proteins or nucleic acid by NMR, X-ray crystallography or homology modeling, identification of potential ligand molecules and evaluation of biological properties, such as potency, affinity and efficacy, as carried out through various experimental analyses (Fang 2012). It also provides the structural descriptions of the target-ligand complex for understanding the binding mode and conformations, characterization of key molecular interaction, characterization of unknown binding sites, mechanistic studies and elucidation of ligand-induced conformational changes (Kahsai et al. 2011). Methods used in SBDD such as molecular dynamics give insight into not only how ligands bind with target proteins but also consider the target flexibility and interaction of pathway. SBDD has contributed to several compounds reaching the clinical trial stage and getting FDA approval to go into the market (Burger and Abraham 2006; Wang et al. 2010; Hanson et al. 2015). Thus, SBDD is a cyclic process consisting of several steps, starting from a known target structure, then going on to several in silico studies, which are conducted to identify potential ligands. The mechanism of structure-based drug design is explained in Fig. 11.1, which shows the binding site feature of the protein (Fig. 11.1a); the available drug molecules displaying the binding phenomenon with the binding site, with a few empty spaces that may be filled with water molecules (Fig. 11.1b); and finally the new drug, designed as per the binding site feature that perfectly fits with the binding site (Fig. 11.1c).

11.3.2 Ligand-Based Drug Design (LBDD)

LBDD is an one the often used method in computer aided drug design effectively used in the absence of the 3D structure of the target and the binding site is not accurately known, then a ligand-based drug design (LBDD) approach is a popular technique in the case of experimentally active compounds that bind to the biological target of interest. The common assumption in drug identification is that similar compounds with similar chemical properties may exhibit similar biological activity. Ligand-based virtual screening (LBVS) is based on the exploration of molecular descriptors gathered from known active compounds. In general, similar characteristics of a compound series are identified and subsequently applied as molecular filters. These filtering methods are used to discover potential lead molecules for experimental evaluation and reduce the chemical space to be explored in further screening steps (Geppert et al. 2010; Sliwoski et al. 2013). This is the main principle and motivation of LBDD, where a compound with interesting biological properties can act as template for finding potential lead molecules. Basically, three approaches –2D fingerprints, 3D methods and pharmacophores—are widely used for defining and quantifying chemical similarity in LBDD.

11.3.2.1 Pharmacophore Modeling

Pharmacophore model prediction is an essential way to describe those steric and electronic features needed for optimal interaction of lead with receptor molecules. According to the International Union of Pure and Applied Chemistry (IUPAC), pharmacophore is "the ensemble of steric and electronic features … necessary to ensure the optimal supramolecular interactions with a specific biological target structure to trigger or to block its biological activity." (Kaserer et al. 2015). In drug discovery approaches with small molecules, it is important to analyze the assignment of proper protonation and tautomeric states of the lead molecules. Pharmacophore describes a set of interactions required to bind in the cavity of target molecules and a set of spatially arranged spheres of a certain type and diameter. These spheres are commonly known as pharmacophoric features (Fig. 11.2). They include hydrophobic centroids, hydrogen-bond acceptor, hydrogen-bond donor, positively ionizable groups and negatively ionizable groups— all common features which target their

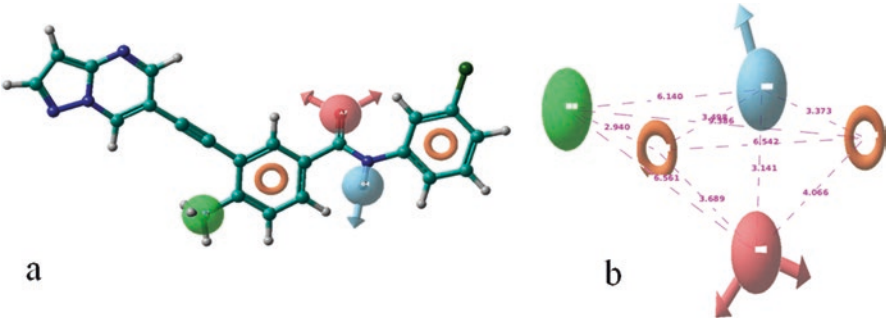


Fig. 11.2 Basic pharmacophore features (**a**) and (**b**) show the superimposed lead molecule with the pharmacophore model

corresponding sites. For example, a hydrophobic feature corresponds to hydrophobic protein side chains in the cavity; and a hydrogen-bond acceptor feature has a hydrogen bond-donating counterpart in the protein (Langer and Hoffmann 2006; Wolber and Langer 2005). A pharmacophore model was built from a collection of known partial agonists, and it was validated with a newly discovered partial agonist. Pharmacophore models are frequently employed in virtual screening processes to find a potential lead molecule. For example, Mustata et al. developed a potential lead molecule against Myc-Max via a pharmacophore model generated using known disruptors. In another study, Petersen et al. identified a novel PPARγ partial agonist using a pharmacophore model (Mustata et al. 2009; Petersen et al. 2011). Pharmacophore-based screening processes match all the atoms or functional groups and the geometric relations between them to the pharmacophore in the query. Basically, two steps are involved in a pharmacophore-based search: in the first step, software checks all the lead molecules as to whether it has the atom type or functional groups required by the pharmacophore; then it checks whether the spatial arrangement of this element matches the query.

2D pharmacophore searching

Searching of a 2D database to find potential lead molecules is one of the crucial steps in drug discovery. Pharmacophore-based virtual screening has been used for the identification of potential hit molecules in drug development process. This approach can used to screen virtually millions of compounds for hit identification. However, problems can arise from substructure when the number of compounds identified reaches into the thousands. This problem can be rectified by collecting these compounds based on similarity between compound in the database and in the query (Vyas et al. 2008). The structure activity relationship of these compounds can be generated in these processes even before synthetic pans are made for lead optimization based on the biochemical data (Enyedy et al. 2003). Beyond structure similarity, activity similarity has also been the subject of several studies.

3D pharmacophore searching

3D pharmacophore modeling acts as an efficient filter for virtual screening of large compound libraries due its simplicity and abstract nature. The computational complexity of the hit identification process in virtual screening is greatly reduced by the sparse pharmacophoric representation of ligand-protein interaction. The generation of a query pharmacophore model that specifies the type and geometric constraints of the chemical feature is the first step in a typical pharmacophore-based virtual screening experiment. Both ligand-based and structure-based models can be created and used separately or in combination via parallel virtual screening. Ligand-based screening is generally used when crystallographic solution structure or modeled structure is lacking. Both ligand-based and structure-based pharmacophores

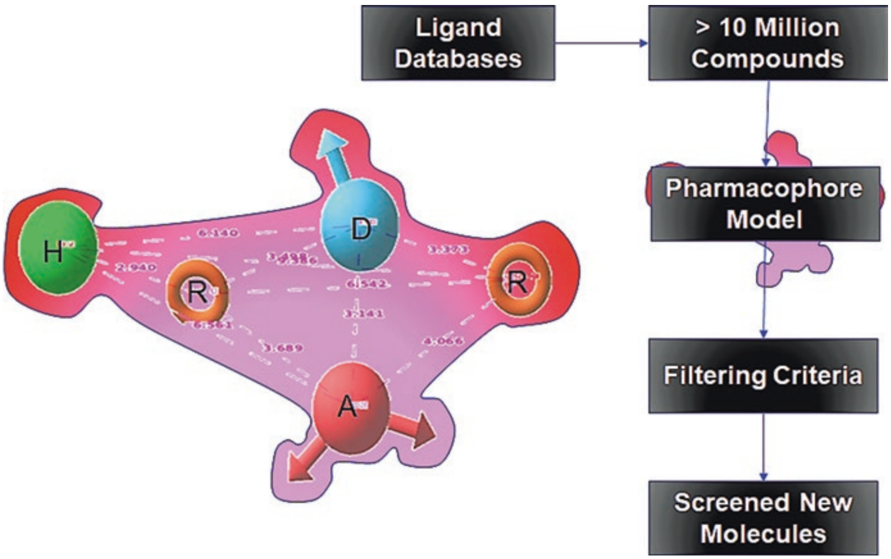


Fig. 11.3 Working method of 3D pharmacophore searching against small molecule databases

significantly screen the potential novel compounds with similar features and activity that can bind the same site of the proteins based on the features of the known compounds as mentioned in the Fig. 11.3. Several software products such as Catalyst, Sybyl/Unity, MOE and Phase are widely used methods for ligand-based pharmacophore building. Structure-based methods in pharmacophore modeling have gained significant interest in recent years, and several new approaches have been described, including the application of pharmacophore fingerprints for lead identification (Karnachi and Kulkarni 2006; Langer and Hoffmann 2006).

Fingerprinting

Pharmacophore fingerprints are defined as the binary encoded information about the presence or absence of pharmacophore features such as the centers and the three inter-center distances between them. By default, the seven center types that are probably the most important for the ligand-receptor interactions defined are: hydrogen-bond acceptor (A) and donor (D), groups with formal negative (N) and positive (P) charges, hydrophobic (H) and aromatic ring (R), and distance in a single molecule or a compound collection. Generally, fingerprinting focuses two or four-point pharmacophore fingerprints, but a larger number can be used, and utilization of up to nine pharmacophores has been described (Martin and Hoeffel 2000; Cato 2000). Traditionally, pharmacophore triplets are a widely used method and are most effective in terms of information content versus complexity; they are usually generated for a set of compounds instead of an individual one. For each compound, the flow energy conformer is calculated by every possible combination of three or four features and used to set the corresponding bit in the fingerprint. The obtained fingerprint is termed the ‘union key’ (Cato 2000). The generation of pharmacophore fingerprints for proteins with known binding site can be calculated from complementary site-points in the binding site. Methods such as ChemProtein module of Chem-X or the GRID program are often used for generation of site-points using a variety of probe atoms (Mason and Cheney 2000; Mason and Beno 2000). Chem-X is one of the most popular software packages. The fingerprinting in this module is defined according to all the potential pharmacophores that can be present in some low-energy conformer of the molecules. Another method, the Oriented Substituent Pharmacophore PRopErtY space (OSPPREYS) approach, introduced by Martin and Hoeffel, is aimed towards better representation of diversity and similarity in combinatorial libraries in the 3D pharmacophore space (Martin and Hoeffel 2000). Pharmacophore fingerprint methods have a wide range of applications; they can be used to measure molecular similarity (Willett 2006), to design libraries, to assess their diversity and to search them for novel active compounds (Beno and Mason 2001).

11.3.2.2 QSAR Modeling

Quantitative structure-activity relationship (QSAR) is a highly popular approach for ligand-based drug designing. This method significantly quantifies the correlation between the chemical structures of a series of compounds and a chemical or biological process. The basic mechanism underlying the QSAR method is that structurally similar molecules or those compounds having similar physiochemical properties yield similar activity (Akamatsu 2002; Verma and Hansch 2009). The first step of developing a QSAR model is identification of a group of chemical entities or potential lead molecules which show the desired biological activity. The developed QSAR model is then used to optimize the active compounds to maximize the relevant activity, and then it is tested experimentally for the desired activity. Mainly, four steps are involved in QSAR model prediction (Fig. 11.4). In the first step, potential lead molecules are identified with experimentally measured values of the desired biological activity. In second step, molecular descriptors associated with various structural and physiochemical properties of the molecules are identified, and in the third step, the correlation between molecular description and biological activity is discovered to explain the variation in activity in the dataset. Finally, the statistical stability and predictive power of the QSAR model is tested.

In the classical or the 2D QSAR method, various electronic, hydrophobic and steric features are correlated with biological activity for a congeneric series of compounds (Acharya et al. 2011). In the classical method the molecular descriptors used for correlation with activity are mostly representative of fragments of the parent molecule. The major advantage of the classical method is that it is more effective for a congeneric series of molecules; however, the fragment-based descriptors are

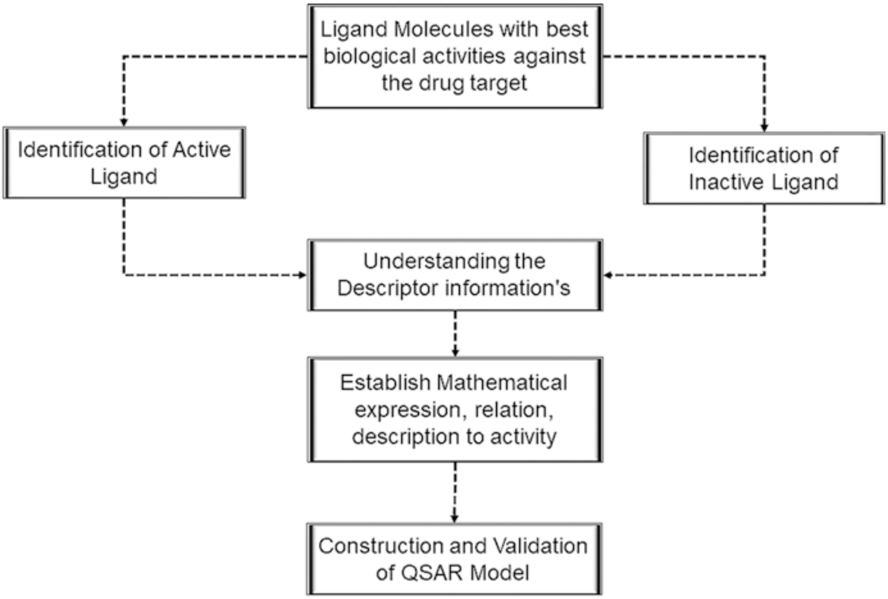

Ligand Molecules with best biological activities against the drug target
Identification of Active Ligand
Identification of Inactive Ligand
Understanding the Descriptor information's
Establish Mathematical expression, relation, description to activity
Construction and Validation of QSAR Model

Fig. 11.4 Working method of QSAR modeling and predictions

usually inadequate to capture 3D conformational features of the crucial step for its activity (Winkler 2002; Bernard et al. 2005; González et al. 2009). To describe the 3D features of molecules the new 3D QSAR method was developed in which various geometric, physical characteristics and quantum chemical descriptors are used to describe the 3D features of a molecule; those descriptors are then combined to create a pharmacophore that can explain the biological activity of ligands (Chang and Swaan 2006). Then, a developed pharmacophore model is subjected to stability and statistical analysis to obtain the final 3D QSAR model. Several techniques including CoMFA, CoMSIA and catalyst are currently used for this drug designing approach.

11.3.2.3 CoMFA

Comparative molecular field analysis (CoMFA) is one of the 3D QSAR techniques mainly used to describe structure activity relationships in a quantitative manner. In this method a set of molecules is identified and aligned based on their 3D structures on a 3D grid and the values of steric and electrostatic potential energies are calculated at each grid point. The identified lead molecules should have a similar binding mode (identical binding) to the same kind of receptor. In the next step, a certain group of molecules is selected as a training set to derive the CoMFA model. The residual molecules are considered a test set, which independently proves the validity of the derived models. A pharmacophore hypothesis of this method is generated to orient the superposition of all molecules and to afford a rational and consistent alignment. It calculates the values in each grid point, i.e., the energy of molecules via a carbon atom, a positively or negatively charged atom, a hydrogen-bond donor or acceptor, or a lipophilic probe, correlating these values with the biological activity. Principle component analysis (PCA) and partial least squares (PLS) are the most widely used methods for development of pharmacophore in CoMFA. The developed model is then tested for statistical significance and robustness (Gohda et al. 2000; Akamatsu 2002; Yasuo et al. 2009). The result of this approach can be represented as counter maps that indicate points of the lattice where variations in field values are related to variations in biological activity. These maps can be used to estimate the regions of molecules where some types of interactions have a favorable or unfavorable influence on the biological activity. Recently, several modifications have been described which significantly are used as alternatives to CoMFA (Sen et al. 2012).

11.3.2.4 Comparative Molecular Similarity Indices Analysis (CoMSIA)

CoMSIA is another 3D QSAR method, introduced by Klebe and his coworkers (1994) based on the calculation of similarity indices between the alignment's molecules and a common probe atom placed at the interaction grid. Most of the features of CoMSIA are similar to CoMFA; however, there are differences: The molecular field expression includes five different properties such as hydrophobic, hydrogen-bond donor and acceptor terms in addition to steric and coulombic contributions, and it calculates similarity indices instead of interaction energies by comparing each ligand molecule with a common probe. The statical evaluation of these field properties are correlated with the biological property by PLS analysis, but the counter maps are more contiguous and easier to interpret in CoMSIA because they are no cut-off values (Flower 2002; Klebe et al. 1994). To calculate the similarity indices, a Gaussian-type functional form is used to describe steric, electrostatic and hydrophobic compounds of the energy function, and it avoids using the arbitrary cut-off value for the energy calculation (Acharya et al. 2011). The Gaussian function also provides a smoother description of potential energy in regions near the van der Waals radius atom (Klebe et al. 1994).

11.4 Virtual Screening (VS) for Lead Discovery

The discovery of novel leads with potential interaction with targets is one of the important steps in drug discovery. This approach is conventionally achieved by wet-lab high throughput screening (HTS) in many pharmaceutical industries, but due to the high cost and low hit rate, the alternative method is developed with broad application of the cheaper and faster screening of in silico approaches (Clark 2008;

Ripphausen et al. 2010). Alternative virtual screening (VS) uses computational power to test a large set of small molecules in a limited time at low cost. VS is a stepwise process with a cascade of sequential filters able to narrow down and choose a set of lead-like hits with potential biological activity against intended drug targets. It can be broadly classified into two categories, ligand-based virtual screening (LBVS) and structure-based virtual screening (SBVS). A broad range of computational techniques that can be applied in this process includes drug likeness screening, counting scheme, functional group filters, topological drug classification, pharmacophore points filter and pharmacophore-based virtual screening. Molecular docking is a computationally intensive method that has been applied to very large databases of chemical structures.

Protein-ligand docking has become one of the widely used tools in modern drug discovery approaches to predict the most likely binding mode of small molecules at a particular receptor to explore specific interactions that may be formed and to estimate ligand-binding affinity. A number of protein-ligand methods are available to date, from academic groups to commercial software vendors. The binding free energy between protein and ligand molecules employs rather heuristic terms and these functions are referred as scoring function. Scoring functions is a very important step, which includes protein preparation, ligand database preparation, docking calculation and post processing. Basically, the scoring process composed of three different aspects relevant to docking and design. The first aspect is the ranking of the conformations generated by the docking research for one ligand interacting with a given protein; this aspect is crucial for detecting the binding mode that best approximates the experimentally observed situation. The second aspect is ranking the different ligands with respect to binding to one protein; that is, prioritizing ligands according to their affinity, which is essential in virtual screening and the third aspect is ranking one or different ligands with respect to their binding affinity to different compounds which is essential for the consideration of selectivity and specificity of ligands (Leach and Hann 2000; Lewis et al. 2000). The amount and quality of available information on the target protein is one of the key factors in designing a virtual screening project (Klebe 2006). The information on the coordinates of the features of the 3D structure of the known targets is valuable data and can be used to improve the quality of the results. The predictions of 3D structure of biomolecules are obtained by the three exemplary methods of NMR spectroscopy, X-ray crystallography and homology modeling. Currently, PDB contains more than 70,000 experimentally solved 3D structures of proteins that can be used as targets in VS and in homology modeling.

11.4.1 Protein Modeling

Proteins are the fundamental structural elements in living organisms; they act as catalytic agents, signal transmitters, transporters and molecular machines in cells (Nelson et al. 2008). Mostly, most the proteins are not functions individually; they must interact with other molecules to carry out their cellular roles, if any alteration in the protein interface leads to a pathological condition. Hence, the protein interface may be used as potential targets for rational drug designing approaches (Rask-Andersen et al. 2011; Jubb et al. 2015). Many experimental methods including NMR and X-ray crystallography have been used to identify and characterize the protein-protein interface at the level of individual atoms and residues, and various mass spectrometry-based approaches such as chemical cross-linking and hydrogen/deuterium exchange have been used, which typically report the location of interface at lower resolution (Hoofnagle et al. 2003; Kaveti and Engen 2006; Gobl et al. 2014; Shi 2014). Though these experiments provide valuable knowledge of the protein recognition mechanism, technical challenges such as expressing and purifying aggregation-prone protein samples, obtaining high quality crystals and protein size constraints are both labor-intensive and time-consuming. Hence, in the absence of an experimentally determined structure, an alternative computational approach such as comparative or homology modeling is used to predict the 3D model of proteins related to at least one known protein structure. The model gives the 3D structure based on its alignment to one or more known protein structures (Pieper et al. 2002).

11.4.1.1 Homology Modeling

Comparative or homology modeling is one the easiest methods among the three-structure prediction approach. In homology modeling, the structure process consists of fold assignment, target-template alignment, model building and model evaluation. There are several computer programs and web servers that automate the comparative modeling of proteins. Generally, the 3D structure of proteins can be achieved by several different approaches and is strongly dependent on the sequence identity (SI) or the percentage of identical amino acid residues present among the target sequence and their templates (Santos Filho and Alencastro 2003). Ab inito is the another method used for prediction of 3D structure of protein and mostly suitable, when there is no suitable template with significant sequential identity to the target sequence. If the sequence identity between target and template protein is above 30%, comparative or homology modeling is a suitable approach (Baker and Sali 2001; D'Alfonso et al. 2001). In practice, homology modeling consists of the seven important steps, which are template recognition and initial alignment, alignment correction, backbone generation, loop modeling, side chain modeling, model optimization and model validation (Peitsch et al. 2000; Westbrook et al. 2002; Orengo et al. 2002; Lo Conte et al. 2002).

Template selection is the initial step in safe homology modeling. The percentage of sequence identity between the sequence of interest (query) and a possible template can be detected by different software. The template model can be found using the query sequence from a database such as the protein data bank (Westbrook et al. 2002), SCOP (Lo Conte et al. 2002) and CATH (Orengo et al. 2002). Three main classes of protein comparison methods are involved in fold identification. Initially, the target sequence is subjected to pairwise sequence alignment with each database sequence independently to find its homologous sequence (Fiser 2010). Computational programs such as BLAST (Schäffer et al. 2001), FASTA (Srivastava et al. 2009) and CDART are frequently used methods for searching the related protein sequence and structure of the template. The second class of method employed is a multiple sequence alignment profile to compare the sequence using profile analysis profile-profile comparisons, Hidden Markov models and intermediate sequence search (Rychlewski et al. 2000 Yona and Levitt 2002; Zhou and Zhou 2005; Fiser 2010). SAM and PSI-BLAST (Karplus et al. 2003) are the most often used programs for this approach. The third class of method is also a pairwise alignment method, where the target sequence adopts any one of the many known 3D -folds predicted by an optimization of the alignment with respect to a structure-dependent scoring function independently for each sequence-structure pair; i.e., the target sequence is threaded through a library of 3D-folds (Kelley et al. 2000).

The next important step is a sequence alignment between the target and template structure. Mostly, fold assignment methods are widely used in this process and it is agreed that profile-based alignment produce better quality models than sequence-based alignments. In addition, HMM-based alignments produce higher quality model than PSSM-based method alignments produced by PSI-BLAST (Yan et al. 2013). A pairwise comparison of protein sequence and protein structure is matched against a library of 3D profiles, this method is also known as fold assignment. Once a list of potential templates is obtained using different searching methods, it is necessary to select a potential template more appropriate for the modeling problem. The selection of highest sequence similarity is the simplest template selection rule for modeling the protein (Retief 2000). After the selection of a potential template, a suitable method is used to construct the 3D model from template and alignments. Generally rigid-body assembly, segment matching, spatial restraint and artificial evolution are used for model building. This rigid-body assembly model relies on the natural dissection of the protein into conserved core regions, variable loops that connect them and side chains that decorate the backbone. The segment matching based on the construction of a model by using a subset of atomic positions from template structure and by identifying and assembling short. All atom segments in the model that fit the guiding positions can evaluated by scanning all the known protein structures (Xiang 2006). Several programs are available for modeling the query sequence. Andrej et al. developed MODELLER, which remains one of the most widely used comparative modeling methods. The spatial restraints approach is implemented in MODELLER. It starts by aligning the target sequence with the related known 3D structure, and the output obtained by this method contains a molecular structure that includes main chain and side chain non-hydrogen atoms similar to the known structure. In addition to MODELLER, other tools including Swiss Model, RAMP, PrISM, COMPOSER, CONGEN+2 and DISGEO/Co-sensus are often used in homology modeling (Schwede et al. 2003; Vyas et al. 2012). This homology modeling approach is described in several available programs, both in the commercial and public arena.

Model evaluation and validation is necessary to construct a model with good stereochemistry; the most important factor in the assessment of constructed models is the scoring function, and programs evaluate the location of each residue in a model with respect to the expected environment as found in the high-resolution X-ray structure. The stereochemistry of the modeled protein can be verified by the analysis of parameters like bond lengths and angles, torsional angles and chirality of residues using PROCHECK (Laskowski et al. 1993), WHATCHECK (Hooft et al. 1996), PROSA (Sippl 1993) and Molprobity (Davis et al. 2007; Chen et al. 2010). The reliability of a predicted model is also subject to a check of other parameters such as planarity of the peptide bond, chirality of the Cα, bond length and angles in the main chain, the planarity of aromatic system, the inner backing of globular proteins and the elements of the secondary structure, hydrophobic and hydrophilic residues of the predicted protein structure (Schwartz et al. 2001).

11.4.1.2 Threading

In comparative modeling it has been observed that the careful alignment of the corresponding amino acid residues of the unknown proteins with a similar sequence, often closely related homologues, tend to have similar 3D structure with similar conformations. When no sequences are clearly related to the modeling target, the alternative method of threading is employed to predict structure via fold recognition. Protein threading, i.e., sequence-structure alignment, is a promising template based on fold recognition, which identifies a suitable fold from a structure library for the query sequence and provides an alignment between the query protein and the fold (Shan et al. 2001). The word ‘threading’ was first coined by Jones et al. (1992); the original term was ‘optimal sequence threading,’ later it shortened ‘threading.’ In this method, the query sequence is threaded onto the backbones of the template structures. Threading requires four basic components: (1) a template library representing the 3D protein structure to be used as the template; (2) an energy function to describe the fitness of any template; (3) a threading algorithm to search for the lowest energy among the possible alignments for a given sequence-template pair; (4) a criterion to estimate the confidence level of the predicted structure. The treading method is further classified into two broad categories, singleton threading, in which the threading considers only the preference of amino acids in the query sequence at single sites of the templates; and a category that uses the preference on pairs of amino acids in the query sequence within contact distance when they are aligned to a given structure. Singleton threading constructs a 1D structure profile for each amino acid residue position in a template using the 3D structural information, such as secondary structure type, degree of environmental polarity and fraction of residue surface accessible to solvent. Typically in threading, it is assumed that the backbones of the structures are rigid and only the amino acid side chains of the query and the template are different. Threading exploits the fact that proteins with different functions can possess a similar structure even though they may have little to no sequence similarity. Loopp and therader are software (learning, observing and outputting protein patterns (Tobi and Elber 2000; Meller and Elber 2001; Teodorescu et al. 2004) can be used for structure prediction via fold recognition. Both loop and threader rely on similar strategies, yet they use different energy and scoring functions to generate possible alignments with feasible templates. THREADER uses solved protein structure as a scaffold on which to place the target protein sequence and analyze secondary structure information about the target sequence used to force alignment between predicted secondary structures of the target. It uses a set of basic knowledge-based potentials such as statistical data compiled from known protein structure and pairwise pseudo-energy to indicate misfolded proteins.

The strategy of LOOPP is similar to THREADER, but it differs in its implementation of an empirical energy function and its scoring method. The most notable aspect of LOOPP is its extensive parameterization, which is based on the structure from the protein data bank (PDB) and a database of close to five million decoy structures (Berman et al. 2000; Tobi and Elber 2000). Three novel implementations of common protocol—the pairwise contact model, gap penalties and Z-scores—differentiate LOOPP from other threading methodologies. It creates a new pairwise interaction model (empirical energy function) acting as the key to devising a truly novel threading algorithm. Basically, two main types of empirical energy functions exist in this method: (1) those that pairwise residues contacts for residues within a specified distance of one another; (2) those based on the environment of an amino acid residue at a point in the structural lattice (Meller and Elber 2001). Several threading programs including the NCBI threading package (Bryant and Lawrence 1993), PROFIT (Sippl and Weitckus 1992), PROSPECT (Xu et al. 1998), CASP-3 (CASP 1999), TOPITS (Rost and Sander 1995) and SAS (Milburn et al. 1998) are used for singleton and pairwise interactions. The NCBI threading package provides a good statistical assessment of a threading result, and recently CASP-3 was used as a top performer in threading with pairwise interactions.

11.4.1.3 Ab Initio Method

Ab initio method is one of the modeling technique often used for structure prediction when the sequence of the query proteins has either no or a low amount of similarity and in this method the query protein is folded with a random conformation. The ab initio method is based on the thermodynamic hypothesis proposed by Anfinsen, according to which the native structure corresponds to the global free energy minimum under a given set of conditions (Floudas et al. 2006). Basically, the ab initio category has two subclasses, fragment-based and biophysics-based methods. These are often called, respectively, first-principles methods that employ database information and first-principles methods without database information (Floudas 2007). All types of proposed approaches rely on minimization of the energy function over the conformation parameters. The typical method has four basic steps for finding the conformation with the lowest energy: (1) start with an unfolded/arbitrarily folded conformation; (2) generate alternative conformations using some heuristics; (3) estimate their corresponding energy; and (4) again, generate the alternative conformation until the final criterion is reached. Parameters like energy function accuracy, search algorithm efficiency and selection of the best models play a crucial role in the structure prediction ab initio method. In the basic modeling, folding process, and quantum mechanics is used to model and estimate the interactions of atoms. Currently, a high performance computing facilities force field (FF) or energy function are employed to express a variety of atomic interactions such as van der Waals, torsion angles, electrostatics and bond length. Energy functions are usually associated with the search procedure to locate the conformation that has the minimum energy function value. The most popular optimization methods are molecular dynamics and Monte Carlo simulation (Adcock and McCammon 2006). The category of ab initio prediction with database information focuses only on predicting as accurately as possible a protein's final configuration. In this approach, the structure prediction starts with the primary amino acid sequence, which is searched for different conformations, leading to the prediction of native folds. After the folds have been recognized and predicted, the model assessment is performed to verify the quality of the structure. ROSETTA and I-TASSER are widely used fragment-based enhanced methodologies for ab initio structure prediction of a protein. TASSER was initially created in 2004 by Zhang and Skolnick (2004), and later the enhanced versions Chunk-TASSER (Zhou and Skolnick 2007) and I-TASSER were developed in structure prediction (Wu et al. 2007). TASSER is a hierarchical approach that encompasses three phases, thus its name: threading/assembly/refinement ("TASSER"). The first step, threading, is an iterative sequence-structure alignment algorithm that uses the program PROSPECTOR_3 (Skolnick et al. 2004). The second step, assembly, uses parallel hydrophobic Monte Carlo sampling by rearranging the template fragments (Zhang et al. 2005). The final step, refinement, is performed using a clustering program called SPICKER (Zhang and Skolnick 2004), and the full atom optimization is conducted using the CHARMM22 force field. ROSETTA prediction involves the identification of small fragments from the structural databases consistent with a local sequence preference.

11.4.1.4 Protein Validation Server

Protein structure has proved to be a crucial piece of information for biochemical research. From the millions of currently sequenced proteins only a small fraction is experimentally solved for structure, and the only feasible way to bridge the gap between sequence and structure data is computational modeling. Unlike experimental structure, the accuracy of a computationally modeled structure can be estimated by a broad range of the accuracy spectrum. Over the past two decades, several approaches have been developed to analyze the accuracy of the protein structure and model. They use stereochemistry checks, molecular mechanics energy-based functions and statistical potentials to tackle problems. Typically, features like molecular environment, hydrogen bonding, secondary structure, solvent exposure, planarity, chirality, phi/psi preference, chi angles, non-bonded contact distances, unsatisfied donors/acceptors, pairwise residue interaction and molecular packing are analyzed in these approaches. A good quality protein should resemble a native protein, with spatial features of the residues complying with empirically characterized constraints on torsional angles captured in Ramachandran plots (Ramachandran et al. 1963). PROCHECK (Laskowski et al. 1993) and MolProbity (Chen et al. 2010) are widely used programs for determining whether a modeled protein structure has native-like features. Traditionally, several studies have examined protein structures using an all atom-based description. Ramachandran's plot with backbone dihedral angle ϕ (N-Cα) and ψ (Cα-C) is a representative microscopic description of the protein structure. Dihedral angle prediction has several applications in protein structure prediction; which include secondary structure prediction (Rost 2001; Wood and Hirst 2005; Kountouris and Hirst 2009), generation of multiple alignments (Huang and Zou 2006a, b; Miao et al. 2008), identification of protein fold (Karchin et al. 2003; Zhang et al. 2008) and fragment-free tertiary structure prediction (Faraggi et al. 2009). Quality assessment is an important step in the modeling process, wherein processes like template level, alignment level, selected fragment level and structural level error are analyzed. A template structure for a target sequence is identified by considering the significance of the score that indicates the fitness of the target to the template. In principle, most frequently the statistical significance of a raw score is considered as either in the form of the E-value (homology search) or the Z-score (used in threading algorithms). Z-score are calculated as measured value minus population mean, divided by the standard deviation of the population. So, a Z-score is negative if the value of X is less than the mean, and it is positive if the measured value is greater than the mean value. WHAT IF uses this criterion a lot to calculate Z-score. The Z-score provides basic information about the root mean square of a population with a Z value and it should be 1.0.

11.4.2 Protein and Ligand Preparation

The success of the various drug designing approaches depends largely on whether reasonable starting structures are used for both the protein and the ligand. The protein structure that is retrieved from PDB (X-ray structure) consists of heavy atoms and may contain water molecules, cofactors, activators, ligands and metal ions as well as several protein subunits and does not have the information on bond orders, topologies. Because of the above structural issues, several protein preparation approaches have been developed (Sastry et al. 2013; Pitt et al. 2013). The determination of protonation states of the amino acid in protein molecules is the first crucial step in protein preparation. Several freely available software packages including PROPKA (Li et al. 2005), H++ (Anandakrishnan et al. 2012) and SPORES (ten Brink and Exner 2010) are widely used for determining the first step of the protein preparation. The next important step is to assign hydrogen atoms and optimize protein hydrogen bonds according to an optimal hydrogen bond network. PDB2PRO software is a widely used tool for these tasks (Dolinsky et al. 2007). The next step is assignment of partial charges, capping of residues, treating metals, filling missing loops and missing side chains and minimizing the protein structure to relieve steric clashes; also, a crucial decision must be made regarding whether water molecule will be left in or removed from the binding site. To tackle the above mentioned challenging problems, freely available tools such as 3D-RISM (Kovalenko 2003; Young et al. 2007; Abel et al. 2008), SZMAP (Myrianthopoulos et al. 2016), JAWS (Michel et al. 2009) and WaterMap (Young et al. 2007; WaterMap, Schrödinger 2014) are utilized in commercial software (Jorgensen and Tirado-Rives 2005; SZMAP Sofware Inc.). In the case of a co-crystallized protein structure with substrates and cofactors, Protein Preparation Wizard of Maestro (Maestro, Schrödinger, LLC) is used to assign proper bond orders and generate accessible tautomer and ionization states prior to virtual screening.

The selection of the type of ligand molecule chosen for docking is another important step in virtual screening. The type can be obtained from various databases like ZINC or pubchem, or it can be sketched by means of Chemsketch or Chemdraw tools (Dias and de Azevedo 2008). A wide variety of small molecule databases are available for virtual screening-based drug designing. Many of them are free and possess desirable characteristic lead molecules. ZINC is a public access database, contains number of commercially available compound that are mostly developed in the pharmaceutical chemistry department at the University of California, San Francisco. NCI is an another open database developed by the Developmental Therapeutics program of the National Cancer Institute, NIH; it currently contains over 250,000 molecules from both organic synthesis and natural sources. ASINEX is a regularly updated commercial database currently containing 600,000 screening compounds, 27,000 macrocycles, 23, 000 fragments and 7000 building blocks. SPECS is a monthly updated database containing more than 240,000 novel drugs—drug-like small molecules obtained from an academic research institute. MAYBRIDGE is one of the widely used commercial databases containing a screening hit discovery collection more than 53,000 and offering a fragment library of 30,000. CHEMBRIDGE encompasses one million drug-like and lead-like molecules in two non-overlapping collections of respectively 460,000 and 620,000 compounds. After selection of potential lead molecules, it should be preprocessed before docking. There are several thousand small molecules in a ligand database, so one must avoid performing manual steps in data preparation. Typically, information on available ligands is stored in 2D form in databases, serving as a data repository. Currently, several thousand small molecules are available in various databases; Table 11.2 shows widely used small molecule repositories. The 2D structure retrieved from these repositories of atom and bond types must be checked and corrected; protonation states and charges have to be assigned. Then, 3D structures must be converted for calculating ligand conformation like rotational barriers or side-chain rotamers allowed. In addition, protein-ligand interactions including site-points that guarantee proper hydrogen-bonding directionality must be assigned (Claussen et al. 2001). LigPrep is the most widely used module for ligand preparation implemented in Schrödinger (LigPrep, Schrödinger 2011). In this module, ionization/tautomeric states are generated with either a pair of fast rule-based programs or with Epik, which is based on the more accurate Hammett and Taft methodologies (Shelley et al. 2007; Epik, Schrödinger 2011).

11.4.3 Active Site Prediction

Binding site prediction and characterization of small molecules is more important for drug discovery. Often, possible binding sites for potential small molecules are known for co-crystal structures of the target or a closely related protein with natural ligand molecules. Recently, Hajduk and coworkers used heteronuclear-NMR-based screening to identify and characterize the ligand binding site on a protein surface (Hajduk et al. 2005). By screening a large number of lead-like molecules against 23 target proteins, the results revealed that 90% of the ligand molecules bonded to specific locations on the protein surface, depicting that certain properties of small-molecule binding sites should be common to general molecular recognition. Mostly computational studies have been used to predict the binding site for an unknown or if a new binding site is to be identified, e.g., allosteric molecules. Computational methods like Q-SITEFINDER, POCKET (Levitt and Banaszak 1992), SURFNET (Laskowski 1995), APROPOS (Peters et al. 1996), LIGSITE (Hendlich et al. 1997), CAST, CASTp (Binkowski et al. 2003) and PASS (Brady and Stouten 2000) are often used for binding site prediction. Computational methods for the identification of a binding site can be categorized into three major classes: (1) geometric algorithms to find the shape concave invagination in the protein molecules; (2) energies-based method; and (3) method considering dynamic of protein structures. Geometric algorithms find a putative binding site through detection of cavities on a protein surface. In this algorithm, grids are used to describe the molecular surface of the protein, and the boundary of the binding site is determined by rolling a spherical probe over the grid surface. This kind of algorithm is used in SURFNET, LIGSITE and POCKET, where spheres are placed between all pairs of target atoms and then the radius of sphere is reduced until each sphere contains only a pair of atoms. An et al. (2005) developed the Pocket Finder algorithm and expanded the geometric method by countering a smoothed van der Waals potential for the target protein to identify candidate ligand binding sites. The new technique of Sitemap, developed by Schrödinger, Inc., identifies the known binding site in >96% of cases by linking together site-points that contribute to tight protein ligand binding. Sitemap provides quantitative and geographical information that helps guide efforts to modify ligand structure to enhance properties (Halgren 2007; Halgren 2009) (Table 11.1).

11.4.4 Molecular Docking

In a modern drug discovery approach, protein-ligand and protein-protein interaction mechanisms play a significant role in predicting orientation of the ligand when it is bound to a protein receptor or enzyme using shape and electrostatic interaction to quantify it. Molecular docking is an attractive scaffold for understanding protein-ligand interaction in a rational drug design and drug discovery; in the mechanistic study a molecule is placed into the binding site of the receptor molecules mainly in

Table 11.1 Widely used small molecule repositories with basic information about the class of the compounds and their size

Database	Type	Size	Citations
PubChem	Biologic activities of small molecules	40,000,000	Wheeler et al. (2006)
Accelrys Available Chemicals Directory (ACD)	Consolidated catalog from major chemical suppliers	7,000,000	Accelrys (2012)
PDBeChem	Ligands and small molecules referred in PDB	14,572	Dimitropoulos et al. (2006)
Zinc	Annotated commercially available compounds	21,000,000	Irwin and Shoichet (2005)
LIGAND	Chemical compounds with target and reactions data	16,838	Goto et al. (2002)
DrugBank	Detailed drug data with comprehensive drug target information	6711	Wishart et al. (2006)
ChemDB	Annotated commercially available molecules	5,000,000	Chen et al. (2005, 2007)
WOMBAT Data base	Bioactivity data for compounds reported in medicinal chemistry journals	331,872	Ekins et al. (2007); Hristozov et al. (2007)
MDDR (MDL Drug Data Report)	Drugs under development or released; descriptions of therapeutic	180,000	Hristozov et al. (2007)
3D MIND	molecules with target interaction and tumor cell line screen data	100,000	Mandal et al. (2009)

a non-covalent fashion to form a stable complex of potential efficacy and more specificity (Rohs et al. 2005; Guedes et al. 2014). The information obtained from a docking study can be used to study the binding energy, free energy and stability of drug-biomolecular complexes with optimized conformation and with the intention of possessing less binding free energy. The basic two steps involved in molecular docking, usually related to sampling methods and scoring schemes, are (1) prediction of ligand conformation and position and orientation within these sites (usually referred as pose) and (2) assessment of binding affinity (Fig. 11.5).

Most of the docking tools employed the searching algorithms including genetic algorithms (GA), Monte Carlo algorithms, molecular dynamics algorithms and conformational search algorithms in the molecular docking method. Conformational search algorithms perform in the docking approach by applying systematic and stochastic search methods (Agrafiotis et al. 2007; Yuriev et al. 2011). The basic methodology of molecular docking falls into three categories: induced fit docking, where both ligand and receptor molecules are flexible; rigid body docking, where ligand and receptor molecules are rigid; and flexible docking method, in which it is also the case that both interacting molecules are flexible (Meng et al. 2011). The molecular docking process involves the following major steps: (1) Preparation of protein—

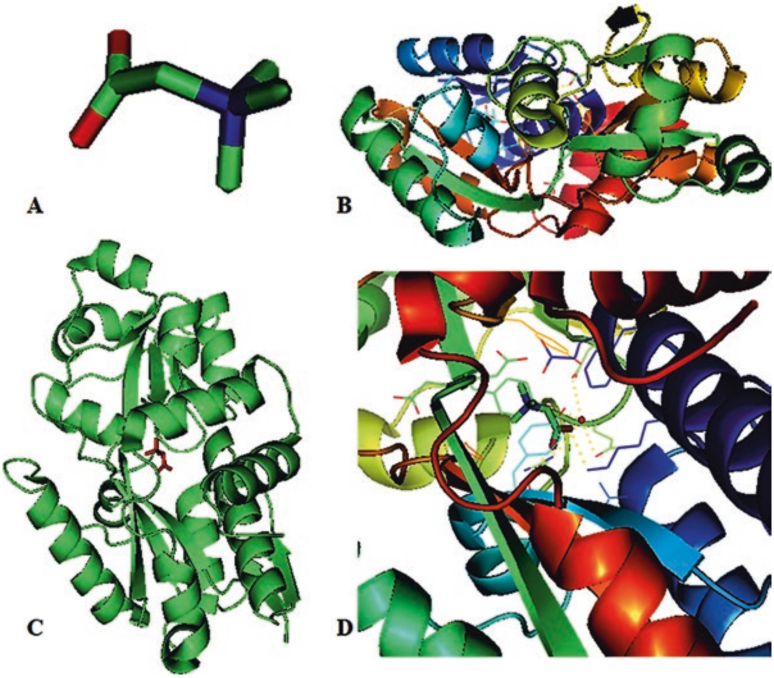


Fig. 11.5 Basic steps involved in molecular docking approach. (**a**) Three-dimensional structure of lead molecules; (**b**) three-dimensional structure of the protein; (**c**) ligand is docked into the binding site of the protein; (**d**) binding affinity and interactions of ligand molecules with protein

before docking, the 3D structure of the receptor molecule (retrieved from either PDB or molecular modeling) should be pre-processed by stabilizing the charges, filling the missing residues, and generating and removing free water molecules from the cavity. (2) Active site prediction—the binding site of the receptor molecules should be predicted in this step; the water molecules and hetero atoms are removed. (3) Ligand preparation—the small molecules can be retrieved from small molecule databases while choosing the ligand molecules; the LIPINSKY'S RULE OF 5 should be utilized. (4) Docking—the final step, where the ligand is docked against the protein and the interactions are analyzed; the scoring function finds the docking scores based on best pose of docked ligands complex. Over the last two decades, approximately 60 different docking tools and programs have been developed for both academic and commercial use, including DOCK (Venkatachalam et al. 2003), Auto Dock (Österberg et al. 2002), FlexX (Rarey et al. 1996), Surflex (Jain 2003), GOLD (Jones et al. 1997), ICM (Schapira et al. 2003), Glide (Friesner et al. 2004), Cdocker, LigandFit (Venkatachalam et al. 2003), MCDock, FRED (McGann et al. 2003), MOE-Dock (Corbeil et al. 2012), LeDock (Zhao and Caflisch 2013), AutoDock Vina (Trott and Olson 2010), Dock (Ruiz-Carmona et al. 2014) and UCSF Dock (Allen et al. 2015). Table 11.2 shows the basic information on the currently used docking tools and scoring functions.

Table 11.2 Basic characteristics of widely used docking tools

S. No	Docking programs	Docking approach	Scoring function
1	DOCK	Shape-fitting (sphere sets)	Chem Score, GB/SA solvation scoring, other
2	Auto Dock	Genetic algorithm Lamarckian genetic algorithm simulated annealing	Auto Dock (force-field methods)
3	Flex X	Incremental construction	FlexX Score, PLP, Screen Score, Drug Score
4	FRED	Shape-fitting (Gaussian)	Screen Score, PLP, Gaussian shape score, user-defined
5	Glide	Monte Carlo sampling	Glide Score, Glide Comp
6	GOLD	Genetic algorithm	Gold Score, Chem Score user defined
7	Ligand Fit	Monte Carlo sampling	Lig Score, PLP, PMF
8	Surflex	Surflex-Dock search algorithm	Bohm's scoring function
9	ICM (Internal Coordinate Modelling)	Monte Carlo minimization	Virtual library screening scoring function
10	MVD (Molegro Virtual Docker)	Evolutionary algorithm	MolDock score
11	FITTED (Flexibility Induced Through Targeted Evolutionary Description)	Genetic algorithm potential of mean force	(PMF), Drug Score
12	GlamDock	Monte Carlo method	ChillScore
13	vLifeDock	Genetic algorithm	PLP score, XCscore
14	iGEMDOCK	Genetic algorithm	Empirical scoring function

11.4.5 Scoring Methods

Molecular docking approaches use scoring functions to calculate the binding energies of the predicted ligand-receptor complexes. Scoring function is a key element of a protein-ligand docking algorithm, determining the accuracy of the algorithms (Gohlke and Klebe 2001; Schulz-Gasch and Stahl 2004; Jain 2006; Rajamani and Good 2007; Gilson and Zhou 2007). Speed and accuracy are the important aspects basic to a scoring function. Several scoring functions have been used mainly to delineate correct poses from incorrect poses, or binders from inactive compounds within a reasonable computation time. Overall, scoring functions can be divided in the three categories of as force field-based, empirical-based and knowledge-based scoring functions (Kitchen et al. 2004). A classical force-field scoring function estimates the binding energy of a complex by calculating the sum of bonded terms such as bond stretching, angle bending and dihedral variation, and non-bonded terms including electrostatic and van der Waals interactions. Electrostatics terms use a set of derived force-field parameters such as AMBER or CHARMM (Miller et al. 2017) and are calculated by a coulombic formulation. In addition to the above electrostatic terms, the force field-based scoring function also considers hydrogen bond, solvation and entropy contributions. The software such as DOCK (Kuntz et al. 1982), GLOD (Shoichet et al. 1993) and Auto Dock (Morris et al. 1998) offer users such functions. Force fields are mathematical expressions describing the dependence of energy of a system on the coordinates of its particles. The force-feild scoring function shows some differences in the treatment of hydrogen bonds in terms of the energy function used, and it is further refined with other techniques such as linear interaction energy (Michel et al. 2006) and free-energy perturbation method (FEP) (Kollman 1993; Briggs et al. 1996) to improve accuracy in predicting binding energies. To reduce computational expense, alternative approaches such as Poisson-Boltzmann/surface area (PB/SA) and the generalized-Born/surface area (GB/SA) models were used to measure accuracy by treating water as a continuum dielectric medium (Rocchia et al. 2002; Liu and Zou 2006; Lyne et al. 2006; Thompson et al. 2008; Guimaraes and Cardozo 2008).

Empirical scoring function is another method to evaluate the types of physical events involved in the formation of the ligand-receptor complex. The binding energy of a complex is calculated by summing up a set of empirical energy terms including van der Waals energy, electrostatic energy, hydrogen bonding energy and desolvation terms. Each empirical energy term component is multiplied, and corresponding coefficients are determined by reproducing the binding affinity data of a training set of protein-ligand complexes with known three-dimensional structure using least squares fitting (Ballester and Mitchell 2010). Due to the simple energy terms and the nature of their fitting to known binding affinities of the training set, empirical scoring functions are computationally more efficient and faster than force-field-based methods. Molecular docking tools such as Surflex and FlexX and Glidescore (Friesner et al. 2004; Halgren et al. 2004), PLP (Gehlhaar et al. 1995; Gehlhaar et al. 1999), SYBYL/F-Score (Rarey et al. 1996), LigScore (Kramer et al. 1998) and Chemscore are some examples of programs that use empirical scoring functions (Jain 2003). Table 11.3 provides the widely used scoring functions implemented in the most frequently used molecular docking programs.

Table 11.3 Provides widely used empirical scoring functions in frequently used molecular docking tools

Force-field-based	Empirical	Knowledge-based
DOCK	Auto Dock	SMoG
Auto Dock	Gold Score	Drug Score
Glide Score	Chem Score	PMF_Score
ICM	X_Score	Motif Score
LigandFit	F_Score	RF_Score
Molegro Virtual Docker	Fresno	PESD_SVM
SYBYL_G-Score	SCORE	Pose Score

A third approach includes knowledge-based scoring functions that use statistical analysis, which are directly derived from the structural information in an experimentally determined protein-ligand complex to obtain interatomic contact frequencies and distance between the ligand and protein. Further, this approach uses pairwise energy potentials derived from a known ligand-receptor complex to obtain a general function (Huang et al. 2006). These potentials are constructed by considering the frequency distribution and the score is calculated by summing up of the individual interactions. Compared to force field and empirical scoring functions, knowledge-based scoring functions offer a good balance between accuracy and speed and are relatively robust and also enable the scoring process to be as fast as the empirical scoring function (Muegge 2006; Huang and Zou 2006a, b). Recently, a consensus scoring method has been developed which combines several scores to assess the docking conformation.

11.4.6 Molecular Dynamics (MD) Simulations

Molecular dynamics (MD) simulations of recent years play a critical role in computational drug discovery. Simulation studies can provide detail concerning individual particle motion as a function of time and use physics-based energy functions and explicit representations of atomic systems to model protein dynamics. MD simulation studies provide basic information to evaluate the stability and functions of the protein and to monitor the specific behaviors over the course of many simulations and provide information about target structure or properties unobtainable from static native structure. MD simulation was first developed in the late 70s when Alder and Wainwright performed it using a hard-sphere model. The first molecular simulation of BPTI was done in 1975 with a crude molecular mechanics potential for only 9.2 ps (Adcock and McCammon 2006). Molecular dynamics simulation mimics the physical motion of each atom in the macromolecule present in the actual environment. Each atom of a protein molecule can interact for a certain period of time, which helps in the computation of their trajectory in and around the protein molecules. A variety of properties such as free energy, kinetics measures and other macroscopic quantities of macromolecules can be calculated by using the trajectories. Several studies revealed the role of classical MD simulations to obtain different conformations of proteins and nucleic acids, including early attempts to stimulate spontaneously complex phenomena such as protein folding (Frenkel and Smit 2001). In recent research, MD simulation has been widely used to overcome the major limitation of static structure-based drug design and also to characterize routinely applied ligand docking calculations which do not sample the major protein conformational rearrangements during ligand binding (Carlson 2002; Fanelli et al. 2008). MD simulation is a multistep process that starts with the knowledge of the potential energy of the system with respect to its position coordinates, and these position coordinates help to compute the force acting on the individual atoms of the system. The next important step is simulation environment, which gives the actual environment including optimum pressure and temperature. In general, protein simulation is done in a canonical ensemble (NVT), particularly the initial equilibrium steps, or it is done in an isothermal-isobaric (NPT) ensemble. For simulation, the protein molecule should be kept in the unit cell and solvated with a suitable explicit solvent. Several explicit water models include TIP3P, TIP4P (Jorgensen et al. 1983), TIP5P (Mahoney and Jorgensen 2001), SPC and SPC/E (Berendsen et al. 1987) are the most popular models used to imitate the specific nature and complexity of molecule hydration, including orientation of solvent dipoles and effective electrostatic shielding, subtle hydrogen bond network rearrangements, saturation of hydrophobic surface and accompanying changes in entropy.

There are two main families of MD simulation methods, classical and quantum simulation, which are distinguished based on the model chosen to represent a physical system. A basic ball-and-stick model of molecules was used in classical molecular simulation, where the atoms correspond to soft balls and elastic sticks correspond to bonds. Several force fields are widely used in the molecular simulation approach. AMBER (Case et al. 2005), NAMD (Phillips et al. 2005), CHARMM (Brooks et al. 1983) and GROMOS (Pronk et al. 2013) are widely used force fields which differ principally in the way they are parameterized, but they generally give similar results. Quantum simulation or first principle MD simulation began in 1980s with the seminal work of Car and Parinello, explicitly taking into account the quantum nature of the chemical bond. Due to the invention of high configurational computer and the advent of graphical processor unit (GPU) architectures, MD simulation software can efficiently run on innovative hardware infrastructures, surpassing alternate conventional methods. Even these methods, running on specialized hardware fails to describe the slow unbinding events. In fast-paced drug discovery programs, this is the major issue limiting the use of MD-based simulation for kinetic prediction (Borhani and Shaw 2012). However, sampling issues have led the development of several innovative algorithms that form the basis of the enhanced sampling method, speeding up the description of slow processes and accelerating the rare events characterized by high-in-free-energy states (Abrams and Bussi 2014). Sampling methods including free energy perturbation (Jorgensen and Thomas 2008), umbrella sampling, replica exchange, meta-dynamics (Laio and Parrinello 2002), steered MD (Isralewitz et al. 2001), accelerated MD (Hamelberg et al. 2004) milestoning (Faradjian and Elber 2004), transition-path sampling (Bolhuis et al. 2002), Monte Carlo sampling of conformational space, quantum mechanics/molecular mechanics (QM/MM) and molecular docking simulation are recently used methods for studying protein-ligand binding and estimating the associated energy and kinetics (Durrant and McCammon 2011; Harvey and Fabritiis 2012).

11.4.7 QM/MM Simulations

Most of biological systems such as enzymes are heavy atoms, too large to be described at any level of ab initio theory, and classical molecular mechanics force field is not sufficiently flexible to model processes in which chemical bonds are

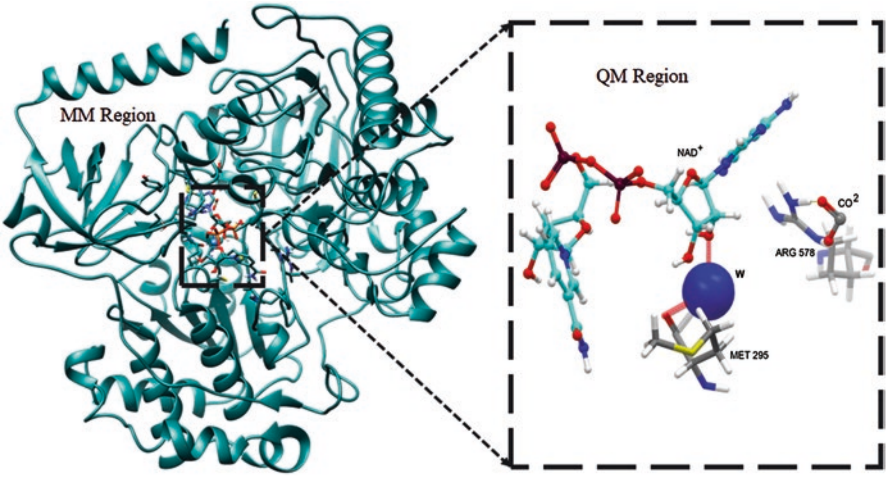


Fig. 11.6 Showing the focused QM region inside the MM region of the whole protein

broken or formed and make a proper model of the complex environment of the reaction, which involves efficient thermal averaging of the energy landscape. To overcome these issues, an alternative approach has been developed that treats a small part of the system at the level of quantum chemistry (QM) while retaining the computationally cheaper force field (MM) for the large part (Fig. 11.6).

This hybrid strategy QM/MM simulation was introduced by Warshel and Levitt and become a power full tool for the analysis of the enzyme reaction mechanism, playing a significant role in exciting applications like drug design (Gao and Truhlar 2002; Shaik et al. 2010; van der Kamp and Mulholland 2013; Lonsdale and Mulholland 2014). Basically, three classes of interaction are involved in QM/MM potential energy: interaction between atoms in the QM region, interaction between atoms in the MM region and interactions between QM and MM atoms. Quantum mechanics calculations are also an essential complement or alternative in the interpretation of outcomes of experiments by theoretical prediction of a molecular characteristic such as electrical and magnetic ones and properties related to geometrical derivatives (Cohen et al. 2012). QM treats molecules as a collection of nuclei and electrons, without any reference to chemical bonds, which is important in understanding the behavior of system at the atomic level. This method applies the lows of QM to approximate the wave function of Schrödinger equation in terms of the motions of electrons (Atkins and de Paula 2006; Tannor 2008). QM methods are a more accurate but they entail an expensive and time-consuming calculation. Calculations are employed in semi-empirical methods such as AM1 and PM3 only for valence electrons in the system. The combined QM-MM methods provide the accuracy of QM description with the low cost of MM (Lin and Truhlar 2007; Menikarachchi and Gascon 2010; Honarparvar et al. 2014). Quantum mechanics-based methods such as ab initio and the density functional theory (DFT) method fall

Table 11.4 Accuracy of different quantum mechanics methods

S. No	Types of quantum mechanics	Accuracy	Maximum atoms
1	Semi-empirical	Low	2000
2	Hartree–Fock and density functional	Medium	500
3	Perturbation and variation methods	High	50
4	Coupled cluster	Very high	20

within the approximate range of a few picometers to nanometers. These electronic structures allow accurate theoretical studies to be certain to extend to both macromolecules (synthetic polymers and proteins) and condensed matter (liquid and solids). DFT provides all the information on the system and avoids the wave function calculation. DFT is rooted in the Hoenberg–Kohn theorems, according to which the exact energy of a molecular system depends on its electron density; the latter being a function of the electronic coordinates. The total energy of a system can be calculated by the sum of several functionals such as kinetic energy, nucleus-electron potential energy, electron-electron repulsion energy and exchange-correlation functional. The choice of QM method, choice of MM force field, segregation of the system into QM and MM regions, simulation types and the advanced conformational sampling are the five important aspects of QM-MM calculation of an enzyme. The choice of QM method is crucial: there are different QM methods ranging from fast, semi-empirical methods to more accurate and more computationally expensive methods; however, not all the methods are applicable to all systems for reasons of accuracy, practicality or due to lack of parameters. The Table 11.4 shows the accuracy of different quantum methods.

11.5 Drug Delivery Approach Using Computational Methods

In drug delivery approach, potential drug molecule must have the capability to sustain its effectiveness, posing key challenges to effective drug delivery; an administered drug must penetrate obstacles such as endo or epithelial membranes and also survive the host's defenses to be effective. Hence, to overcome these challenges requires some form of drug encapsulation such as the unique molecular encapsulation architecture known as a drug delivery system (Allen et al. 2004; Blanco et al. 2015). This new approach of controlling the pharmacokinetics, thermodynamics, non-specific toxicity, immunogenicity, biorecognition and efficacy of drugs was generated to minimize drug degradation and loss and to prevent harmful side effects and increase drug bioavailability and the fraction of the drug that accumulates in the required zone (Reddy and Swarnalatha 2010). Several mechanisms are involved in a drug delivery system such as drug formulation, medical device or dosage technology to carry the drug inside the body and a mechanism for the release. Most of the commercial applications of nanoparticles in medicine are directed to drug delivery, for which several solutions have been proposed, including liposomal and lipid-based colloidal nanoDDS, nanoparticulate polymeric micelles (as drug carrier and polymer-based nanoparticulate DDS. Molecular modeling and computational chemistry provide several tools such as quantum mechanical ab initio methods, molecular dynamics, free energy perturbation and docking to quantify drug-carrier, carrier-medium and drug-medium interactions (Neumann et al. 2004).

11.6 Polymer Used as Carrier

Polymers are naturally occurring substances with high molar masses and a large number of repeating units; they play a significant role in the development of drug delivery systems by releasing both hydrophilic and hydrophobic drug molecules. Covalent bond formation of polymers with drug molecules carries the drug molecules to their respective site. Hence, there are several advantages of polymers acting as inert carriers to which a drug can be conjugated; for example, polymers improve pharmacokinetic and pharmacodynamic properties of drug molecules. Polymers is an important constituents of pharmaceutical forms such as solid dosage as in tablets and capsules; they can be dispersed in a system like a suspension, emulsion, cream or ointment; and they can be made into a particulate system, microcapsules, microparticles and nanoparticles; and they are accepted that formulation in clinical performance of pharmaceutical dosage forms (Duncan 2003; Raizada et al. 2010). The main function of a polymeric carrier is to carry and transport drug molecules to the site of action. This polymeric drug delivery system significantly protects the drug molecule from interaction with other macromolecules including proteins and nucleic acids, which could alter the chemical structure of the drug molecules. Both non-biodegradable and biodegradable polymers have been used in drug delivery systems. Based on their desirable physical properties, polymers are selected and used in both non-biological and biological settings. Polymers such as polymethyl methacrylate, polyvinyl alcohol, polyurethane and polyethylene are a few examples of polymer use in non-biological processes. In recent years, polymers have been used as carrier molecules due to their unique features such as chemical inertness, freedom from impurities, appropriate physical structure and ability to be processed readily. Polyethylene-co-vinyl acetate, polymethyl methacrylate, polyvinyl alcohol, poly-N-vinyl pyrrolidine, polyacrylic acid and polyacrylamide are often used in controlled drug delivery system (Poddar et al. 2010; Harekrishna Roy et al. 2013). Smart polymers are those having the capability to change their properties in response to the changes in biological conditions (Yang and Pierstorff 2012). Several stimuli including temperature, pressure, pH electric field, magnetic field, light, change in concentration, ionic strength and potential may influence the changes in nature of polymer properties (Schmaljohann 2006). For example, a temperature-responsive polymer brings about changes in hydrophilicity/hydrophobicity of polymers, enhancing their membrane permeation. This alteration in polymer properties can be used to allow adhesion to a cell surface, to break down a cellular membrane and to release biologically active compounds. Recently, polymers have been used for developing controlled drug release systems and sustained release formulations, which help regulate drug administration by preventing under- or overdosing. These advanced drug-releasing systems play a significant role in improving bioavailability, minimizing side effects and other types of inconveniences (Liechty et al. 2010).

11.6.1 Drug-Polymer Interaction

Most computational studies for drug delivery use molecular dynamics simulation, which mimics the natural pathway of molecular motion to sample successive configuration. Newton's law and Maxwell–Boltzmann distribution assign initial velocity of molecules at a given temperature. The interactions between molecules at each time are computed and then equations of motion are solved numerically with an appropriate time step to update the velocities and position for the next successive steps (Frenkel and Smit 2002).

In classical molecular dynamics simulations, the interaction of molecules can be described by a force field with certain functional forms and several parameters. A force field such as AMBER (Cornell et al. 1995), OPLS (Jorgensen et al. 1996) and CHARMM (Mackerell et al. 1998) is widely used to study polymer and peptide drug interactions. Interactions such as hydrogen bonding (Zhang et al. 2012; Miyazaki et al. 2011), dipole-dipole interaction (Marsac et al. 2009; Khougaz and Clas 2000), ionic interaction (Yoo et al. 2009; Kindermann et al. 2011) and van der Waals interaction (Marsac et al. 2009) generally occur between drug and polymer. Dissipative particle dynamics (DPD) is a widely used mesoscale simulation for identifying and defining chemically distinct components and defining interaction parameters between various chemical species. In this model, a fluid system is simulated using a set of interacting particles. Each particle represents a cluster of small molecules instead of a single molecule. Drug, polymer, surfactant and solvent are represented as distinct bead types. Polymer bead number length is determined by

$$N_{\mathrm{DPD}} = \frac{Mp}{Mm\,C\infty},$$

where Mp is polymer molecular weight, Mm monomer molecular weight and $C\infty$ polymer characteristic ratio. However, a detailed mechanism on drug-polymer interactions is lacking, such as how chemically substituted cellulosic polymers interact with drug molecules at a molecular level and how different structural variables such as molecular weight and substitution pattern affect the drug-polymer interaction. In addition to the classical MD and DPD, another two levels of molecular models such as coarse-grained molecular dynamics (CG-MD) simulations, which are used to model excipients such as modified cellulosic polymers at a monomer level resolution and drugs at a similar level. The full spectrum of the CG-MD approach contains contributions from several different fields and continuum transport modeling, in which diffusion equations for transport of polymer, drug and solvent through a capsule are determined by solving the relevant differential equation. Several software packages can integrate these equations, including the popular GROMACS (Van der Spoel et al. 2005), NAMD (Phillips et al. 2005), CHARMM (Klauda et al. 2010) and AMBER (Wang et al. 2004) packages. Many of the coarse-grained methods utilize one of these integrators to perform simulations.

11.7 Computational Methods Used in Toxicity Studies

Toxicity is a measurement of the adverse effect of chemicals, and specific types of these adverse effects are known as toxicity endpoints, for example, carcinogenicity or genotoxicity. These adverse effects can be quantitatively or qualitatively measured to identify harmful effects caused by substances on humans and animals (Rowe et al. 2010). A number of factors determine the toxicity of chemicals, including route of exposure, dose, duration of exposure, ADME properties (absorption, distribution, metabolism and excretion), biological properties and chemical properties (Raies and Bajic 2016). A number of in vitro models have been used to determine toxicity such as high throughput screening (AltTox) and in vivo animal models. Recently, computational toxicity methods have been widely used to potentially minimize the need for animal testing and reduce the cost and time of the toxicity test to improve toxicity prediction and safety assessment. The major advantage of computational toxicity methods is their ability to estimate chemicals for toxicity even before they are synthesized (Madan et al. 2013). In silico toxicology analysis encompasses a wide range of computational tools including database storage of chemical data, their toxicity and chemical properties, and software for generating molecular descriptors, simulation tools for systems biology and molecular dynamics and modeling methods for toxicity. Rule-based and structural alerts are often-used computational methods for determining toxicity based on chemical properties and how drugs should be altered to reduce their toxicity. Another method, read-across, is used to predicting the unknown toxicity of a chemical through the use of similar chemicals (analogs) with known toxicity from the same chemical category (Dimitrov and Mekenyan 2010; Modi et al. 2012; Benigni et al. 2013; Venkatapathy and Wang 2013;). There are two approaches—an analog, or one-to-one approach, and a category, or many-to-one approach—for developing a read-across method. Both approaches are quite sensitive, identifying similar chemicals by calculating their properties and the similarities between them. The main advantage of read-across is its transparency (Cronin 2011): it is easy to interpret and implement (Enoch 2009), and it can model quantitative and qualitative toxicity endpoints and allow for a wide range of types of descriptors and similarity measures to be used to express similarity between chemicals (Dimitrov and Mekenyan 2010).

Quantitative structure-activity relationship (QSAR) is another widely used method that employs molecular descriptors to predict a chemical's toxicity. Generally, the QSAR method predicts toxicity (T) of a lead molecule using a vector feature of chemical properties (θp) and a function f that calculates T given θp is

$$T = f\left(\theta \mathrm{p}\right).$$

There are two QSAR models: local QSAR, which is generated from congeneric chemicals, and global QSAR, which is made from diverse chemicals. Local QSAR is used to predict toxicity based on the mode of action of specific chemicals, hence, local QSAR are more accurate as they are customized for specific chemicals (Valerio 2009). Mainly two basic steps are involved in the development of a QSAR model: the generation of molecular descriptors and then of models to fit the data. The number of molecular descriptors, as based on simulated annealing, generic algorithm or principal component analysis, can be used to determine the chemicals (Deeb and Goodarzi 2012; Devillers 2013). If there are a small number of descriptors, using two-dimensional scatterplots of each descriptor versus its biological activity can help identify significant descriptors (Devillers 2013). There are many tools available that provide pre-built QSAR model such as OECD QSAR Toolbox (OECD 2015), TopKat (Accelrys 2015) and METEOR (Lhasa Limited, Meteor Nexus 2014). The major advantage of QSAR is that it's easy to interpret and it can model categorical and continuous toxicity endpoints and molecular descriptors and toxic and non-toxic chemicals. However, it may not be always employable, as a large number of chemicals are needed in the model development for QSAR to achieve statistical significance (Valerio 2009; Deeb and Goodarzi 2012).

Pharmacokinetic (PK) models relate to the concentration of drug molecules in tissues to time, estimating the amount of chemicals in different parts of the body and quantifying ADME (absorption, distribution, metabolism and excretion) processes (Jack et al. 2013; Sung et al. 2014). Mainly, the PK models are used to relate chemical concentration in a part of the body to time of toxic responses. A PK model can be categorized as two models: compartment and non-compartmental (Sung et al. 2014). A compartment model consists of one more compartments, with each compartment represented by differential equations (Sung et al. 2014). One compartment model represents the whole body as a single compartment, assuming rapid equilibrium of chemical concentration within the body but not considering the time to distribute of the chemical. Two-compartment models consist of two compartments, the central and peripheral with both compartments represented by differential equations. These models provide mechanistic insight based on pharmacokinetic models including concentration and time, physiological descriptors of tissues and ADME processes such as volumes, blood flows, chemical binding/partitioning, metabolism and excretion (Jack et al. 2013; El-Masri 2013).

11.8 Outcome of Drug Research in Bacterial Inhibitors

Bacterial infection is one of the major threats to human health because it frequently causes severe diseases not only in the form of primary agents but also after pathologies caused by other agents. Compared to Gram-negative bacteria, Gram-positive bacteria have a much thicker peptidoglycan layer, which is responsible for the increasing occurrence of bacterial resistance to antibiotics in medicinal practice (Springer et al. 2010; Nikaido 2003). Since the discovery of several antibiotics in the mid-twentieth century, resistance has been a concern (Peters et al. 2008). Although the emergence of antibacterial resistance is not new, it continues to be a major health concern. The report from the Centers for Disease Control and Prevention on antimicrobial resistance revealed that more than 21% of hospital-acquired infections were caused by an antimicrobial resistant pathogen. Hence, there is a need for new alternatives in the treatment of infections by multi- resistant bacteria. Among the several pathogens, *Staphylococcus aureus* resistant to methicillin (MRSA), *Streptococcus pneumonia*, resistant to penicillin, glycopeptide-intermediately-resistant *S. aureus* (GISA), methicillin-resistant *S. epidermis*, glycopeptide-resistant *enterococcus spp* and vancomycin-resistant *Enterococci* (VRE) are the more important etiological agents of hospital and community infections and are responsible for high rates of morbidity and mortality in hospitalized patients (Woodfor and Livermore 2009; Livermore 2009; Arias and Murray 2009). Several fluoroquinolones, ramoplanin, beta-lactams and the quinupristin/dalfopristin are currently used in the market. Moellering et al. (1999) studied the clinical efficacy and safety of quinupristin-dalfopristin in the treatment of a patient with a vancomycin-resistant infection. From the studies it was noted that the overall clinical and bacteriologic success rate was 66%. In another study, Nichols et al. (1999) compared quinupristin-dalfopristin with cefazolin, oxacillin and vancomycin in two randomized, open-label clinical trials.

Oxazolidinones, an antimicrobial class of agents, are a unique family of drug molecule possessing activity against *Staphylococcus aureus* and glycopeptide-intermediately-resistant *S. aureus* (Rybak et al. 2000; Wootton et al. 2000) and they are also more effective against a wide range of Gram-positive bacteria and *Mycobacterium tuberculosis*. Linezolid was the first approved derivative with acceptable tolerability in humans for the treatment of pneumonia, skin and soft tissues infections caused by VRE (Cammarata et al. 2000). Daptomycin is another antibacterial agent used to treat a wide range of Gram-positive bacteria. Recent studies from the US and Europe revealed that daptomycin was active against all *Staphylococcus aureus* and Gram-negative bacteria such as Leuconostoc, which are characteristically resistant to glycopeptides (Barry et al. 2001; King and Phillips 2001). The effectiveness of daptomycin has been proved in various animal models of Gram-positive infection. Several global randomized, double blind phase II trials have investigated the efficacy of daptomycin in the treatment of community-acquired pneumonia (Pertel et al. 2008).

11.9 Future Aspects of Computational Methods in Targeting Bacterial Infections

The drug-resistant capability of Gram-positive bacteria is a serious issue in clinical practice, and several antibacterial agents have already been approved by the US Federal Drug Administration for several infections, while other agents are still undergoing clinical trials. However, a lack of effective antibiotics in development implies that future treatment strategies for the resistant bacteria may have to show enhanced therapeutic efficacy. The battle against antibiotic resistance can be carried out on two fronts: either in advancing research efforts toward the discovery of novel and potential agents or by enhancing the effectiveness of the currently available ones. With the increasing prevalence of bacterial resistance, there is need to identify potential lead molecules to combat them. Conventional drug development research requires huge investment and at least 12–15 years experimentation, and even so, it often does not reach the market; hence; alternative approaches and strategies are required to develop safe and effective novel antimicrobial therapies. The current scenario of antibiotic research and development is not very effective, so a computational approach such as structure-based drug design, ligand-based drug design, pharmacophore modeling and molecular docking are useful for understanding the mechanism of bacterial resistance to antibiotics. In addition to the experimental approach, computational biology combination therapy has great potential in the future discovery of antimicrobial drugs.

References

Abel R, Young T, Farid R, Berne BJ, Friesner RA (2008) Role of the active-site solvent in the thermodynamics of factor Xa ligand binding. J Am Chem Soc 130:2817–2831

Abrams C, Bussi G (2014) Enhanced sampling in molecular dynamics using metadynamics, replica-exchange, and temperature-acceleration. Entropy 16:163–199

Accelrys Available Chemicals Directory (ACD) (2012) Accelrys, Inc., San Diego, CA

Accelrys. TOPKAT (TOxicity Prediction by Komputer Assisted Technology) (2015). Available at: http://accelrys.com/products/discovery-studio/admet.html

Acharya C, Coop A, Polli JE, MacKerell AD (2011) Recent advances in ligand-based drug design: relevance and utility of the conformationally sampled pharmacophore approach. Curr Comput Aided Drug Des 7:10–22

Adcock SA, McCammon JA (2006) Molecular dynamics: survey of methods for simulating the activity of proteins. Chem Rev 106:1589–1615

Agrafiotis DK, Gibbs AC, Zhu F, Izrailev S, Martin E (2007) Conformational sampling of bioactive molecules: a comparative study. J Chem Inf Model 47:1067–1086

Akamatsu M (2002) Current state and perspectives of 3D-QSAR. Curr Top Med Chem 2:1381–1394

Albrich WC, Monnet DL, Harnarth S (2004) Antibiotic selection pressure and resistance in Streptococcus pneumoniae and Streptococcus pyogenes. Emerg Infect Dis 10:514–517

Allen TD, Eby LT, Poteet ML, Lentz E, Lima L (2004) Career benefits associated with mentoring for protégeé: a meta-analysis. J Appl Psychol 89:127–136

Allen GD, Breshears DD, McDowell NG (2015) On underestimation of global vulnerability to tree mortality and forest die-off from hotter drought in the Anthropocene. Ecosphere 6:1–55

AltTox. Toxicity testing overview. Available at: http://alttox.org/mapp/toxicity-testing-overview/

An J, Totrov M, Abagyan R (2005) Pocketome via comprehensive identification and classification of ligand binding envelopes. Mol Cell Proteomics 4:752–761

Anandakrishnan R, Aguilar B, Onufriev AV (2012) H++ 3.0: automating pK prediction and the preparation of biomolecular structures for atomistic molecular modeling and simulation. Nucleic Acids Res 40:W537–41–W541

Arias CA, Murray BE (2009) Antibiotic-resistant bugs in the 21 st century: a super-clinical challenge. N Engl J Med 360:439–443

Ashrafuzzaman M (2014) Aptamers as both drugs and drug-carriers. Biomed Res Int 2014:697923

Atkins P, de Paula J. (2006) Atkins' Physical Chemistry. 8th ed. New York: Macmillan.

Baker D, Sali A (2001) Protein structure prediction and structural genomics. Science 294(5540):93–96

Ballester PJ, Mitchell JBO (2010) A machine learning approach to predicting protein-ligand binding affinity with applications to molecular docking. Bioinformatics (Oxford, England) 26:1169–1175

Barry AL, Fuchs PC, Brown SD (2001) In vitro activities of daptomycin against 2,789 clinical isolates from 11 North American medical centers. Antimicrob Agents Chemother 45:1919–1922

Benigni R, Battistelli CL, Bossa C, Colafranceschi M, Tcheremenskaia O (2013) Mutagenicity, carcinogenicity, and other end points. In: Reisfeld B, Mayeno AN (eds) Computational toxicology, vol 930. Humana Press, New York, pp 67–98

Beno B, Mason J (2001) The design of combinatorail libraries using properties and 3D-pharmacophore fingerprinits. Drug Discov Today 6:251–258

Berendsen HJC, Grigera JR, Straatsma TP (1987) The missing term in effective pair potentials. J Phys Chem 91(24):6269–6271

Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE (2000) The protein data bank. Nucleic Acids Res 28:235–242

Bernard D, Coop A, MacKerell AD (2005) Conformationally sampled pharmacophore for peptidic delta opioid ligands. J Med Chem 48:7773–7780

Binkowski T, Naghibzadeh S, Liang J (2003) CASTp: computed atlas of surface topography of proteins. Nucleic Acids Res 31:3352–3355

Blanco E, Shen H, Ferrari M (2015) Principles of nanoparticle design for overcoming biological barriers to drug delivery. Nat Biotechnol 33:941–951

Bleicher KH, Bohm HJ, Muller K, Alanine AI (2003) Hit and lead generation: beyond high-throughput screening. Nat Rev Drug Discov 2:369–378

Bolhuis PG, Chandler D, Dellago C, Geissler PL (2002) Transition path sampling: throwing ropes over rough mountain passes, in the dark. Annu Rev Phys Chem 53:291–318

Borhani DW, Shaw DE (2012) The future of molecular dynamics simulations in drug discovery. J Comput Aided Mol Des 26:15–26

Brady GP, Stouten PFW (2000) Fast prediction and visualization of protein binding pockets with PASS. J Comput Aided Mol Des 14:383–401

Briggs JM, Marrone TJ, McCammon JA (1996) Computational science new horizons and relevance to pharmaceutical design. Trends Cardiovasc Med 6:198–206

Bronzwaer SL, Bronzwaer SL, Cars O, Buchholz U, Mölstad S, Goettsch W, Veldhuijzen IK, Kool JL, Sprenger MJ, Degener JE (2002) A European study on the relationship between antimicrobial use and antimicrobial resistance. Emerg Infect Dis 8:278–282

Brooks BR, Bruccoleri RE, Olafson BD, States DJ, Swaminathan S, Karplus M (1983) CHARMM – a program for macromolecular energy, minimization, and dynamics calculations. J Comput Chem 4:187–217

Bryant SH, Lawrence CE (1993) An empirical energy function for threading protein sequence through the folding motif. Proteins 16(1):92–112

Burger A, Abraham DJ (2006) Burger's medicinal chemistry and drug discovery, drug discovery and drug development; Wiley, 2003. In: Langer T, Hoffmann RD (eds) Pharmacophores and Pharmacophore Searches. Wiley-VCH Verlag GmbH & Co, Weinheim

Cammarata SK, Timm JA, Hempsall KA, Todd WM, Oliphant TH, Hafkin B (2000) Efficacy of linezolid in pneumonia due to penicillin intermediate and resistant Streptococcus pneumoniae. [abstract SO.OlO]. In: Programs and abstracts of the 9th International Congress on Infectious Diseases; Buenos Aires, Argentina, April 10–13. International Society of Infectious Diseases, Boston

Carlson HA (2002) Protein flexibility and drug design: how to hit a moving target. Curr Opin Chem Biol 6:447–452

Case DA, Cheatham TE 3rd, Darden T, Gohlke H, Luo R (2005) The Amber biomolecular simulation programs. J Comput Chem 26:1668–1688

CASP (1999). Proteins, Structure, Function, and Genetics. CASP3 Proceeding 37:1–237

Cato S (2000) Pharmacophore perception, development, and use in drug design. International University Line, La Jolla, CA, Chapter Exploring Pharmacophores with Chem-X,, pp 107–125

Chang C, Swaan PW (2006) Computational approaches to modeling drug transporters. Eur J Pharm Sci 27:411–424

Chen J, Swamidass SJ, Dou Y, Bruand J, Baldi P (2005) ChemDB: a public database of small molecules and related chemoinformatics resources. Bioinformatics 21(22):4133–4139

Chen JH, Linstead E, Swamidass SJ, Wang D, Baldi P (2007) ChemDB update-full-text search and virtual chemical space. Bioinformatics 23:2348–2351

Chen VB, Arendall WB, Headd JJ, Keedy DA, Immormino RM, Kapral GJ, Murray LW, Richardson SJ, Richardson DC (2010) MolProbity: allatom structure validation for macromolecular crystallography. Acta Crystallogr D Biol Crystallogr D66:12–21

Clark DE (2006) What has computer-aided molecular design ever done for drug discovery? Expert Opin Drug Discovery 1:103–110

Clark DE (2008) What has virtual screening ever done for drug discovery? Expert Opin Drug Discovery 3:841–851

Claussen H, Buning C, Rarey M, Lengauer T (2001) FlexE: efficient molecular docking considering protein structure variations. J Mol Biol 308:377

Cohen AJ, Mori-Sánchez P, Yang W (2012) Challenges for density functional theory. Chem Rev 112:289–320

Collignon P, Powers JH, Chiller TM, Aidara-Kane A, Aarestrup FM (2009) World Health Organization ranking of antimicrobials according to their importance in human medicine: a critical step for developing risk management strategies for the use of antimicrobials in food production animals. Clin Infect Dis 49:132–141

Corbeil CR, Williams CI, Labute P (2012) Variability in docking success rates due to dataset preparation. J Comput Aided Mol Des 26:775–786

Cornell WD, Cieplak P, Bayly CI, Gould IR, Merz KM, Ferguson DM, Spellmeyer DC, Fox T, Caldwell JW, Kollman PA (1995) A 2nd generation force-field for the simulation of proteins, nucleic acids and organic molecules. J Am Chem Soc 117:5179–5197

Cronin MTD (2011) In silico tools for toxicity prediction. In: Wilson AGE (ed) New horizons in predictive toxicology: current status and application. Royal Society of Chemistry, Cambridge

D'Alfonso G, Tramontano A, Lahm A (2001) Structural conservation in single-domain proteins: implications for homology modeling. J Struct Biol 134:246–256

Davis IW, Leaver-Fay A, Chen VB, Block JN, Kapral GJ, Wang X, Murray LW, Arendall WB, Snoeyink J, Richardson JS, Richardson DC (2007) MolProbity: all-atom contacts and structure validation for proteins and nucleic acids. Nucleic Acids Res 35:W375–W383

Deeb O, Goodarzi M (2012) In silico quantitative structure toxicity relationship of chemical compounds: some case studies. Curr Drug Saf 7:289–297

Devillers J (2013) Methods for building QSARs. In: Reisfeld B, Mayeno AN (eds) Computational toxicology, vol 930. Humana Press, New York, pp 3–27

Dias R, de Azevedo WF Jr (2008) Molecular docking algorithms. Curr Drug Targets 9(12):1040–1047

DiMasi JA, Hansen RW, Grabowski HG (2003) The price of innovation: new estimates of drug developmentcosts. J Health Econ 22:151–185

Dimitropoulos D, Ionides J, and Henrick K (2006) Using PDBeChem to Search the PDB Ligand Dictionary, in Current Protocols in Bioinformatics; John Wiley & Sons, 14.13.11–14.13.13.

Dimitrov S, Mekenyan O (2010) An introduction to readacross for the prediction of the effects of chemicals. In: Cronin MTD, Madden JC (eds) In silico toxicology: principles and applications. The Royal Society of Chemistry, Cambridge, pp 372–384

Doernberg SB, Lodise TP, Thaden JT, Munita JM, Cosgrove SE, Arias CA, Boucher HW, Corey GR, Lowy FD, Murray B, Miller LG, Holland TL (2017) Gram-positive bacterial infections: research priorities, accomplishments, and future directions of the antibacterial resistance leadership group. Clin Infect Dis 64:S24

Dolinsky TJ, Czodrowski P, Li H, Nielsen JE, Jensen JH, Klebe G, Baker NA (2007) PDB2PQR: expanding and upgrading automated preparation of biomolecular structures for molecular simulations. Nucleic Acids Res 35:W522–W525

Duncan R (2003) The dawning era of polymer therapeutics. Nat Rev Drug Discov 2:347–360

Durrant J, McCammon JA (2011) Molecular dynamics simulations and drug discovery. BMC Biol 9:71

Ekins S, Mestres J, Testa B (2007) In silico pharmacology for drug discovery: methods for virtual ligand screening and profiling. Br J Pharmacol 152:9–20

El-Masri H (2013) Modeling for regulatory purposes (risk and safety assessment). In: Reisfeld B, Mayeno AN (eds) Computational toxicology, vol 930. Humana Press, New York, pp 297–303

Enoch SJ (2009) Chemical category formation and readacross for the prediction of toxicity. In: Puzyn T, Leszczynski J, Cronin MT (eds) Recent advances in QSAR studies, vol 8. Springer, Dordrecht, pp 209–219

Enyedy IJ, Sakamuri S, Zaman WA, Johnson KM, Wang S (2003) Pharmacophore-based discovery of substituted pyridines as novel dopamine transporter inhibitors. Bioorg Med Chem Lett 13:513–517

Epik v2.2 (2011) Portland, OR: Schrödinger, Inc

Ernst A, Obrecht D (2008) Case study of parallel synthesis in hit identification, hit exploration, hit-to-lead, and lead-optimization programs in high-throughput lead optimization in drug discovery. In: Kshirsagar T (ed) CRC Press, Boca Raton, USA, pp 99–116

Fanelli F, Gohlke H, Kuhn LA, Morris GM, Orozco M, Pertinhez TA, Rizzi M, Sotriffer C (2008) Target flexibility: an emerging consideration in drug discovery and design. J Med Chem 51:6237–6255

Fang Y (2012) Ligand-receptor interaction platforms and their applications for drug discovery. Expert Opin Drug Discovery 7:969–988

Faradjian AK, Elber R (2004) Computing time scales from reaction coordinates by milestoning. J Chem Phys 120:10880–10889

Faraggi E, Xue B, Zhou Y (2009) Improving the prediction accuracy of residue solvent accessibility and real-value backbone torsion angles of proteins by guided-learning through a two-layer neural network. Proteins 74:847–856

Fiser A (2010) Template-based protein structure modeling. Methods Mol Biol (Clifton, NJ) 673:73–94

Floudas CA (2007) Computational methods in protein structure prediction. Biotechnol Bioeng 97:207–213

Floudas C, Fung H, McAllister S, Moennigmann M, Rajgaria R (2006) Advances in protein structure prediction and de novo protein design: a review. Chem Eng Sci 61:966

Flower DR (2002) Drug design: cutting edge approaches. Royal Society of Chemistry, Cambridge

Frenkel D, Smit B (2001) Understanding molecular simulation. Academic Press, Inc, San Diego, p 638

Frenkel D, Smit B (2002) Understanding molecular simulation: from algorithms to applications. Academic Press, San Diego

Friesner RA, Banks JL, Murphy RB, Halgren TA, Klicic JJ, Mainz DT, Repasky MP, Knoll EH, Shelley M, Perry JK, Shaw DE, Francis P, Shenkin PS (2004) Glide: a new approach for rapid, accurate docking and scoring. 1. Method and assessment of docking accuracy. J Med Chem 47:1739–1749

Gao JL, Truhlar DG (2002) Quantum mechanical methods for enzyme kinetics. Annu Rev Phys Chem 53:467–505

Gehlhaar DK, Verkhivker GM, Rejto PA, Sherman CJ, Fogel DB, Freer ST (1995) Molecular recognition of the inhibitor AG-1343 by HIV-1 protease: conformationally flexible docking by evolutionary programming. Chem Biol 2:317–324

Gehlhaar DK, Bouzida D, Rejto PA (1999) In: Parrill L, Reddy MR (eds) In rational drug design: novel methodology and practical applications, vol 719. American Chemical Society, Washington, DC, pp 292–311

Geppert H, Vogt M, Bajorath J (2010) Current trends in ligand-based virtual screening: molecular representations, data mining methods, new application areas, and performance evaluation. J Chem Inf Model 50:205–216

Gilson MK, Zhou HX (2007) Calculation of protein-ligand binding affinities. Annu Rev Biophys Biomol Struct 36:21–42

Giske CG, Monnet DL, Cars O, Carmeli Y (2008) Clinical and economic impact of common multidrug-resistant Gram-negative bacilli. Antimicrob Agents Chemother 52:813–821

Göbl C, Madl T, Simon B, Sattler M (2014) NMR approaches for structural analysis of multidomain proteins and complexes in solution. Prog Nucl Magn Reson Spectrosc 80:26–63

Gohda K, Mori I, Ohta D, Kikuchi T (2000) A CoMFA analysis with conformational propensity: an attempt to analyze the SAR of a set of molecules with different conformational flexibility using a 3D-QSAR method. J Comput Aided Mol Des 14:265–275

Gohlke H, Klebe G (2001) Statistical potentials and scoring functions applied to protein-ligand binding. Curr Opin Struct Biol 11:231–235

González PM, Acharya C, MacKerell AD Jr, Polli JE (2009) Inhibition requirements of the human apical sodium-dependent bile acid transporter (hASBT) using aminopiperidine conjugates of glutamyl-bile Acids. Pharm Res 26:1665–1678

Goto E, Yagi Y, Matsumoto Y, Tsubota K (2002) Impaired functional visual acuity of dry eye patients. Am J Ophthalmol 133:181–186

Guedes IA, de Magalhães CS, Dardenne LE (2014) Receptor-ligand molecular docking. Biophys Rev 6:75–87

Guimaraes CRW, Cardozo M (2008) MM-GB/SA rescoring of docking poses in structure-based lead optimization. J Chem Inf Model 48:958–970

Hajduk PJ, Huth JR, Fesik SW (2005) Druggability indices for protein targets derived from NMR-based screening data. J Med Chem 48:2518–2525

Halgren T (2007) New method for fast and accurate binding-site identification and analysis. Chem Biol Drug Des 69:146–148

Halgren TA (2009) Identifying and characterizing binding sites and assessing druggability. J Chem Inf Model 49:377–389

Halgren TA, Murphy RB, Friesner RA, Beard HS, Frye LL, Pollard WT, Banks JL (2004) Glide: a new approach for rapid, accurate docking and scoring. 2. Enrichment factors in database screening. J Med Chem 47:1750–1759

Hamelberg D, Mongan J, McCammon JA (2004) Accelerated molecular dynamics: a promising and efficient simulation method for biomolecules. J Chem Phys 120:11919–11929

Hanson SM, Newstead S, Swartz KJ, Sansom MS (2015) Capsaicin interaction with TRPV1 channels in a lipid bilayer: molecular dynamics simulation. Biophys J 108:1425–1434

Harvey MJ, De Fabritiis G (2012) High-throughput molecular dynamics: the powerful new tool for drug discovery. Drug Discov Today 17:1059–1062

Hawkey PM, Jones AM (2009) The changing epidemiology of resistance. J Antimicrob Chemother 64:i3–i10

Hendlich M, Rippmann F, Barnickel G (1997) LIGSITE: automatic and efficient detection of potential small moleculebinding sites in proteins. J Mol Graph Model 15:359–363

Herrling PL (2005) The drug discovery process. Prog Drug Res 62:1–14

Hileman B (2006) Accounting for R&D, Many doubt the $800 million pharmaceutical price tag. Chemical Eng News 84:50–51

Honarparvar B, Kruger HG, Maguire GE, Govender T, Soliman ME (2014) Integrated approach to structure-based enzymatic drug design: molecular modeling, spectroscopy, and experimental bioactivity. Chem Rev 114:493–537

Hoofnagle AN, Resing KA, Ahn NG (2003) Protein analysis by hydrogen exchange mass spectrometry. Annu Rev Biophys Biomol Struct 32:1–25

Hooft R, Vriend WG, Sander C, Abola EE (1996) Errors in protein structures. Nature 381:272

Hristozov D, Oprea TI, Gasteiger J (2007) Ligand-based virtual screening by novelty detection with self-organizing maps. J Chem Inf Model 47(6):2044–2062

Huang SY, Zou X (2006a) An iterative knowledge-based scoring function to predict protein-ligand interactions: I. Derivation of interaction potentials. J Comput Chem 27:1866–1875

Huang SY, Zou X (2006b) An iterative knowledge-based scoring function to predict protein-ligand interactions: II. Validation of the scoring function. J Comput Chem 27:1876–1882

Huang X, Lin J, Demner-Fushman D (2006) Evaluation of PICO as a knowledge representation for clinical questions. AMIA Annual Symposium proceedings. AMIA Symposium 2006:359–363

Irwin JJ, Shoichet BK (2005) ZINC - a free database of commercially available compounds for virtual screening. J Chem Inf Model 45:177–182

Isralewitz B, Gao M, Schulten K (2001) Steered molecular dynamics and mechanical functions of proteins. Curr Opin Struct Biol 11:224–230

Jack J, Wambaugh J, Shah I (2013) Systems toxicology from genes to organs. In: Reisfeld B, Mayeno AN (eds) Computational toxicology, vol 930. Humana Press, New York, pp 375–397

Jain AN (2003) Surflex: fully automatic flexible molecular docking using a molecular similarity-based search engine. J Med Chem 46:499–511

Jain AN (2006) Scoring functions for protein-ligand docking. Curr Protein Pept Sci 7:407–420

Jones DT, Taylor WR, Thornton JM (1992a) The rapid generation of mutation data matrices from protein sequences. Comput Appl Biosci 3:275–282

Jones DT, Taylor WR, Thornton JM (1992b) A new approach to protein fold recognition. Nature 358:86–89

Jones G, Willett P, Glen RC, Leach AR, Taylor R (1997) Development and validation of a genetic algorithm for flexible docking. J Mol Biol 267:727–748

Jorgensen WL, Thomas LL (2008) Perspective on free-energy perturbation calculations for chemical equilibria. J Chem Theory Comput 4:869–876

Jorgensen WL, Tirado-Rives J (2005) Molecular modeling of organic and biomolecular systems using BOSS and MCPRO. J Comput Chem 26:1689–1700

Jorgensen WL, Chandrasekhar J, Madura JD, Impey RW, Klein ML (1983) Comparison of simple potential functions for simulating liquid water. J Chem Phys 79:926–935

Jorgensen WL, Maxwell DS, Tirado-Rives J (1996) Development and testing of the OPLS all atom force field on conformational energetics and properties of organic liquids. J Am Chem Soc 118:11225–11236

Jubb H, Blundell TL, Ascher DB (2015) Flexibility and small pockets at protein–protein interfaces: new insights into druggability. Prog Biophys Mol Biol 119:2–9

Kahsai AW, Xiao K, Rajagopal S, Ahn S, Shukla AK, Sun J, Oas TG, Lefkowitz RJ (2011) Multiple ligand-specific conformations of the β2-adrenergic receptor. Nat Chem Biol 7:692–700

Kalyaanamoorthy S, Chen YP (2011) Structure-based drug design to augment hit discovery. Drug Discov Today 16:831–839

Karchin R, Cline M, Mandel-Gutfreund Y, Karplus K (2003) Hidden Markov models that use predicted local structure for fold recognition: alphabets of backbone geometry. Proteins 51:504–514

Karnachi P, Kulkarni A (2006) Application of pharmacophore fingerprints to structure based design and data mining. In: Langer T, Hoffmann RD (eds) Pharmacophores and pharmacophore searches. Wiley-VCH, Weinheim

Karplus K, Karchin R, Draper J, Casper J, Mandel-Gutfreund Y, Diekhans M, Hughey R (2003) Combining local-structure, fold-recognition and new fold methods for protein structure prediction. Proteins 53:491–496

Kaserer T, Beck KR, Akram M, Odermatt A, Schuster D (2015) Pharmacophore models and pharmacophore-based virtual screening: concepts and applications exemplified on hydroxysteroid dehydrogenases. Molecules 20:22799–22832

Kaveti S, Engen JR (2006) Protein interactions probed with mass spectrometry. Methods Mol Biol 316:179–197

Kelley LA, MacCallum RM, Sternberg MJ (2000) Enhanced genome annotation using structural profiles in the program 3d-pssm. J Mol Biol 299:499–520

Khougaz K, Clas SD (2000) Crystallization inhibition in solid dispersions of MK-0591 and poly(vinylpyrrolidone) polymers. J Pharm Sci 89:1325–1334

Kindermann C, Matthee K, Strohmeyer J, Sievert F, Breitkreutz J (2011) Tailor-made release triggering from hot-melt extruded complexes of basic polyelectrolyte and poorly watersoluble drugs. Eur J Pharm Biopharm 79:372–381

King A, Phillips I (2001) The in vitro activity of daptomycin against 514 Gram-positive aerobic clinical isolates. J Antimicrob Chemother 48:219–223

Kitchen DB, Decornez H, Furr JR, Bajorath J (2004) Docking and scoring in virtual screening for drug discovery:methods and application. Nat Rev Drug Discov 3:935–949

Klauda JB, Venable RM, Freites JA, O'Connor JW, Tobias DJ, Mondragon-Ramirez C, Vorobyov I, MacKerell AD, Pastor RW (2010) Update of the CHARMM all-atom additive force field for lipids: validation on six lipid types. J Phys Chem B 114:7830–7843

Klebe G (2006) Virtual ligand screening: strategies, perspectives and limitations. Drug Discov Today 11:580–594

Klebe G, Abraham U, Mietzner T (1994) Molecular similarity indices in a comparative analysis (CoMSIA) of drug molecules to correlate and predict their biological activity. J Med Chem 37:4130–4146

Kollman PA (1993) Free energy calculations: applications to chemical and biochemical phenomena. Chem Rev 93:2395–2417

Kountouris P, Hirst JD (2009) Prediction of backbone dihedral angles and protein secondary structure using support vector machines. BMC Bioinformatics 10:437

Kovalenko A (2003) Three-dimensional RISM theory for molecular liquids and solid-liquid interfaces. In: Hirata F, Mezey PG (eds) 360, vol 24. Kluwer Academic Publishers, Dordrecht, pp 169–275

Kramer MS, Cutler N, Feighner J, Shrivastava R, Carman J, Sramek JJ, Reines SA, Liu G, Snavely D, Wyatt-Knowles E, Hale JJ, Mills SG, MacCoss M, Swain CJ, Harrison T, Hill RG, Hefti F, Scolnick EM, Cascieri MA, Chicchi GG, Sadowski S, Williams AR, Hewson L, Smith D, Carlson EJ, Hargreaves RJ, Rupniak NM (1998) Distinct mechanism for antidepressant activity by blockade of central substance P receptors. Science 281:1640–1645

Kuntz ID, Blaney JM, Oatley SJ, Langridge R, Ferrin TE (1982) A geometric approach to macromolecule-ligand interactions. J Mol Biol 161:269–288

Laio A, Parrinello M (2002) Escaping free-energy minima. Proc Natl Acad Sci U S A 99: 12562–12566

Langer T, Hoffmann RD (2006) Pharmacophore modeling: application in drug discovery. Expert Opin Drug Discovery 1:261–267

Laskowski RA (1995 Oct) SURFNET: a program for visualizing molecular surfaces, cavities, and intermolecular interactions. J Mol Graph 13(5):323–330, 307–308

Laskowski RA, Moss DS, Thornton JM (1993) Main-chain bond lengths and bond angles in protein structures. J Mol Biol 231:1049–1067

Leach AR, Hann MM (2000) The in silica world of virtual libraries. Drug Discov Today 5:326–336

Levitt DG, Banaszak LJ (1992 Dec) POCKET: a computer graphics method for identifying and displaying protein cavities and their surrounding amino acids. J Mol Graph 10(4):229–234

Lewis K (2013) Platforms for antibiotic discovery. Nat Rev Drug Discov 12:371–387

Lewis RA, Pickett SD, Clark DE (2000) Computer-aided molecular diversity analysis and combinatorial library design. Rev Comput Chem 16:1–51

Lhasa Limited. Meteor Nexus (2014). Available at: http://www.lhasalimited.org

Li H, Robertson AD, Jensen JH (2005) Very fast empirical prediction and rationalization of protein pKa values. Proteins 61:704–721

Liechty WB, Kryscio DR, Brandon VS, Peppas NA (2010) Polymers for drug delivery. Annu Rev Chem Biomol Eng 1:149–173

LigPrep v2.5 (2011) Portland, OR: Schrödinger, Inc

Lin H, Truhlar DG (2007) QM/MM: what have we learned, where are we, and where do we go from here? Theor Chem Acc 117:185–199

Liu HY, Zou X (2006) Electrostatics of ligand binding: parametrization of the generalized born model and comparison with the Poisson-Boltzmann approach. J Phys Chem B 110:9304–9313

Livermore DM (2009) Has the era of untreatable infections arrived? J Antimicrob Chemother 64:i29–i36

Lo Conte L, Brenner SE, Hubbard TJ, Chothia C, Murzin AG (2002) SCOP database in 2002: refinements accommodate structural genomics. Nucleic Acids Res 30:264–267

Lonsdale R, Mulholland AJ (2014) QM/MM modelling of drugmetabolizing enzymes. Curr Top Med Chem 14:1339–1347

Lyne PD, Lamb ML, Saeh JC (2006) Accurate prediction of the relative potencies of members of a series of kinase inhibitors using molecular docking and MM-GBSA scoring. J Med Chem 49:4805–4808

MacKerell AD, Bashford D, Bellott M, Dunbrack RL, Evanseck JD, Field MJ (1998) All-atom empirical potential for molecular modeling and dynamics studies of proteins. J Phys Chem B 102:3586–3616

Madan AK, Bajaj S, Dureja H (2013) Classification models for safe drug molecules. In: Reisfeld B, Mayeno AN (eds) Computational toxicology, vol 930. Humana Press, New York, pp 99–124

Mahoney MW, Jorgensen WL (2001) A five-site model for liquid water and the reproduction of the density anomaly by rigid, nonpolarizable potential functions. J Chem Phys 112:8910–8922

Malabarba A, Goldstein BP (2005) Origin, structure, and activity in vitro and in vivo of dalbavancin. J Antimicrob Chemother 55:15–20

Mandal S, Moudgil M, Mandal SK (2009) Rational drug design. Eur J Pharmacol 625:90–100

Marsac PJ, Li T, Taylor LS (2009) Estimation of drugpolymer miscibility and solubility in amorphous solid dispersions using experimentally determined interaction parameters. Pharm Res 26:139–151

Martin EJ, Hoeffel TJ (2000) Oriented Substituent Pharmacophore PRopErtY Space (OSPPREYS): a substituent-based calculation that describes combinatorial library products better than the corresponding product-based calculation. J Mol Graph Model 18:383–403

Mason JS, Beno BR (2000) Library design using BCUT chemistry-space descriptors and multiple four-point pharmacophore fingerprints: simultaneous optimization and structure-based diversity. J Mol Graph Model 18:438–451

Mason JS, Cheney DL (2000) Library design and virtual screening using multiple 4-point pharmacophore fingerprints. Pac Symp Biocomput 5:576–587

McGann MR, Almond HR, Nicholls A, Grant JA, Brown FK (2003) Gaussian docking functions. Biopolymers 68:76–90

Meller J, Elber R (2001) Linear programming optimization and a double statistical filter for protein threading protocols. Proteins: Struct Funct Bioinf 45:241

Meng X-Y, Zhang H-X, Mezei M, Cui M (2011) Molecular docking: a powerful approach for structure-based drug discovery. Curr Comput Aided Drug Des 7(2):146–157

Menichetti F (2005) Current and emerging serious Gram-positive infections. Clin Microbiol Infect 11(s3):22–28

Menikarachchi LC, Gascon JA (2010) QM/MM approaches in medicinal chemistry research. Curr Top Med Chem 10:46–54

Miao X, Waddell PJ, Valafar H (2008) TALI: local alignment of protein structures using backbone torsion angles. J Bioinform Comput Biol 6:163–181

Michel J, Verdonk ML, Essex JW (2006) Protein-ligand binding affinity predictions by implicit solvent simulations: a tool for lead optimization? J Med Chem 49:7427–7439

Michel J, Tirado-Rives J, Jorgensen WL (2009) Prediction of the water content in protein binding sites. J Phys Chem B 113:13337–13346

Milburn D, Laskowski RA, Thornton JM (1998) Sequence annotated by structure: a tool facilitate the use of structural information in sequence analysis. Protein Eng 11:855–859

Miller MS, Lay WK, Li S, Hacker WC, An J, Ren J, Elcock AH (2017) Reparameterization of protein force field nonbonded interactions guided by osmotic coefficient measurements from molecular dynamics simulations. J Chem Theory Comput 13:1812–1826

Miyazaki T, Aso Y, Yoshioka S, Kawanishi T (2011) Differences in crystallization rate of nitrendipine enantiomers in amorphous solid dispersions with HPMC and HPMCP. Int J Pharm 407:111–118

Modi S, Hughes M, Garrow A, White A (2012) The value of in silico chemistry in the safety assessment of chemicals in the consumer goods and pharmaceutical industries. Drug Discov Today 17:134–142

Moellering RC, Linden PK, Reinhardt J, Blumberg EA, Bompart F (1999) The efficacy and safty of quinupristin/dalfopristin for the treatment of infection caused by vancomycin-resistant Enterococcus faecium. J Antimicrob Chemother 44:251–261

Morris GM, Goodsell DS, Halliday RS, Huey R, Hart WE, Belew RK, Olson AJ (1998) Automated docking using a Lamarckian genetic algorithm and an empirical binding free energy function. J Comput Chem 19:1639–1662

Muegge I (2006) PMF scoring revisited. J Med Chem 49:5895–5902

Mustata G, Follis AV, Hammoudeh DI, Metallo SJ, Wang H, Prochownik EV, Lazo JS, Bahar I (2009) Discovery of novel Myc-Max heterodimer disruptors with a three-dimensional pharmacophore model. J Med Chem 52:1247–1250

Myrianthopoulos V, Gaboriaud-Kolar N, Tallant C, Hall ML, Grigoriou S, Brownlee PM, Fedorov O, Rogers C, Heidenreich D, Wainer M, Drosos N, Mexia N, Savitsky P, Bagratuni T, Kastritis E, Terpos E, Mikros E (2016) Discovery and optimization of a selective ligand for the switch/sucrose nonfermenting-related bromodomains of polybromo protein-1 by the use of virtual screening and hydration analysis. J Med Chem 59(19):8787–8803

Nelson DL, Lehninger AL, Cox MM (2008) Lehninger principles of biochemistry. W.H. Freeman, New York

Neumann D, Lehr C-M, Lenhof H-P et al (2004) Computational modeling of the sugar-lectin interaction. Adv Drug Deliv Rev 56:437–457

Nichols RL, Graham DR, Barriere SL, Rodgers A, Wilson SE, Zervos M, Dunn DL, Kreter B (1999) Treatment of hospitalized patients with complicated Gram-positive skin and skin structure infections: two randomized, multicentre studies of quinupristin/dalfopristin versus cefazolin, oxacillin or vancomycin. Synercid Skin and Skin Structure Infection Group. J Antimicrob Chemother 44:263–273

Nikaido H (2003) Molecular basis of bacterial outer membrane permeability revisited. Microbiol Mol Biol Rev 67:593–656

OECD (2015) The OECD QSAR toolbox. Available at: http://www.oecd.org/chemicalsafety/riskassessment/theoecdqsartoolbox.htm

Orengo CA, Bray JE, Buchan DW, Harrison A, Lee D (2002) The CATH protein family database: a resource for structural and functional annotation of genomes. Proteomics 2:11–21

Österberg F, Morris GM, Sanner MF, Olson AJ, Goodsell DS (2002) Automated docking to multiple target structures: incorporation of protein mobility and structural water heterogeneity in AutoDock. Proteins 46:34–40

Pallen MJ, Loman NJ, Penn CW (2010) High-throughput sequencing and clinical microbiology: progress, opportunities and challenges. Curr Opin Microbiol 13:625–631

Palumbi SR (2001) Humans as the world's greatest evolutionary force. Science 293:1786–1790

Peitsch MC, Schwede T, Guex N (2000) Automated protein modelling—the proteome in 3D. Pharmacogenomics 1:257–266

Pertel PE, Bernardo P, Fogarty C, Matthews P, Northland R, Benvenuto M, Thorne GM, Luperchio SA, Arbeit RD, Alder J (2008) Effects of prior effective therapy on the efficacy of daptomycin and ceftriaxone for the treatment of community-acquired pneumonia. Clin Infect Dis 46(8):1142–1151

Peters JM, King RW, Höög C, Kirschner MW (1996 Nov 15) Identification of BIME as a subunit of the anaphase-promoting complex. Science 274(5290):1199–1201

Peters KN, Dixon DM, Holland SM, Fauci AS (2008) The research agenda of the National Institute of Allergy and Infectious Diseases for antimicrobial resistance. J Infect Dis 197:1087–1093

Petersen RK, Christensen KB, Assimopoulou AN, Fretté X, Papageorgiou VP, Kristiansen K, Kouskoumvekaki I (2011) Pharmacophore-driven identification of PPARγ agonists from natural sources. J Comput Aided Mol Des 25:107–116

Phillips JC, Braun R, Wang W, Gumbart J, Tajkhorshid E, Villa E, Chipot C, Skeel RD, Kalé L, Schulten K (2005) Scalable molecular dynamics with NAMD. J Comput Chem 26:1781–1802

Pieper U, Eswar N, Ilyin VA, Stuart A, Sali A (2002) ModBase, a database of annotated comparative protein structure models. Nucleic Acids Res 30:255–259

Pitt WR, Calmiano MD, Kroeplien B, Taylor RD, Turner JP, King MA (2013) Structure-based virtual screening for novel ligands. Methods Mol Biol 1008:501–519

Poddar RK, Rakha P, Singh SK, Mishra DN (2010) Bioadhesive polymers as a platform for drug delivery: possibilities and future trends. Res J Pharm Dosage Forms Technol 1:40

Pronk S, Pall S, Schulz R, Larsson P, Bjelkmar P (2013) GROMACS 4.5: a high-throughput and highly parallel open source molecular simulation toolkit. Bioinformatics 29:845–854

Schrödinger Suite 2014-1 Protein Preparation Wizard, Epik, version 2.7; Schrödinger, LLC: New York, NY, USA, 2013

Raies AB, Bajic VB (2016) In silico toxicology: computational methods for the prediction of chemical toxicity. Wiley Interdiscip Rev Comput Mol Sci 6:147–172

Raizada A, Bandari A, Kumar B (2010) Polymers in drug delivery: a review. Int J Pharm Res Dev 2:9–20

Rajamani R, Good AC (2007) Ranking poses in structure-based lead discovery and optimization: current trends in scoring function development. Curr Opin Drug Discov Devel 10:308–315

Ramachandran GN, Ramakrishnan C, Sasisekharan V (1963) Stereochemistry of polypeptide chain configurations. J Mol Biol 7:95–99

Rarey M, Kramer B, Lengauer T, Klebe GA (1996) Fast flexible docking method using an incremental construction algorithm. J Mol Biol 261:470–489

Rask-Andersen M, Almén MS, Schiöth HB (2011) Trends in the exploitation of novel drug targets. Nat Rev Drug Discov 10:579–590

Reddy PD, Swarnalatha D (2010) Recent advances in novel drug delivery systems. Int J PharmTech Res 2:2025–2027

Retief JD (2000) Phylogenetic analysis using PHYLIP. Methods Mol Evol 132:143–158

Ripphausen P, Nisius B, Peltason L, Bajorath J (2010) Quo vadis, virtual screening? A comprehensive survey of prospective applications. J Med Chem 53:8461–8467

Rocchia W, Sridharan S, Nicholls A, Alexov E, Chiabrera A, Honig B (2002) Rapid grid-based construction of the molecular surface and the use of induced surface charge to calculate reaction field energies: applications to the molecular systems and geometric objects. J Comput Chem 23:128–137

Rohs R, Bloch I, Sklenar H, Shakked Z (2005) Molecular flexibility in ab-initio drug docking to DNA: binding-site and binding-mode transitions in all-atom Monte Carlo simulations. Nucleic Acids Res 33:7048–7057

Rost B (2001) Review: protein secondary structure prediction continues to rise. J Struct Biol 134:204–218

Rost B, Sander C (1995) Progress of 1D protein structure prediction at last. Proteins 23(3):295–300

Rowe PH (2010) Statistical methods for continuous measured endpoints in in silico toxicology. In silico toxicology: Principles and applications (pp. 228–251). Cambridge, UK: The Royal Society of Chemistry

Roy H, Brahma CK, Kumar R, Nandi S (2013) Formulation of saquinavir mesylate loaded microparticle by counterion induced aggregation method: approach by hyperosmotic technique. Drug Invention Today 5:259–266

Ruiz-Carmona S, Alvarez-Garcia D, Foloppe N (2014) rDock: a fast, versatile and open source program for docking ligands to proteins and nucleic acids. PLoS Comput Biol 10:e1003571

Rybak MJ, Hershberger E, Moldovan T, Grucz RG (2000) In vitro activities of daptomycin, vancomycin, linezolid, and quinupristin-dalfopristin against Staphylococci and Enterococci, including vancomycinintermediate and – resistant strains. Antimicrob Agents Chemother 44:1062–1066

Rychlewski L, Jaroszewski L, Li W, Godzik A (2000) Comparison of sequence profiles: strategies for structural predictions using sequence information. Protein Sci 9:232–241

Salum L, Polikarpov I, Andricopulo AD (2008) Structure-based approach for the study of estrogen receptor binding affinity and subtype selectivity. J Chem Inf Model 48:2243–2253

Santos Filho OA, Alencastro RB (2003) Modelagem de proteinas por homologia. Quím Nova 26:253–259

Sastry GM, Adzhigirey M, Day T, Annabhimoju R, Sherman W (2013) Protein and ligand preparation: parameters, protocols, and influence on virtual screening enrichments. J Comput Aided Mol Des 27:221–234

Schäffer AA, Aravind L, Madden TL, Shavirin S, Spouge JL, Wolf YI, Koonin EV, Altschul SF (2001) Improving the accuracy of PSI-BLAST protein database searches with composition-based statistics and other refinements. Nucleic Acids Res 29:2994–3005

Schapira M, Abagyan R, Totrov M (2003) Nuclear hormone receptor targeted virtual screening. J Med Chem 46:3045–3059

Schmaljohann D (2006) Thermo- and pH-responsive polymers in drug delivery. Adv Drug Deliv Rev 58:1655–1670

Schulz-Gasch T, Stahl M (2004) Scoring functions for protein-ligand interactions: a critical perspective. Drug Discov Today Technol 1:231–239

Schwartz R, Istrail S, King J (2001) Frequencies of amino acid strings in globular protein sequences indicate suppression of blocks of consecutive hydrophobic residues. Protein Sci 10:1023–1031

Schwede T, Kopp J, Guex N, Peitsch MC (2003) SWISS-MODEL: an automated protein homology-modeling server. Nucleic Acids Res 31:3381–3385

Sen S, Farooqui NA, Easwari TS, Roy B (2012) CoMFA-3D QSAR approach in drug design. Int. J. Res. Dev. Pharm. L. Sci 1:167–175

Shaik S, Cohen S, Wang Y (2010) P450 enzymes: their structure, reactivity, and selectivity-modeled by QM/MM calculations. Chem Rev 110:949–1017

Shan Y, Wang G, Zhou H (2001) Fold recognition and accurate query-template alignment by a combination of PSI-BLAST and threading. Proteins: Struct Funct Bioinf 42:23–37

Shelley JC, Cholleti A, Frye LL, Greenwood JR, Timlin MR, Uchimaya M (2007) Epik: a software program for pK (a) prediction and protonation state generation for drug-like molecules. J Comput Aided Mol Des 21:681–691

Shi Y (2014) A glimpse of structural biology through X-ray crystallography. Cell 159:995–1014

Shoichet BK, Stroud RM, Santi DV, Kuntz ID, Perry KM (1993) Structure-based discovery of inhibitors of thymidylate synthase. Science 259:1445–1450

Silverman JA, Perlmutter NG, Shapiro HM (2003) Correlation of daptomycin bactericidal activity and membrane depolarization in Staphylococcus aureus. Antimicrob Agents Chemother 47:2538–2544

Sippl MJ (1993) Recognition of errors in three-dimensional structures of proteins. Proteins 14: 355–362

Sippl MJ, Weitckus S (1992) Detection of native-like models for amino acid sequences of unknown three-dimensional structure in a data base of known protein conformations. Proteins: Struct Funct Bioinf 13:258–271

Skolnick J, Kihara D, Zhang Y (2004) Development and large scale benchmark testing of the PROSPECTOR 3 threading algorithm. Proteins 56:502–518

Sliwoski G, Kothiwale S, Meiler J, Lowe EW (2013) Computational methods in drug discovery. Pharmacol Rev 66:334–395

Spellberg B, Powers JH, Brass EP, Miller LG, Edwards JE Jr (2004) Trends in antimicrobial drug development: implications for the future. Clin Infect Dis 38:1279–1286

Springer J, Safley M, Huber S, Troxclair D, Craver R, Newman WP (2010) Histopathological findings in fatal novel H1N1: an autopsy case series from September-November 2009 in New Orleans. Louisiana J La State Med Soc 162:88–91

Srivastava S, Srivastava SK, Singh RB, Mishra S (2009) Computational haracterization of proteins involved in banana (Musa acuminata) ripening. Open Nutraceuticals J 2:122–126

Steenbergen JN, Alder J, Thorne GM, Tally FP (2005) Daptomycin: a lipopeptide antibiotic for the treatment of serious Gram-positive infections. J Antimicrob Chemother 55:283–288

Sung JH, Srinivasan B, Esch MB, McLamb WT, Bernabini C, Shuler ML, Hickman JJ (2014) Using PBPK guided “Body-on-a-Chip” systems to predict mammalian response to drug and chemical exposure. Exp Biol Med 239:1225–1239

Talele TT, Khedkar SA, Rigby AC (2010) Successful application of computer aided drug discovery: moving drugs from concept to the clinic. Curr Top Med Chem 10:127–141

Tannor DJ (2008) Introduction to quantum mechanics: a time-dependent perspective. University Science Books, Sausalito

ten Brink T, Exner TE (2010) pK(a) based protonation states and microspecies for protein-ligand docking. J Comput Aided Mol Des 24:935–942

Teodorescu O, Galor T, Pillardy J, Elber R (2004) Enriching the sequence substitution matrix by structural information. Proteins: Struct Funct Bioinf 54:41

Thomas CM, Nielsen KM (2005) Mechanisms of, and barriers to, horizontal gene transfer between bacteria. Nat Rev Microbiol 3:711–721

Thompson DC, Humblet C, Joseph-McCarthy D (2008) Investigation of MM-PBSA rescoring of docking poses. J Chem Inf Model 48:1081–1091

Tobi D, Elber R (2000) Distance-dependent, pair potential for protein folding: results from linear optimization. Proteins: Struct Funct Bioinf 41:40

Tollman PA (2001) Revolution in R&D: how genomics and genetics are transforming the biopharmaceutical industry. Boston Consulting Group, Boston

Trott O, Olson AJ (2010) AutoDock Vina: improving the speed and accuracy of docking with a new scoring function, efficient optimization, and multithreading. J Comput Chem 31:455–461

Umasankar V, Gurunathan S (2015) Drug discovery: an Appraisal. Int J Parm Pharm Sci 4:59–66

Valerio LG Jr (2009) In silico toxicology for the pharmaceutical sciences. Toxicol Appl Pharmacol 241:356–370

van der Kamp MW, Mulholland AJ (2013) Combined quantum mechanics/molecular mechanics (QM/MM) methods in computational enzymology. Biochemistry 52:2708–2728

Van der Spoel D, Lindahl E, Hess B, Groenhof G, Mark AE, HJC B (2005) GROMACS: fast, flexible, and free. J Comput Chem 26:1701–1718

Venkatachalam CM, Jiang X, Oldfield T, Waldman M (2003) LigandFit: a novel method for the shape-directed rapid docking of ligands to protein active sites. J Mol Graph Model 21:289–307

Venkatapathy R, Wang NCY (2013) Developmental toxicity prediction. In: Reisfeld B, Mayeno AN (eds) Computational toxicology, vol 930. Humana Press, New York, pp 305–340

Verma RP, Hansch C (2009) Camptothecins: a SAR/QSAR study. Chem Rev 109:213–235

Vyas V, Jain A, Jain A, Gupta A (2008) Virtual screening: a fast tool for drug design. Sci Pharm 76:333–360

Vyas VK, Ukawala RD, Ghate M, Chintha C (2012) Homology modeling a fast tool for drug discovery: current perspectives. Indian J Pharm Sci 74(1):1–17

Walsh C (2003) Antibiotics: actions, origins, resistance. American Society for Microbiology (ASM) Press, Washington, DC

Wang J, Wolf RM, Caldwell JW, Kollman PA, Case DA (2004) Development and testing of a general amber force field. J Comput Chem 25:1157–1174

Wang Y, Shaikh SA, Tajkhorshid E (2010) Exploring transmembrane diffusion pathways with molecular dynamics. Physiology 25:142–154

WaterMap (2014) New York, NY: Schrödinger, LLC

Westbrook J, Feng Z, Jain S, Bhat TN, Thanki N, Veerasamy R, Gilliland GL, Bluhm WF, Weissing H, Greer DS, Bourne PE, Berman HM (2002) The protein data bank: unifying the archive. Nucleic Acids Res 30:245–248

Wheeler SR, Kearney JB, Guardiola AR, Crews ST (2006) Single-cell mapping of neural and glial gene expression in the developing Drosophila CNS midline cells. Dev Biol 294:509–524

Willett P (2006) Similarity-based virtual screening using 2D fingerprints. Drug Discov Today 11(23–24):1046–1053

Wilson GL, Lill MA (2011) Integrating structure-based and ligand-based approaches for computational drug design. Future Med Chem 3:735–750

Winkler DA (2002) The role of quantitative structure-activity relationships (QSAR) in biomolecular discovery. Brief Bioinform 3:73–86

Wishart DS, Knox C, Guo AC, Shrivastava S, Hassanali M, Stothard P, Chang Z, Woolsey J (2006) DrugBank: a comprehensive resource for in silico drug discovery and exploration. Nucleic Acids Res 34(Database issue):D668–D672

Wolber G, Langer T (2005) LigandScout: 3-D pharmacophores derived from protein bound ligands and their use as virtual screening filters. J Chem Inf Model 45:160–169

Wood MJ, Hirst JD (2005) Protein secondary structure prediction with dihedral angles. Proteins 59:476–481

Woodfor N, Livermore DM (2009) Infections caused by Gram-positive bacteria: a review of the global challenge. J Infect 59:S4–S16

Wootton M, Howe RA, Walsh TR, Bennett PM, Macgowan AI (2000) In-vitro activity of a range of old and new antimicrobials to hetero-vancomycin-intermediate Staphylococcus aureus (hVISA) [abstract 23131. In: Program and abstract of the 40th Interscience Conference on Antimicrobial agents and chemotherapy; September 17–20, Toronto, Canada

Wu S, Skolnick J, Zhang Y (2007) Ab initio modeling of small proteins by iterative TASSER simulations. BMC Biol 5:17

Xiang Z (2006) Advances in homology protein structure modeling. Curr Protein Pept Sci 7:217–227

Xu Y, Xu D, Uberbacher EC (1998) An efficient computational method for globally optimal threading. J Comput Biol 5:597–614

Yan R, Xu D, Yang J, Walker S, Zhang Y (2013) A comparative assessment and analysis of 20 representative sequence alignment methods for protein structure prediction. Sci Rep 3:2619

Yang WW, Pierstorff E (2012) Reservoir-based polymer drug delivery systems. J Lab Autom 17:50–58

Yang L, Wang W, Sun Q, Xu F, Niu Y, Wang C, Liang L, Xu P (2016) Development of novel proteasome inhibitors based on phthalazinone scaffold. Bioorg Med Chem Lett 26:2801–2805

Yasuo K, Yamaotsu N, Gouda H, Tsujishita H, Hirono S (2009) Structure-based CoMFA as a predictive model – CYP2C9 inhibitors as a test case. J Chem Inf Model 49:853–864

Yona G, Levitt M (2002) Within the twilight zone: a sensitive profile-profile comparison tool based on information theory. J Mol Biol 315:1257–1275

Yoo SU, Krill SL, Wang Z, Telang C (2009) Miscibility/stability considerations in binary solid dispersion systems composed of functional excipients towards the design of multi-component amorphous systems. J Pharm Sci 98:4711–4723

Young T, Abel R, Kim B, Berne BJ, Friesner RA (2007) Motifs for molecular recognition exploiting hydrophobic enclosure in protein-ligand binding. Proc Natl Acad Sci U S A 104:808–813

Yuriev E, Agostino M, Ramsland PA (2011) Challenges and advances in computational docking: 2009 in review. J Mol Recognit 24:149–164

Zhang Y, Skolnick J (2004) Automated structure prediction of weakly homologous proteins on a genomic scale. Proc Natl Acad Sci U S A 101:7594–7599

Zhang Y, Arakaki A, Skolnick J (2005) TASSER: an automated method for the prediction of protein tertiary structures in CASP6. Proteins 61:91–98

Zhang W, Liu S, Zhou Y (2008) SP5: improving protein fold recognition by using torsion angle profiles and profile-based gap penalty model. PLoS One 3:e2325

Zhang M, Li H, Lang B, O'Donnell K, Zhang H, Wang Z, Dong Y, Wu C, Williams RO (2012) Formulation and delivery of improved amorphous fenofibrate solid dispersions prepared by thin film freezing. Eur J Pharm Biopharm 82:534–544

Zhao H, Caflisch A (2013) Discovery of ZAP70 inhibitors by high-throughput docking into a confirmation of its kinase domain generated by molecular dynamics. Bioorg Med Chem Lett 23:5721–5726

Zhou H, Skolnick J (2007) Ab initio protein structure prediction using chunk-TASSER. Biophys J 93:1510–1518

Zhou H, Zhou Y (2005) Fold recognition by combining sequence profiles derived from evolution and from depth-dependent structural alignment of fragments. Proteins 58:6321–6328

Chapter 12
Combinatorial Designing of Novel Lead Molecules Towards the Putative Drug Targets of Extreme Drug-Resistant *Mycobacterium tuberculosis*: A Future Insight for Molecular Medicine

Nikhil Bachappanavar and Sinosh Skariyachan

Contents

N. Bachappanavar · S. Skariyachan (✉)
Department of Biotechnology, Dayananda Sagar College of Engineering,
Dayananda Sagar Institutions, Bengaluru, Karnataka, India
e-mail: sinosh-bt@dayanandasagar.edu

N. A. Shaik et al. (eds.), *Essentials of Bioinformatics, Volume II*,
https://doi.org/10.1007/978-3-030-18375-2_12

12.1 Introduction

The emergence of extensively and totally drug-resistant (XDR and TDR) tuberculosis (TB) is a growing global concern. Various strains of XDR *Mycobacterium tuberculosis* (XDR-Mtb) have exhibited resistance to most of the currently prescribed first and second line of anti-tuberculosis drugs at an alarming rate (WHO 2018). Resistance to isoniazid and rifampicin is termed as multidrug-resistant tuberculosis (MDR-TB), and further resistance to fluoroquinolones and any one of the injectable drugs such as amikacin, kanamycin or capreomycin is termed extensively drug-resistant tuberculosis (XDR-TB) (Coll et al. 2018).

XDR-Mtb cases have been reported in more than 123 countries, and tuberculosis remains to be a leading cause of death (1.7 million annually) in developing countries (Quan et al. 2017). On an average, around 7% of patients with MDR-Mtb have XDR-Mtb (Maitre et al. 2017). The emergence of XDR-TB strains is due to the mismanagement of MDR cases; hence, new cases of XDR-TB can be prevented by early detection and proper treatment of existing patients with XDR-TB and the correct management of MDR-TB patients (Dheda et al. 2017; Matteelli et al. 2014). Further, TB has been associated with an increase (fourfold) in the mortality rates in population of patients infected with HIV infections (Bell and Noursadeghi 2018). It has also been reported that patients with highly drug-resistant TB are at an increased risk of longer and expensive treatments (Coll et al. 2018; Quan et al. 2017).

Computer-aided drug discovery (CADD) serves as an ideal platform for the identification of potential drug targets and screening of novel lead molecule against XDR-Mtb. The present chapter emphasizes that serine hydroxymethyltransferase (EC 2.1.2.1) (GlyA) has been identified as a putative drug target by Kyoto Encyclopedia of Genes and Genomes (KEGG) pathway analysis (Kanehisa et al. 2016). This enzyme is crucial for the survival of the bacteria and plays an important role in causing virulence (Raman et al. 2008). It has also been reported as a potential drug target for pathogenic *Plasmodium falciparum* and *Plasmodium vivax*, which causes malaria in humans (Sopitthummakhun et al. 2012). Thus, identification of effective inhibitors against this target enzyme is crucial in mitigating infections due to Mtb.

The development of novel drugs against TB has drawn significant attention to plant-based therapeutic agents with high medicinal value and their metabolites for potential antibacterial properties. Medicinal plants host innumerable bioactive compounds, and these compounds have progressively highlighted their importance in tackling invasive infections caused by MDR bacteria (Mohamad et al. 2018). Several natural compounds and their derivatives have been reported to show growth inhibitory activity against Mtb. Plant extracts from *Artemisia capillaris*, *Tinospora crispa*, *Zingiber officinale*, *Micromelum minutum*, *Clausena harmandiana*, *Aegle marmelos*, *Rollinia mucosa*, *Piper betle*, *Piper sarmentosum*, *Vitex trifolia*, *Piper nigrum* and many others have previously demonstrated growth inhibitory activities against MDR-Mtb (Mohamad et al. 2018; Sanusi et al. 2017). Some of the identified and isolated anti-mycobacterial compounds include allicin, β-sitosterol, friedelin, gallic acid, taraxerol, anthocyanidin, decarine, ellagic acids and 1-epicatechol, to name a few (Chinsembu 2016). Recently, studies have reported the in vitro anti-TB activity of several phytochemicals isolated from *Costus speciosus*, *Cymbopogon citratus* and *Tabernaemontana coronaria* (Mohamad et al. 2018). Similarly, in another recent study, the in vitro activity of *Phyllanthus niruri* (Bhumyamalaki) against Mtb has been described (Putri et al. 2018).

In the recent decades, this bacterium has exhibited resistance to a broad range of antibiotics, and the current approaches for the treatment and control of tuberculosis caused by XDR-Mtb are not sustainable (Quan et al. 2017). Hence, there is an urgent need, and it is paramount to identify novel drug targets and screen potential therapeutics against XDR-Mtb in order to overcome the global burden of TB caused by this pathogen (Bell and Noursadeghi 2018). Thus, screening of potential herbal-based lead molecules against drug targets of Mtb provides profound insights into the development of novel more efficacious antibacterial agents. This chapter illustrates the scope and application of computer-aided virtual screening for the identification of potential drug targets and screening of novel herbal-based lead molecules by various computational approaches to combat the global spread of XDR-Mtb.

12.2 Recent Perspectives on Drug-Resistant *Mycobacterium tuberculosis*

World Health Organization (WHO) declared TB as a global public health emergency in 1993, and since then, efforts are continuously being made to control the occurrence and spread of this pathogen. Dr. Margaret Chan, the Director-General of WHO, suggests that everyone with TB should have access to the innovative tools and services they need for rapid diagnosis, treatment and care. Further, high quality and complete care must be provided to combat drug-resistant TB. It has been estimated that there were 600,000 new cases with resistance to rifampicin—the most effective first-line antibiotic—out of which 490,000 had MDR-TB and 37,200 had XDR-TB. Similarly, in 2016, 87% of new TB cases were reported in the 30 high TB-burden countries, and 7 countries including India, Indonesia, Pakistan, Nigeria, South Africa, China and Philippines accounted for 64% of the new TB cases. Tuberculosis kills 5000 people every day, and efforts are being made to end the ‘global tuberculosis epidemic’. To take a lead in this direction, WHO has initiated ‘End TB Strategy’ (2015–2035) with the goals and milestones aligned in a way to reduce the number of TB deaths by 95% in number compared with 2015.

Similarly, according to the European Centre for Disease Prevention and Control (ECDC) (https://www.ecdc.europa.eu/en/home), 58,994 cases of TB were reported in 30 European Union and European Economic Area (EU/EEA) countries. Out of these, around 20% of the total TB cases have been XDR-TB in 2016. Similarly, 70.4% of the cases were newly diagnosed (ECDC 2018). Dr. Andrea Ammon, Director of ECDC, has asked all the healthcare systems to remain vigilant about TB, especially in low-incidence settings. She has also suggested making use of recent technologies such as whole-genome sequencing (WGS) to investigate cross-border outbreaks of MDR-TB. Likewise, ECDC launched a pilot project in 2017 to address the threat due to this pathogen. It has been suggested that with the advent of WGS technology, the detection and investigation of Mtb in the EU/EEA can be improved vastly. Further, this project will also establish guidelines for WGS in investigating MDR-TB and XDR-TB bacterial strains to trace outbreaks.

In the United States, according to Centers for Disease Control and Prevention (CDC) (https://www.cdc.gov/), a total of 9272 TB cases were reported in 2016 (CDC 2018). CDC, along with several academic institutions and Division of Tuberculosis Elimination (DTBE), has developed 'The Tuberculosis Epidemiologic Studies Consortium II' (TBESC-II). The goal of this consortium is to develop strategies and tools to increase diagnosis and treatment of latent tuberculosis infection in high-risk populations.

India accounts for about a quarter of the global TB burden with an estimated 2.79 million cases every year. India is the country with the highest burden of both TB and MDR-TB. Out of the total, 79,000 MDR-TB patients are notified with the cases of pulmonary TB each year. Hence, in March 2017, the Government of India announced to eliminate TB by 2025 by initiating 'National Strategic Plan' (NSP) (2017–2025) (https://tbcindia.gov.in/). According to WHO, elimination can be defined as less than one case of TB for a population of a million people. The Union Ministry of Health and Family Welfare, Government of India, has committed to ensure affordable and quality healthcare to the population in achieving zero TB deaths and to end TB by 2025. Due to the growing concerns by various governmental bodies and awareness across the world, combating infections due to Mtb must be of paramount importance. Hence, by identifying and studying major metabolic pathways responsible for the pathogenesis, conventional therapies and associated drug resistance, potential drug targets and screening of novel lead molecules by CADD, efforts can be made towards mitigating TB globally.

12.3 Metabolic Pathways in Tuberculosis

Metabolic pathways that are unique to the pathogen and absent in the host help in identifying proteins associated with virulence, important for persistence, or vital for mycobacterial metabolism and causing pathogenesis of tuberculosis (et al. 2016). Biochemical pathways such as polyketide sugar unit biosynthesis, mycobactin biosynthesis, peptidoglycan biosynthesis, methane metabolism, alanine metabolism, thiamine metabolism and C5-branched dibasic acid metabolism are present in the bacteria and absent in the host and, hence, can be studied for their role in pathogenesis of tuberculosis (Bushra and Adem 2016).

Mtb has the ability to survive in the nutrient-poor environment by switching to fatty acids and lipids as a sole source of carbon. Utilization of fatty acids has been possible due to the presence of aceA gene which encodes for the enzyme isocitrate lyase that converts isocitrate to succinate in the fatty acid metabolism pathway (Shukla et al. 2017). This process is further assisted by upregulation of several essential genes such as fadD3, fadD26, fadE5, echA19 and fadB2 that code for major enzymes in the pathway (Forrellad et al. 2013). Furthermore, genes such as gltA1, rv1130 and pckA were also highly expressed for the survival and pathogenesis of XDR-Mtb (Raman and Chandra 2008).

Similarly, it has been reported that pantothenate is a key nutrient involved in the biosynthesis of fatty acids, synthesis of CoA and other cellular processes. In Mtb, the genes PanC and PanD were reported to be upregulated in virulent strains (Mukhopadhyay et al. 2012). These genes code for pantothenate synthetase and aspartate-1-decarboxylase, respectively, and play a vital role in the virulence of drug-resistant Mtb. It has been observed that DdlA, EmbA, EmbB, AftA, AftB and MurG have been suggested as attractive targets in the biosynthesis of mycobacterial cell wall (Seidel et al. 2007; Alderwick et al. 2006; Belanger et al. 1996).

ArgA, an essential enzyme catalysing the initial step of arginine biosynthesis, and AroB in the shikimate pathway have been associated with virulence and pathogenesis in tuberculosis (de-Mendonça et al. 2007; Errey and Blanchard 2005). Another important pathway associated with causing virulence and pathogenesis in Mtb is the iron acquisition pathway (Mukhopadhyay et al. 2012). There are several genes that get upregulated during the survival of this pathogen and hence have been identified as an important pathway. Further, in the pantothenic acid biosynthesis synthesis pathway, the enzymes PanB, PanC and PanD have been identified for the survival and metabolism of fatty acids and lipids in Mtb (Sambandamurthy et al. 2002). Similarly, in the pathway of two-component system, DevR and DevS have been identified as key proteins in causing virulence in Mtb (Saini and Tyagi 2005). DevR is also essential for growth of Mtb under low oxygen conditions and DevS plays a key role in signal transduction. Hence, the study of the major pathways and associated genes related to the pathogenesis and virulence in tuberculosis has profound scope in anti-tuberculosis drug discovery.

12.4 Conventional Treatments Against Tuberculosis

12.4.1 First-Line Anti-tuberculosis Drugs

Isoniazid is a prodrug which gets activated by the catalase/peroxidase enzyme encoded by the KatG gene (Dookie et al. 2018). Activated isoniazid inhibits the synthesis of essential mycolic acid (involved in synthesis of mycobacterial cell wall) via the NADH-dependent enoyl-acyl enzyme, encoded by the InhA gene (Seifert et al. 2015). Isoniazid resistance is usually mediated by mutations in the KatG and InhA gene (Mukhopadhyay et al. 2012). Isoniazid-resistant isolates are reported more frequently than any other anti-tuberculosis drugs (Seifert et al. 2015). Mutations in InhA and KatG not only cause high resistance to isoniazid but also result in cross-resistance to structurally related second-line drug ethionamide (Mukhopadhyay et al. 2012).

Rifampicin is a bactericidal antibiotic that acts actively against growing and stationary phase Mtb isolates (Mukhopadhyay et al. 2012). It binds to the β-subunit of the DNA-dependent RNA polymerase, inhibiting the elongation of messenger RNA. Rifampicin-resistant Mtb show mutations within the RpoB gene that code for the β-subunit of RNA polymerase (Vidyaraj et al. 2017). Majority of rifampicin-resistant strains also show resistance to isoniazid; hence, detection of rifampicin resistance is considered as an indicator for MDR-Mtb (Forrellad et al. 2013). Further, cross-resistance cases between rifampicin and rifabutin have been linked to the mutations in the hotspot region of RpoB gene (Vidyaraj et al. 2017).

Pyrazinamide is an important first-line prodrug that is used along with isoniazid and rifampicin for treatment of tuberculosis. The mode of action for pyrazinamide is similar to that of isoniazid. Majority of pyrazinamide-resistant Mtb strains (72–97%) have mutations in the PncA gene (Njire et al. 2016). PncA gene codes for the enzyme pyrazinamidase which converts pyrazinamide to pyrazinoic acid (Dookie et al. 2018). Pyrazinamide is effective against Mtb and shows no activity against other *Mycobacterium* species (Njire et al. 2016). Hence, pyrazinamide resistance in *Mycobacterium* species is a specific indicator of drug-resistant Mtb (Rajendran and Sethumadhavan 2014).

Ethambutol is an important anti-tuberculosis drug that is effective against multiplying bacilli (Zhao et al. 2015). However, this bacteriostatic agent fails to show effect against non-replicating bacilli (Mukhopadhyay et al. 2012). In Mtb, the EmbB gene encodes for the enzyme arabinosyl transferase which is further involved in the biosynthesis of arabinogalactan, a vital component of mycobacterial cell wall (Forrellad et al. 2013). Ethambutol inhibits the biosynthesis of the arabinogalactan thereby disintegrating the bacterial cell wall (Dookie et al. 2018). It has been observed that the majority of ethambutol-resistant Mtb strains have mutations in the EmbB gene (Zhao et al. 2015).

Streptomycin, a bactericidal antibiotic which is effective against stationary phase bacilli, inhibits protein synthesis by binding to the 30S ribosomal subunit of the bacteria (Sun et al. 2016). The genes RpsL and Rrs in Mtb encode for the ribosomal protein S12 and 16S rRNA, respectively. It has been reported that mutations in these genes are a major cause of streptomycin resistance in Mtb (Sun et al. 2016). However, mutations in the GidB gene, encoding a 7-methylguanosine methyl transferase, have also been associated in low-level streptomycin resistance.

12.4.2 Second-Line Anti-tuberculosis Drugs

Amikacin and kanamycin are prescribed as second line of antibiotics for the treatment of tuberculosis. Several studies have reported cross-resistance between amikacin and kanamycin or between kanamycin and capreomycin or viomycin (Krüüner et al. 2003). Resistance to amikacin and kanamycin in Mtb has been majorly associated with a mutation in *rrs* gene, which codes for 16S rRNA of the bacteria (Mukhopadhyay et al. 2012; Maus et al. 2005a). Further, mutations in the *eis* gene (promotor region), which codes for acetyltransferase enzyme, have also been reported to result in low-level resistance to kanamycin (Forrellad et al. 2013). According to previous studies, viomycin and capreomycin have exhibited full cross-resistance due to the mutations in the gene *tlyA*, which codes for rRNA methyl transferase (Maus et al. 2005b).

Fluoroquinolones, specifically ciprofloxacin, moxifloxacin and levofloxacin, are currently used as second-line treatment for TB. These antibiotics play an important role in the treatment of TB as they show high bactericidal activity with fewer side effects in comparison with other TB drugs (Jabeen et al. 2015). These antibiotics inhibit the function of mycobacterial DNA gyrase (type II topoisomerase) encoded by the *gyrA* gene (Forrellad et al. 2013). Mtb resistant to the entire class of fluoroquinolones has often been associated with genetic mutations in *gyrA* and *gyrB* genes (Jabeen et al. 2015).

Ethionamide, prothionamide and isoniazid are structurally similar prodrugs prescribed for the treatment of TB (Dookie et al. 2018). The mechanism of action towards the treatment of TB is also similar to that of isoniazid (Vilchèze and Jacobs 2014). These drugs inhibit the expression of *inhA* gene present in the mycolic acid synthesis pathway (Mukhopadhyay et al. 2012). Hence, ethionamide-resistant Mbt strains have been associated to mutations in the *ethA* and *inhA* genes which are also the responsible genes for isoniazid resistance (Tan et al. 2017). However, another gene *MshA*, encoding glycosyltransferase enzyme involved in mycothiol biosynthesis, has also been suggested as a potential drug target for ethionamide (Vilchèze and Jacobs 2014).

Cycloserine is a structural analogue of D-alanine, and it inhibits the biosynthesis of mycobacterial cell wall by inhibiting the action of D-alanine ligase enzyme (Chen et al. 2017). However, the drug target of cycloserine has not yet been experimentally elucidated in Mtb, yet overexpression of *alrA* gene in *M. smegmatis* resulted in high resistance to cycloserine.

12.4.3 Third-Line Anti-tuberculosis Drugs

Delamanid belongs to the nitroimidazole class of antibiotics and is structurally similar to pretomanid. The mechanism of action towards TB is by inhibiting the synthesis of mycolic acids which are vital for biosynthesis of mycobacterial cell wall (Mukhopadhyay et al. 2012). This prodrug specifically inhibits methoxy-mycolic and keto-mycolic acids while isoniazid also inhibits α-mycolic acid (D'Ambrosio et al. 2017). It has also been reported that mutations in the fbiA and fdg1 genes are associated to delamanid resistance (D'Ambrosio et al. 2017).

Bedaquiline belongs to diarylquinolines, a new class of drugs. It acts by inhibiting mycobacterial ATP synthase, which further affects the survival of Mtb. The *AtpE* gene encodes for an important mycobacterial F1F0 proton ATP synthase, a vital enzyme for ATP synthesis and generation of membrane potential. Hence, mutations in the *AtpE* gene have been attributed with increased resistance to bedaquiline (Dookie et al. 2018). *P*-amino salicylic acid along with isoniazid and streptomycin was one of the first antibiotics used in the treatment of TB. *P*-amino salicylic acid has been classified as a part of third line of antibiotics in the treatment of TB. In Mtb, *p*-aminosalicylic acid inhibits dihydropteroate synthase, an important enzyme in the folate biosynthesis (Forrellad et al. 2013). Similarly, the main reason for *p*-amino salicylic acid resistance has been attributed to mutations occurring in the *thyA* gene that encodes for thymidylate synthase (Almeida-Da-Silva and Palomino 2011).

Linezolid belongs to the class of drugs known as oxazolidinones, and it has been approved for the treatment of TB. The mechanism of action in Mtb is by inhibiting the synthesis of proteins by binding to the V domain of the 50S ribosomal subunit in the bacteria. Linezolid-resistant Mtb strains are rarely reported, yet resistant strains have been identified with mutations in the *rrl* and *rplC* gene, encoding the 50S ribosomal sequence (Almeida-Da-Silva and Palomino 2011). An overview of currently prescribed antibiotics against XDR-Mtb, their mode of action, associated resistance mechanisms (genes involved) and commonly occurring side effects of the antibiotics have been depicted in Table 12.1.

12.5 Scope of Computer-Aided Drug Discovery (CADD) and Associated Challenges

Modern drug discovery and development focuses on understanding disease mechanisms which further leads to target identification, validation and screening of potential leads. In this process, computational tools offer tremendous potential in target identification, virtual screening, de novo synthesis and integration of data on multiple levels (Katsila et al. 2016). Similarly, state-of-the-art network-based computational algorithms pharmacophore substructure similarity searching, data mining through machine learning, molecular docking, molecular dynamic (MD) simulations and bioactivity spectra-based algorithms and systems biology approaches help in integrating information from various databases and optimize the process of drug development (Katsila et al. 2016; Engin et al. 2014). The identification of potential drug targets can also be carried out through network-based approaches where information from different databases is integrated to understand the importance and role of proteins in specific disease networks. This approach is highly reliable and includes the amalgamation of data from various fields such as pharmacogenomics, genomics, proteomics, transcriptomics, microbiome and metabolomics, to name a few. It also makes use of computational biology tools for data correlation and interpretation (Anastasio 2017; Engin et al. 2014). Similarly, the application of connectivity maps has recently helped several researchers and pharmaceutical industries to find a common link between functionally associated genes in disease prognosis and

Table 12.1 Conventional therapies and associated drug resistance. An overview of first, second and third line of antibiotics prescribed against extremely drug-resistant *Mycobacterium tuberculosis*, their mode of action, associated resistance mechanisms (genes involved) and commonly occurring side effects of the antibiotics

Anti-tuberculosis drugs	Activity of antibiotic	Gene(s) involved in resistance	Mode of action	Side effects	Year of discovery	Reference
First line of antibiotics						
Rifampicin	Bactericidal	*rpoB*	Inhibits transcription by binding to RpoB, the β-subunit of DNA-dependent RNA polymerase	Gastrointestinal disturbances, rashes and allergic reactions	1965	Vidyaraj et al. (2017)
Isoniazid	Bacteriostatic	*katG* *inhA* *kasA*	Inhibits the synthesis of mycolic acids, which is a vital component of mycobacterial cell wall	Headache, weight gain, peripheral neuropathy and gastrointestinal disturbances	1952	Seifert et al. (2015)
Pyrazinamide	Bactericidal	*pncA* *rpsA* *panD*	Pyrazinoic acid disrupts mycobacterial cell wall and interferes with energy production	Nausea, vomiting, anorexia, sideroblastic anaemia, skin rashes and arthralgia	1936	Njire et al. (2016)
Ethambutol	Bacteriostatic	*embB* *ubiA*	Inhibits cell wall arabinogalactan biosynthesis	Optic neuritis, red-green colour blindness, peripheral neuropathy, arthralgia, nausea, headache, and allergic reactions	1961	Zhao et al. (2015)
Streptomycin	Bactericidal	*rpsL* *rrs* *gidB*	Interferes with the binding of formyl-methionyl-tRNA to the 30S subunit	Vomiting, rashes and numbness of the face	1943	Sun et al. (2016)
Second line of antibiotics						
Amikacin	Bactericidal	*Rrs* *tlyA* *eis*	Inhibits the protein synthesis by binding to 16S rRNA	Vertigo, numbness, hearing loss and kidney problems	1971	Krüüner et al. (2003)
Kanamycin	Bactericidal	*Rrs* *tlyA* *eis*	Inhibits the protein synthesis by binding to 30S rRNA	Vertigo, nausea, vomiting, dizziness and loss of balance	1957	Krüüner et al. (2003)

(continued)

Table 12.1 (continued)

Anti-tuberculosis drugs	Activity of antibiotic	Gene(s) involved in resistance	Mode of action	Side effects	Year of discovery	Reference
Capreomycin	Bactericidal	*tlyA* *eis* *rrs*	Inhibits the protein synthesis by binding to 70S rRNA	Kidney problems, hearing problems, poor balance, pain at the site of injection and allergic reactions	1960	Maus et al. (2005a, b)
Viomycin	Bactericidal	*Rrs* *eis* *tlyA*	Inhibits the protein synthesis by binding to 70S rRNA	Gastrointestinal disturbances	1951	Maus et al. (2005a, b)
Ciprofloxacin	Bactericidal	*gyrA* *gyrB*	Inhibits the DNA gyrase, thereby inhibiting cell division	Nausea, vomiting, diarrhoea and rashes	1987	Jabeen et al. (2015)
Moxifloxacin	Bactericidal	*gyrA* *gyrB*	Inhibits DNA gyrase, thereby inhibiting cell division	Diarrhoea, dizziness, and headache	1988	Jabeen et al. (2015)
Levofloxacin	Bactericidal	*gyrA* *gyrB*	Inhibits DNA gyrase, thereby inhibiting cell division	Nausea, diarrhoea and trouble sleeping	1985	Jabeen et al. (2015)
Ethionamide	Bacteriostatic	*etaA* *ethA* *inhA* *mshA* *ndh*	Inhibits the synthesis of mycolic acids, which is a vital component of mycobacterial cell wall	Nausea, diarrhoea, abdominal pain and loss of appetite	1956	Vilchèze and Jacobs (2014)
Prothionamide	Bacteriostatic	*etaA* *inhA*	Inhibits the enzyme InhA	Nausea and vomiting	1956	Tan et al. (2017)
Cycloserine	Bacteriostatic	*alrA* *cycA*	Inhibits the biosynthesis of mycobacterial cell wall	Allergic reactions, seizures, sleepiness, unsteadiness and numbness	1954	Chen et al. (2017)
Terizidone	Bacteriostatic	*alrA*	Inhibits the biosynthesis of mycobacterial cell wall	Nausea, vomiting and skin allergies	1991	Chen et al. (2017)

Third line of antibiotics						
Rifabutin	Bactericidal	*rpoB*	Inhibition of DNA-dependent RNA polymerase	Abdominal pain, nausea, rash, headache, and low blood neutrophil levels	1975	Berrada et al. (2016)
Amoxicillin-clavulanate	Bactericidal	*blaC*	β-lactamase inhibitor	Diarrhoea, vomiting, and allergic reactions	1979	Gonzalo and Drobniewski (2013)
Meropenem	Bactericidal	*blaC*	Inhibits the biosynthesis of mycobacterial cell wall	Nausea, diarrhoea, constipation, headache, rashes and pain at the site of infection	1983	Chambers et al. (2005)
p-amino salicylic acid	Bacteriostatic	*thyA* *falC* *ribD*	Inhibits folic acid and thymine nucleotide metabolism	Nausea, abdominal pain, diarrhoea, liver inflammation and allergic reactions	1902	Zheng et al. (2013)
Imipenem-cilastatin	Bactericidal	*blaC*	Inhibits the biosynthesis of mycobacterial cell wall	Nausea, diarrhoea, pain at the site of injection and allergic reactions	1987	Chambers et al. (2005)
Linezolid	Bactericidal	*rplC* *rrl*	Inhibits the biosynthesis of proteins	Headache, diarrhoea, rashes and nausea	1990	Agyeman and Ofori-Asenso (2016)
Bedaquiline	Bactericidal	*atpE* *pepQ*	Inhibits the mycobacterial ATP synthase	Nausea, joint pain, headaches and chest pain	2004	Field (2015)
Delamanid	Bacteriostatic	*Fgd1* *fbiA* *fbiB* *fbiC*	Inhibits the synthesis of mycolic acids, which is a vital component of mycobacterial cell wall	Headache, dizziness and nausea	2014	D'Ambrosio et al. (2017)

drug interactions for a pathogen. A connectivity map is an assembly of data collected from whole-genome transcriptional expression of cultured human cells treated with bioactive small molecules (Anastasio 2017). The validation of identified targets is a laborious, time-consuming and expensive process. The efficiency of this process can be vastly improved when combined with computational approaches such as step-wise data filtering by biostatistics. High-throughput screening (HTS) usually offers identification of several hits; yet, the success rates are often lower as many of the identified compounds are rejected due to their lack of physicochemical properties. This can be avoided by the combinatorial approach, wherein CADD and HTS are applied together to screen and identify novel leads against a particular target. An overview of CADD in identifying novel leads against potential drug targets of *Mycobacterium tuberculosis* has been depicted in Fig. 12.1.

Recently, a study illustrated structure-based virtual screening of natural compounds to identify potential inhibitors against Mtb isocitrate lyase (Shukla et al. 2017). This enzyme catalyses the first step in the glyoxylate cycle and plays a key role in the survival of Mtb. Hence, structure-based virtual screening of natural compounds from the ZINC database (167,748 compounds) was performed to identify three potential inhibitors (ZINC1306071, ZINC2111081 and ZINC2134917) against this enzyme. These ligands were docked against the isocitrate lyase enzyme and were further subjected to MD simulation to understand ligand binding and the stability of the bound complexes. Similarly, these lead compounds also displayed substantial pharmacological and structural properties to be drug candidates (Shukla et al. 2017). In another study conducted by Silva et al. (2016), it was suggested that carbapenems such as imipenem and meropenem inhibit the activity of L,D-transpeptidase enzyme, which is a key enzyme for synthesis of L,D-transpeptide linkages in the mycobacterial cell wall. Further, molecular modelling approaches were undertaken to study the enzyme/inhibitor interactions. Furthermore, the binding energies for nine commercially available inhibitors were calculated using molecular mechanics/generalized born surface area (MM/GBSA) and solvation interaction energy (SIE) approaches, and the calculated energies corresponded well with the available in vitro analysis.

CADD is an interdisciplinary field that requires collaborative efforts among highly intellectual professionals from systems biology, computational chemistry, chemoinformatics, bioinformatics, computational biology and pharmacogenomics. In scientific computing, in order to make the calculations in a finite period of time, several assumptions, significant approximations and numerous algorithms are applied (Baldi 2010). Hence, these factors weaken the accuracy of any ligand-receptor interaction and are identified as major limitations of CADD. Similarly, screening of large number of compounds leads to the identification of undesired chemical structures which are chemically unstable, synthetically unfeasible or have higher toxicity (Baldi 2010). The handling of large amounts of data generated by these methods is quite difficult and poses as a drawback. However, there have been significant improvements in the development of softwares with user-friendly programs, and with the advent of ultra-fast supercomputers, CADD has been considered as a reliable approach and has been integrated in the process of modern drug design and development (Baldi 2010).

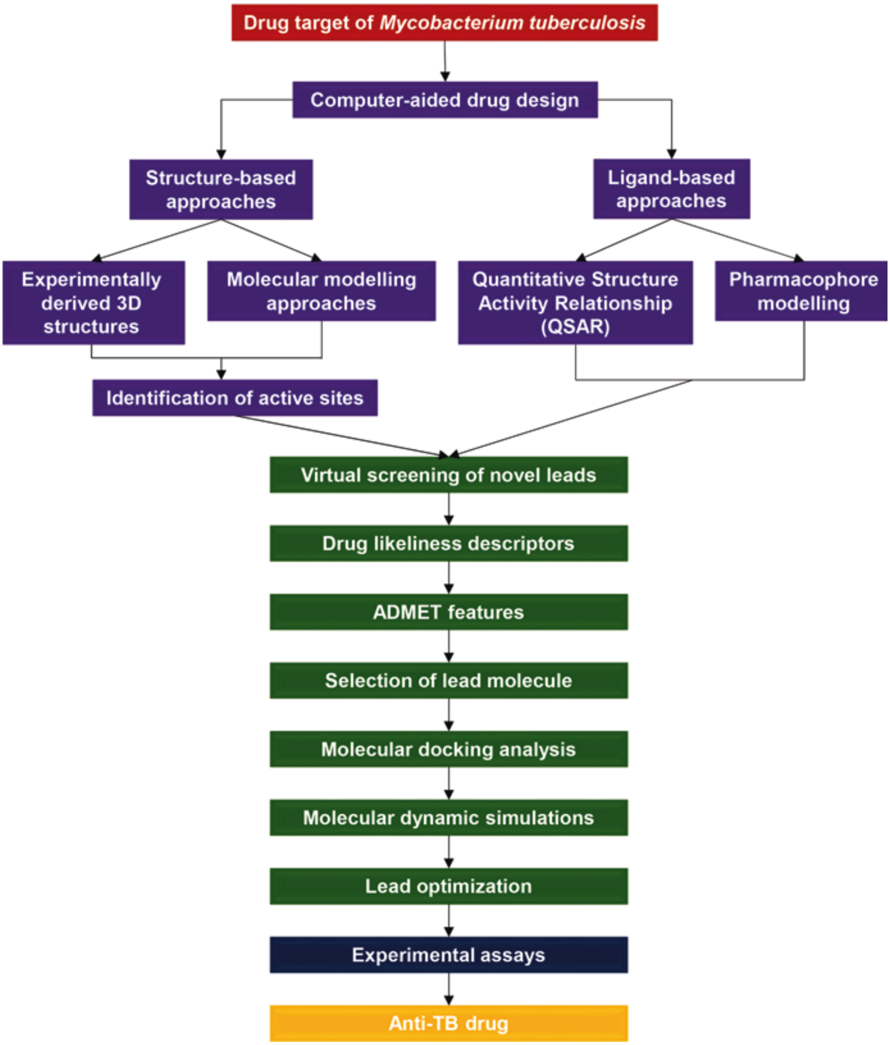

Drug target of *Mycobacterium tuberculosis*
Computer-aided drug design
Structure-based approaches
Ligand-based approaches
Experimentally derived 3D structures
Molecular modelling approaches
Quantitative Structure Activity Relationship (QSAR)
Pharmacophore modelling
Identification of active sites
Virtual screening of novel leads
Drug likeliness descriptors
ADMET features
Selection of lead molecule
Molecular docking analysis
Molecular dynamic simulations
Lead optimization
Experimental assays
Anti-TB drug

Fig. 12.1 An overview of computer-aided drug discovery in identifying novel leads against potential drug targets of *Mycobacterium tuberculosis*. Virtual screening has become a key component of the modern drug discovery and development process

12.6 Computational Biology Approaches for Identification of Putative Drug Targets

Target identification is the first step and the most important step in the drug discovery pipeline. The initial step requires assessing several metabolic pathways to identify a potential biological target (Mehra et al. 2016; Reddy et al. 2007). The outcome of this step can be improved if the molecular mechanisms of disease have been previously elucidated (Chandra 2011). Conventional approaches of in vitro and in vivo identification and validation of targets include whole-cell or animal experiments, gene knockout and site-directed mutagenesis studies. These approaches are time-consuming, laborious and not feasible for screening large number of receptors (Chandra 2011). In the recent decades, in silico approaches help in reducing cost, analysing voluminous amounts of data generated from experiments (gene profiling by microarrays), and they provide an overall picture of the systems involved at the molecular level in the bacteria. Hence, computational approaches offer tremendous potential in screening and selecting drug targets along with conventional methods. Drug targets can be identified by various approaches such as gene network analysis, kinetic modelling, flux balance analysis, topological analysis and rule-based analysis of key enzymes and proteins in the bacteria (Amir et al. 2014; Chandra 2011). Gene network analyses have been playing a key role in screening certain proteins that may cause resistance in pathogenic bacteria (Reddy et al. 2007). Similarly, these proteins along with other virulence factors can be targeted to inhibit the resistance mechanism in Mtb (Lionta et al. 2014; Chandra 2011).

On the other hand, databases also play a significant role in the identification of several drug targets. Some of the commonly used databases include Therapeutic Target Database (TTD) (Li et al. 2018), DrugBank (https://www.drugbank.ca/), DrugMap Central (http://r2d2drug.org/index.html), Gene Ontology Consortium (http://www.geneontology.org/), Reactome (https://reactome.org/), Kyoto Encyclopedia of Genes and Genomes (KEGG) (http://www.genome.jp/kegg/), Panther 13.1 (http://www.pantherdb.org/) and Potential Drug Target Database (PDTD) (Gao et al. 2008). Similarly, some of the Mtb-specific databases include MIRU-VNTRplus (http://www.miru-vntrplus.org/MIRU/index.faces), MycoperonDB (http://cdfd.org.in/mycoperondb/home.html), TB Drug Target (TBDT) (http://www.bioinformatics.org/tbdtdb/) and Mycobrowser (https://mycobrowser.epfl.ch/).

Based on the metabolic pathway analysis, several targets can be selected for structure-based virtual screening, for example, hydroxymethyltransferase (EC 2.1.2.1) (GlyA) has been identified as a potential drug target by KEGG pathway analysis. This enzyme plays a vital role in causing virulence, pathogenesis and survival of Mtb. It is actively involved in several pathways such as biosynthesis of amino acids and secondary metabolites, metabolism of carbon, methane, cyanoamino acid and glyoxylate and dicarboxylate. The native structure of serine hydroxymethyltransferase (PDB ID: 3H7F) possessed two chains (A and B) with molecular weight of 95226.08 Da and a resolution of 1.5 Å (*R*-value free, 0.196) (Fig. 12.2a). Further, gene network analysis for the gene GlyA, performed by STRING database, revealed that genes such as guaA, gcvP, purH, folD, PurH, PurN, cys, gcvH and PurM (Fig. 12.2c) closely interacted with GlyA and performed major functions in metabolic pathway (Szklarczyk et al. 2017). The key proteins associated to serine hydroxymethyltransferase are glycine dehydrogenase, IMP cyclohydrolase, methenyltetrahydrofolate cyclohydrolase, cysteine synthase and phosphoribosylglycinamide formyltransferase (Szklarczyk et al. 2017). The genes and their gene products involved in this network play an important role in the pathogenesis and virulence of Mtb. Serine hydroxymethyltransferase has been reported

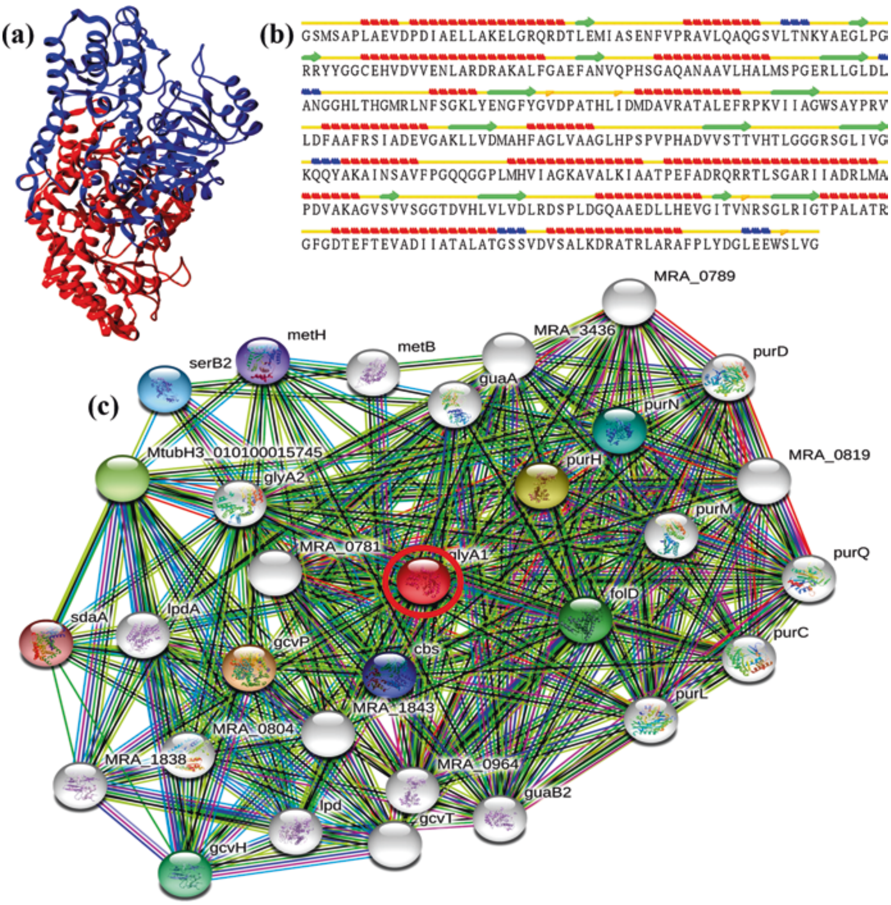

(a)
(b)
GSMSAPLAEVDPDIAELLAKELGRQRDTLEMIASENFVPRAVLQAQGSVLTNKYAEGLPG
RRYYGGCEHVDVVENLARDRAKALFGAEFANVQPHSGAQANAAVLHALMSPGERLLGLDL
ANGGHLTHGMRLNFSGKLYENGFYGVDPATHLIDMDAVRATALEFRPKVIIAGWSAYPRV
LDFAAFRSIADEVGAKLLVDMAHFAGLVAAGLHPSPVPHADVVSTTVHTLGGGRSGLIVG
KQQYAKAINSAVFPGQQGGPLMHVIAGKAVALKIAATPEFADRQRRTLSGARIIADRLMA
PDVAKAGVSVVSGGTDVHLVLVDLRDSPLDGQAAEDLLHEVGITVNRSGLRIGTPALATR
GFGDTEFTEVADIIATALATGSSVDVSALKDRATRLARAFPLYDGLEEWSLVG
(c)
MRA_0789
MRA_3436
purD
metH
metB
serB2
guaA
purN
MRA_0819
MtubH3_010100015745
glyA2
purH
purM
MRA_0781
glyA1
purQ
folD
sdaA
lpdA
gcvP
cbs
purC
purL
MRA_1843
MRA_0804
MRA_1838
MRA_0964
guaB2
lpd
gcvT
gcvH

Fig. 12.2 Visualization, detailed secondary structure analysis and gene interaction network analysis for serine hydroxymethyltransferase from Mtb. (**a**) Visualization of the 3D structure of serine hydroxymethyltransferase by UCSF Chimera. (**b**) The detailed secondary structure alignment visualized using STRIDE web interface revealed 34 α-helices, 59 turns, 29 β-strands and 11 3_{10}-helices. (**c**) Gene network analysis for the gene GlyA, performed by STRING database revealed some of the major interacting genes such as guaA, gcvP, purH, folD, PurH, PurN, cys, gcvH and PurM. The gene product of GlyA has been selected as probable drug target in the study and is highlighted in the figure. In the figure, coloured nodes represent the first shell of interactors, while white nodes represent second shell of interactors and empty nodes represent proteins of unknown 3D structure. The genes and their gene products involved in this network play an important role in the pathogenesis and virulence of Mtb

as a potential drug target for pathogenic *Plasmodium falciparum* and *Plasmodium vivax*, which causes malaria in humans (Sopitthummakhun et al. 2012). Similarly, a comprehensive list of potential drug targets, their involvement in major pathways and associated genes in drug-resistant Mtb have been tabulated (Table 12.2).

The exploration of pathogenesis caused by Mtb and the identification of the related targets is an important step in combating tuberculosis (Geromichalos 2012).

Hence, Raman et al. (2008) have developed a comprehensive in silico target identification pipeline (targetTB) for drug-resistant Mtb by an interactome, reactome and genome-wide structural analysis. In the pipeline, the authors have incorporated a network analysis of the protein-protein interactome, a flux balance analysis of the reactome, experimentally derived phenotype essentiality data, sequence analyses and a structural assessment of target ability by the application of novel algorithms (Raman et al. 2008). Such resources aid in the identification and validation of drug targets of Mtb by computational approaches. Similarly, Amir et al. (2014) have performed an in silico comparative analysis of metabolic pathways of the host *Homo sapiens* and *Mycobacterium tuberculosis* H37Rv strain to identify potential drug targets. The study identified five unique metabolic pathways comprising of 55 enzymes which are essential for the survival and virulence in Mtb and also which are non-homologous to human protein sequences. Further, the functional analysis using UniProt and DEG database revealed the importance of all the unique enzymes in the synthesis of different cellular components (Amir et al. 2014).

12.7 Protein Structure Prediction

Predicting the 3D structure of molecular targets (most cases a receptor or protein) is a key step in structure-based drug discovery as it will assist in designing inhibitors or anti-TB drugs against XDR-Mtb (Qiu et al. 2017). The 3D structures of the identified targets that are not available in their native forms can be computationally predicted by various approaches. The 3D structure of proteins can be predicted by three different approaches, namely, homology modelling, fold recognition and ab initio methods.

12.7.1 Homology Modelling

Homology modelling, also known as comparative modelling, is a process of predicting a three-dimensional structure of the ‘target’ protein from its amino acid sequence and an experimental elucidated 3D structure of a related homologous protein of the identified template. It has been observed that the protein structures among homologous sequences are highly conserved in nature, whereas, sequences that fall under 30% sequence identity exhibit different structures (Vyas et al. 2012). Similarly, proteins that are evolutionarily related have similar sequences, and naturally occurring homologous proteins exhibit similar protein structures (Liu et al. 2011). This method is widely applied in structure-based drug discovery to predict 3D structures of potential drug targets that play a vital role in causing pathogenesis (Vyas et al. 2012). Some of the key steps involved in the process of homology modelling include target retrieval, template identification, structural alignment and

Table 12.2 List of potential drug targets (key proteins and enzymes), associated genes and pathways involved in survival and pathogenesis due to drug-resistant *Mycobacterium tuberculosis*

Protein/enzyme	Enzyme classification	Gene	Pathways involved	KEGG orthology	Reference
D-glycero-D-manno-heptose 1,7-bisphosphate phosphatase	3.1.3.82	*gmhB*	Lipopolysaccharide biosynthesis	K03273	Valvano et al. (2002)
D-glycero-alpha-D-manno-heptose-7-phosphate kinase	2.7.1.168	*hddA*	Lipopolysaccharide biosynthesis	K07031	Valvano et al. (2002)
D-sedoheptulose 7-phosphate isomerase	5.3.1.28	*gmhA*	Lipopolysaccharide biosynthesis	K03271	Kneidinger et al. (2002)
UDP-N-acetylglucosamine 1-carboxyvinyltransferase	2.5.1.7	*murA*	Peptidoglycan biosynthesis Amino sugar and nucleotide sugar metabolism	K00790	Cole et al. (1998)
UDP-N-acetylmuramate dehydrogenase	1.3.1.98	*murB*	Peptidoglycan biosynthesis Amino sugar and nucleotide sugar metabolism	K00075	Cole et al. (1998)
UDP-N-acetylmuramate--alanine ligase	6.3.2.8	*murC*	Peptidoglycan biosynthesis D-Glutamine and D-glutamate metabolism	K01924	Cole et al. (1998)
UDP-N-acetylmuramoylalanine--D-glutamate ligase	6.3.2.9	*murD*	Peptidoglycan biosynthesis D-Glutamine and D-glutamate metabolism	K01925	Cole et al. (1998)
UDP-N-acetylmuramoyl-tripeptide--D-alanyl-D-alanine ligase	6.3.2.10	*murF*	Peptidoglycan biosynthesis, Lysine biosynthesis Vancomycin resistance	K01929	Cole et al. (1998)
D-alanine-D-alanine ligase	6.3.2.4	*ddlA*	D-alanine metabolism, Peptidoglycan biosynthesis Vancomycin resistance	K01921	Bruning et al. (2011)
Alanine racemase	5.1.1.1	*alr*	D-Alanine metabolism Vancomycin resistance	K01775	LeMagueres et al. (2005)
Phospho-N-acetylmuramoyl-pentapeptide-transferase	2.7.8.13	*murX*	Peptidoglycan biosynthesis Vancomycin resistance	K01000	Cole et al. (1998)

(continued)

Table 12.2 (continued)

Protein/enzyme	Enzyme classification	Gene	Pathways involved	KEGG orthology	Reference
UDP-N-acetylmuramoyl-L-alanyl-D-glutamate--2,6-diaminopimelate	6.3.2.13	*murE*	Peptidoglycan biosynthesis Lysine biosynthesis	K01928	Cole et al. (1998)
UDP-N-acetylglucosamine--N-acetylmuramyl-(pentapeptide) pyrophosphoryl-undecaprenol N-acetylglucosamine transferase	2.4.1.227	*murG*	Peptidoglycan biosynthesis Vancomycin resistance	K02563	Cole et al. (1998)
Undecaprenyl-diphosphatase	3.6.1.27	*bacA*	Peptidoglycan biosynthesis	K19302	Kumar et al. (2012)
Serine-type D-Ala-D-Ala carboxypeptidase	3.4.16.4	*dacB*	Peptidoglycan biosynthesis	K07258	Kumar et al. (2012)
Acetolactate synthase	2.2.1.6	*ilvB*	Valine, leucine and isoleucine biosynthesis Butanoate metabolism C5-branched dibasic acid metabolism Pantothenate and CoA biosynthesis Biosynthesis of secondary metabolites 2-Oxocarboxylic acid metabolism Biosynthesis of amino acids	K01652	Fleischmann et al. (2002)
Succinyl-CoA synthetase	6.2.1.5	*sucC*	Citrate cycle (TCA cycle) Propanoate metabolism C5-Branched dibasic acid metabolism Biosynthesis of secondary metabolites Microbial metabolism in diverse environments Carbon metabolism	K01903	Fleischmann et al. (2002)
3-isopropylmalate dehydratase	4.2.1.33	*leuC*	Valine, leucine and isoleucine biosynthesis C5-branched dibasic acid metabolism Biosynthesis of secondary metabolites 2-Oxocarboxylic acid metabolism Biosynthesis of amino acids	K01704	Manikandan et al. (2011)

3-isopropylmalate dehydrogenase	1.1.1.85	*leuB*	Valine, leucine and isoleucine biosynthesis, C5-branched dibasic acid metabolism Biosynthesis of secondary metabolites 2-Oxocarboxylic acid metabolism Biosynthesis of amino acids	K00052	Singh et al. (2005)
2-phospho-L-lactate guanylyltransferase	2.7.7.68	*cofC*	Methane metabolism Microbial metabolism in diverse environments	K14941	Grochowski et al. (2008)
6-phosphofructokinase	2.7.1.11	*pfkA*	Pentose phosphate pathway Fructose and mannose metabolism Galactose metabolism Methane metabolism Metabolic pathways Biosynthesis of secondary metabolites Microbial metabolism in diverse environments Glycolysis/gluconeogenesis	K21071	Baugh et al. (2015)
Fructose-1,6-bisphosphatase	3.1.3.11	*glpX*	Glycolysis/gluconeogenesis Pentose phosphate pathway Fructose and mannose metabolism Methane metabolism Biosynthesis of secondary metabolites Microbial metabolism in diverse environments Carbon metabolism	K02446	Baugh et al. (2015)
Fructose-bisphosphate aldolase	4.1.2.13	*fba*	Glycolysis/gluconeogenesis Pentose phosphate pathway Fructose and mannose metabolism Methane metabolism Biosynthesis of secondary metabolites Microbial metabolism in diverse environments Carbon metabolism	K01624	Bashiri et al. (2016)

(continued)

Table 12.2 (continued)

Protein/enzyme	Enzyme classification	Gene	Pathways involved	KEGG orthology	Reference
D-3-phosphoglycerate dehydrogenase	1.1.1.95	*serA*	Glycine, serine and threonine metabolism Methane metabolism Microbial metabolism in diverse environments Carbon metabolism Biosynthesis of amino acids	K00058	Graham et al. (2002)
Phosphoserine aminotransferase	2.6.1.52	*serC*	Glycine, serine and threonine metabolism Methane metabolism Vitamin B6 metabolism Microbial metabolism in diverse environments Carbon metabolism Biosynthesis of amino acids	K00831	Graham et al. (2002)
Phosphoserine phosphatase	3.1.3.3	*serB*	Glycine, serine and threonine metabolism Methane metabolism Microbial metabolism in diverse environments Carbon metabolism Biosynthesis of amino acids	K01079	Graham et al. (2002)
Serine hydroxymethyltransferase	2.1.2.1	*GlyA*	Glycine, serine and threonine metabolism Cyanoamino acid metabolism Glyoxylate and dicarboxylate metabolism One carbon pool by folate Methane metabolism Biosynthesis of secondary metabolites Microbial metabolism in diverse environments Carbon metabolism Biosynthesis of amino acids	K00600	Baugh et al. (2015)

2,3-bisphosphoglycerate-dependent phosphoglycerate mutase	5.4.2.11	*gmp1*	Glycolysis/gluconeogenesis Glycine, serine and threonine metabolism Methane metabolism Biosynthesis of secondary metabolites Microbial metabolism in diverse environments Carbon metabolism Biosynthesis of amino acids	K01834	Hallam et al. (2004)
S-(hydroxymethyl)glutathione dehydrogenase	1.1.1.284	*adhB*	Glycolysis/gluconeogenesis Fatty acid degradation Tyrosine metabolism Chloroalkane and chloroalkene degradation Naphthalene degradation Methane metabolism Biosynthesis of secondary metabolites Microbial metabolism in diverse environments Carbon metabolism Degradation of aromatic compounds	K00121	Hallam et al. (2004)
Malate dehydrogenase	1.1.1.37	*mdh*	Citrate cycle (TCA cycle) Cysteine and methionine metabolism Pyruvate metabolism Glyoxylate and dicarboxylate metabolism Methane metabolism Biosynthesis of secondary metabolites Microbial metabolism in diverse environments Carbon metabolism	K00024	Ferraris et al. 2015

(continued)

Table 12.2 (continued)

Protein/enzyme	Enzyme classification	Gene	Pathways involved	KEGG orthology	Reference
Acetate kinase	2.7.2.1	*ackA*	Microbial metabolism in diverse environments Carbon metabolism Methane metabolism Pyruvate metabolism Propanoate metabolism Taurine and hypotaurine metabolism	K00925	Ferraris et al. 2015
Phosphate acetyltransferase	2.3.1.8	*pta*	Microbial metabolism in diverse environments Carbon metabolism Methane metabolism Pyruvate metabolism Propanoate metabolism Taurine and hypotaurine metabolism	K13788	Bashiri et al. (2016)
Acetyl-CoA synthetase	6.2.1.1	*acs*	Glycolysis/gluconeogenesis Microbial metabolism in diverse environments Carbon metabolism Methane metabolism Pyruvate metabolism Propanoate metabolism Taurine and hypotaurine metabolism	K01895	Hallam et al. (2004)

superposition, model prediction, loop modelling, side-chain optimization, model refinement and model validation. The most frequently used computational tools/web-based servers for protein structure prediction by homology modelling include 3D-JIGSAW (Bates et al. 2001), MODELLER (Webb and Sali 2017), HHpred (Söding et al. 2005), RaptorX (Peng and Xu 2011), SWISS-MODEL (Schwede et al. 2003) and Phyre2 (Kelley et al. 2015b). Qiu et al. (2017) have reported the homology model of potential drug target serine acetyltransferase (CysE) from Mtb. The study reported the essential amino acids that are associated with enzymatic activity of CysE to design inhibitors (Qiu et al. 2017). Similarly, Ko and Choi (2016) have reported the 3D structure of QcrB from *Mycobacterium tuberculosis* cytochrome bc1 complex by homology modelling to study the effect of new anti-tuberculosis agent Q203 (Ko and Choi 2016).

12.7.2 Fold Recognition

Fold recognition, also known as protein threading, is a method of structure prediction wherein the three-dimensional structure of proteins is predicted on the basis of folds. This process utilizes experimentally elucidated structure of proteins that have similar folds. Both fold recognition and homology modelling are template-based prediction methods (Vyas et al. 2012); however, it differs from the homology modelling approach, as it is used for proteins which do not have similar protein structures deposited in any of the structural databases (Liu et al. 2011). Further, the prediction in this method is carried out by aligning each amino acid in the target sequence to a position in the template structure and evaluating how well the target fits the template (Leelananda and Lindert 2016). The predicted structure is then evaluated by using various scoring methods, and this process is reiterated for all 3D structures in a structural database until the best structural fit is obtained for a given query (Usha et al. 2017). Some of the commonly used computational tools/web servers for protein structure prediction by fold recognition include MUSTER (Wu and Zhang 2008), GenTHREADER (Jones 1999), I-TASSER (Yang et al. 2015) and DescFold (Yan et al. 2009). In a study conducted by Mao et al. (2013), reported the predict 3D fold and structure of several proteins in the genome of Mtb H37Rv strain using Phyre2 tool (Mao et al. 2013).

12.7.3 Ab Initio Prediction

Ab initio or de novo methods predict a 3D structure directly from the amino acid sequence without the usage of the template. Template-based structure prediction methods do not require computationally intensive settings, whereas ab initio methods make use of high GPU, and the prediction is restricted to smaller proteins (<120 residues) (Usha et al. 2017). Ab initio prediction is carried out in two steps: first, by formulating a scoring function (such as energy) that separates accurate (native-like or native) structures from incorrect ones, and, second, by devising a search method for exploring the conformational space (Leelananda and Lindert 2016). It has been observed that the template-based methods predict accurate structures in comparison to de novo methods for structure prediction (Liu et al. 2011). Further, these 3D structures are validated and explored for molecular docking and MD simulation studies in the process of structure-based drug discovery. There are several web servers and stand-alone softwares for both prediction and validation of 3D structures. Some of the commonly used tools/web servers for protein structure prediction by ab initio method include QUARK (Xu and Zhang 2012), I-TASSER (Yang et al. 2015), Rosetta/Robetta (Bradley et al. 2005), CABS-FOLD (Blaszczyk et al. 2013) and EVfold (Marks et al. 2011). These predicted 3D structures are evaluated using various bioinformatics tools or web servers such as ProCheck, WHATIF, ERRAT, PROVE, ANOLEA, GROMOS, Verify3D, ProMotif, DSSP, QMEAN and ProSA (Leelananda and Lindert 2016).

12.8 Virtual Screening of Novel Lead-Like Molecules Against *Mycobacterium tuberculosis*

Conventionally, experimental methods such as high-throughput screening are being employed for rapid identification of lead molecules against drug targets by performing individual biochemical assays for several compounds. However, there are several drawbacks of these processes, such as being time-consuming, expensive and laborious in nature and these can be surmounted by the integration of computer-aided virtual screening (Kar and Roy 2013). Virtual screening is defined as an exhaustive process of screening extensive libraries of compounds for identification of new lead molecules against biological targets by techniques such as computer-aided molecular drug design, pharmacophore searches, homology modelling, high-throughput docking and MD simulations (Lionta et al. 2014; Geromichalos 2012). Recently, virtual screening techniques have become a key component of modern drug discovery and development process. Further, it has also been adopted by pharmaceutical industries and academic groups in the early stages of drug discovery and development to screen undesirable compounds which otherwise result in expensive and time-consuming experimental methods (Cheng et al. 2012; Geromichalos 2012). Virtual screening methods are further divided into structure-based and ligand-based. The underlying principle behind structure-based virtual screening is molecular docking and interaction analysis (usually, protein-ligand interactions). This process involves automated and fast docking of several compounds against a given biological target. In this approach, an accurate understanding about the active site or the binding cavities of the target protein is essential. Similarly, ligand-based virtual screening is based on similarity, topological and pharmacophore substructure searches against various databases (Reddy et al. 2007). Some of the commonly screened databases include PubChem (https://pubchem.ncbi.

nlm.nih.gov/), ZINC (http://zinc.docking.org/), ACD (http://accelrys.com/products/collaborative-science/databases/sourcing-databases/biovia-available-chemicals-directory.html), ChemSpider (http://www.chemspider.com/), ChEMBL (https://www.ebi.ac.uk/chembl/), Enamine (http://www.enamine.net/), ChemNavigator (https://www.chemnavigator.com/), and TCM (http://tcm.cmu.edu.tw/). It is also known as neighbourhood behaviour search and is widely employed for identification of novel drug-like candidates against the disease. It has been observed that there is no universal protocol to follow in the process of virtual screening of novel drug-like molecules. However, having in-depth knowledge about the targets and through identifying the need for the study, the process can be altered to obtain reliable results (Kar and Roy 2013).

In a recent study conducted by Kaur et al. (2018), by drug-targeted virtual screening and MD simulations, the authors have identified several inhibitors of LipU protein (a key protein in the survival of Mtb) (Kaur et al. 2018). Similarly, Sengupta et al. (2015) have performed pharmacophore-based virtual screening and molecular dynamic simulations to identify potential inhibitors of maltosyl transferase (GlgE) in Mtb (Sengupta et al. 2015). Further, Maganti et al. (2015) have reported 3D-QSAR and shape-based virtual screening of novel inhibitors against aryl acid adenylating enzyme (MbtA) involved in the biosynthesis of siderophores (Maganti et al. 2015). This study illustrated the molecular dynamics simulations to gain more insights about the stability of the ligand-receptor complexes. The application of virtual screening, molecular modelling, molecular docking and MD simulations has been well established in identifying novel leads against Mtb (Janardhan et al. 2017; Lone et al. 2017c; Mansuri et al. 2016; Mehra et al. 2016).

Some of the major drawbacks of this approach include ligand/target flexibility, studying multiple binding modes, consideration of solvent parameters, variability in the scoring functions, tautomerization and ionization of ligand and protein residues and the solvation effects (Cheng et al. 2012; Shoichet 2004). Nevertheless, virtual screening has become a valuable and crucial part of drug discovery process and, perhaps, is the most practical approach in identifying novel leads against targets of XDR-Mtb (Reddy et al. 2007).

12.9 Computational Pharmacokinetic Analysis: Prediction of Drug Likeliness and ADMET Properties

Prior to studying the drug-like features, the identified molecules are subjected to several molecular descriptors based on topology (connectivity and balaban indices), constitution (molecular weight, rotatable bonds, and H-bond acceptors/donors), lipophilicity (ocatanol-water partition coefficient), geometry (polar and volume related surface area), thermodynamics (heat of formation and molar refractivity) and electronic descriptors (partial charges and dipole moment) to screen the molecules for further steps (Vyas et al. 2008). Further, the drug-likeliness features of the lead molecules are usually predicted on the basis of Lipinski's rule of five (Lipinski 2004), Lead-like rule (Lipinski 2004), Comprehensive Medicinal Chemistry (CMC)-like rule (Ghose et al. 2006), World Drug Index (WDI)-like rule (Sneader 1990) and MDL Drug Data Report (MDDR)-like rule (Oprea 2000). Similarly, the molecules that qualify the initial screening are further evaluated for adsorption, distribution, metabolism and excretion (ADME) by various statistical models such as human intestinal absorption (HIA%), blood-brain barrier (BBB) penetration (Clark 2003), buffer solubility, heterogeneous human epithelial colorectal adenocarcinoma (caco2) cell permeability, plasma protein binding (Leeson et al. 2004), Madin Darby canine kidney (MDCK) cell permeability, P-glycoprotein inhibition, CYP 2C19 inhibition, pure water solubility and skin permeability assays (Averbukh et al. 2014; Bickerton et al. 2012; Veber et al. 2002). Further, the lead molecules that possess good ADME properties are further selected to predict the toxicity in terms of mutagenicity (based on Ames parameters), rodent carcinogenicity in mouse and rat models, hERG gene inhibition, acute fish toxicity in minnow (*Pimephales notatus*) and medaka (*Oryzias latipes*), acute algae and daphnia toxicity assays.

Some of the commonly used computational tools and web servers for the prediction of drug-like and ADMET (absorption, distribution, metabolism, excretion and toxicity) features include QikProp (https://www.schrodinger.com/qikprop), SwissADME (http://www.swissadme.ch/), PreADMET (https://preadmet.bmdrc.kr/), ADMEWORKS ModelBuilder (http://www.fqs.pl/en/chemistry/products/admeworks-modelbuilder), DrugLogit (http://hermes.chem.ut.ee/~alfx/druglogit.html), AdmetSAR (http://lmmd.ecust.edu.cn/admetsar1/) and ADMET Predictor 8.5 (https://www.simulations-plus.com/software/admetpredictor/).

In this chapter, the authors have illustrated drug-likeliness features and ADMET features of the few herbal leads by PreADMET web server. The drug-likeliness properties of all the identified lead molecules are displayed in Table 12.3. Out of the four herbal-based compounds screened, all the compounds were qualified for drug-like properties according to Lipinski's rule of five and CMC-like rule and displayed mid-structure as per MDDR-like rule. However, only strictamin qualified lead-like rule. Further, ajmalicine and strictamin were in the 90% cut-off range for WDI-like rule. All the four herbal leads possessed drug-likeliness properties. The ADME properties of the lead molecules are displayed in Table 12.4. The computational prediction suggested that all the molecules displayed ideal penetration across the blood-brain barrier (BBB) (low-level penetration to the suggested targets), exhibited higher bioavailability and were easily absorbed by human intestine. The probability values of the four molecules predicted by caco2 cell permeability model demonstrated that they were within the acceptable range of 20–39 which suggested good intestinal absorption. Buffer solubility and pure water solubility for strictamin (a herbal compound present in *Alstonia scholaris*) were predicted to be 649 mg/L and 416 mg/L, respectively. The prediction by skin permeability model suggested that the molecules were in the acceptable range of −2.3 cm/hour (curcumin) to −4.3 (ajmalicine) cm/hour. The toxicity profiles of the identified leads are displayed in Table 12.5. In vitro values of acute algae (algae_at) and daphnia (daphnia_at) toxicity were predicted to be within the acceptable range of 0.01 (curcumin) to 0.1 (limonin). The toxicity prediction for ajmalicine (a herbal compound present in

Table 12.3 Computer-aided drug likeliness prediction of molecules from herbal origin such as *Rauvolfia serpentine*, *Curcuma longa*, *Alstonia scholaris* and *Vitis vinifera* using PreADMET web server

Molecules	PubChem ID (CID)	Herbal source (common name)	Molecular Weight (Da)	Rule of five	CMC-like rule	Lead-like rule	MDDR-like rule	WDI-like rule
Ajmalicine	251561	*Rauvolfia serpentina* (Sarpagandha)	352.42	Suitable	Qualified	Violated	Mid-structure	In 90% cut-off
Curcumin	969516	*Curcuma longa* (Turmeric)	368.37	Suitable	Qualified	Violated	Mid-structure	Out of 90% cut-off
Strictamin	6444325	*Alstonia scholaris* (Saptaparna)	322.40	Suitable	Qualified	Suitable	Mid-structure	In 90% cut-off
Limonin	179651	*Vitis vinifera* (Grapes)	470.51	Suitable	Qualified	Violated	Mid-structure	Out of 90% cut-off

Table 12.4 Computer-aided ADME prediction results (using PreADMET web server) for herbal leads screened against drug-resistant *Mycobacterium tuberculosis*

Ligand	PubChem ID (CID)	BBB (C_{brain}/ C_{blood})[a]	Buffer solubility (mg/L)	Caco2 (nm/s)[b]	CYP 2C19	HIA[c] (%)	MDCK (nm/s)[d]	PPB (%)[e]	Pure water solubility (mg/L)	Skin permeability (log kp, cm/h)[f]
Ajmalicine	251561	1.98	413.32	39.46	Non-inhibitor	93.31	27.77	55.5	184	−4.3
Curcumin	969516	0.0913	7014.27	20.07	Inhibitor	94.4	99.98	88.03	10.8	−2.33
Strictamin	6444325	2.003	649	28	Non-inhibitor	97.6	185	58.9	416	−4.01
Limonin	179651	0.122	804.42	27.67	Non-inhibitor	96.25	0.788	80.27	7.16	−3.73

[a] In vivo blood-brain-barrier penetration—(C_{brain}/C_{blood}) for high absorption to CNS >2.0; middle adsorption to CNS: 2.0 ≈ 0.1; low absorption to CNS <0.1

[b] In vivo caco2 cell permeability—low <4; middle: 4–7; high >7

[c] Human intestinal (HIA%) absorption—poor: 0–20%; moderate: 20–70%; well: 70–100%

[d] In vivo MDCK cell permeability—low <25; middle: 25–500; high >500

[e] In vivo plasma protein binding—weakly bound: <90%; strongly bound: >90%

[f] In vivo skin permeability—low <1; middle: 1–2; high >2.0

Table 12.5 Computer-aided toxicity prediction (using PreADMET web server) for potential lead molecules against drug-resistant *Mycobacterium tuberculosis*

Ligand	PubChem ID (CID)	Acute algae toxicity (algae_at)	Ames test	Carcinogenicity test		Acute daphnia toxicity (daphnia_at)	hERG inhibition
				Mouse	*Rat*		
Ajmalicine	251561	0.0484	Mutagen	Negative	Negative	0.1210	Medium risk
Curcumin	969515	0.0188	Non mutagen	Negative	Positive	0.0387	Medium risk
Strictamin	6444325	0.0954	Mutagen	Negative	Negative	0.2034	Medium risk
Limonin	179651	0.1007	Non mutagen	Negative	Positive	0.5191	Low risk

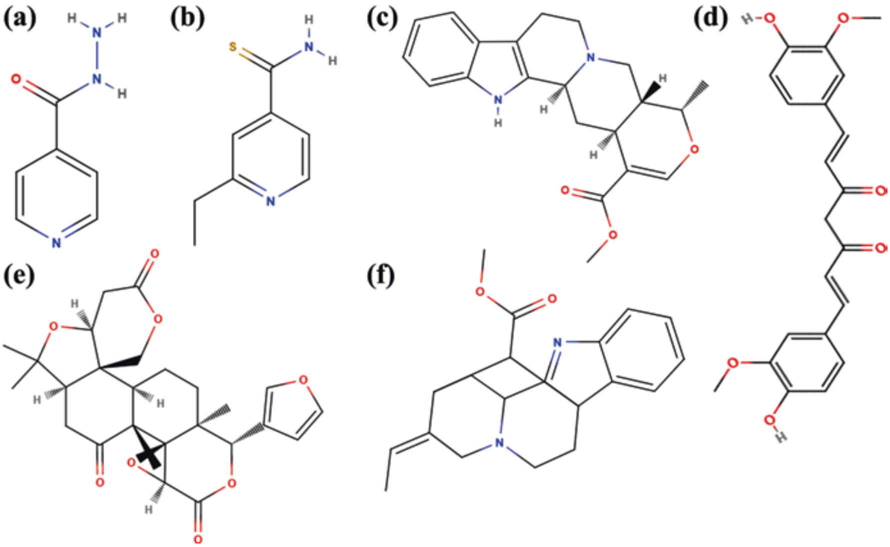


Fig. 12.3 Structural representation of currently prescribed antibiotics and potential natural lead compounds (**a**) isoniazid, (**b**) ethionamide, (**c**) ajmalicine, (**d**) curcumin, (**e**) limonin, (**f**) strictamin

Rauvolfia serpentina) and strictamin (a herbal compound present in *Alstonia scholaris*) were predicted to be non-carcinogenic in both mouse and rat models making them potential leads. Similarly, the prediction suggested that limonin (a herbal compound present in *Vitis vinifera*) displayed low risk for the inhibition of hERG gene, while the other leads displayed medium risk. The hERG gene codes for the α-subunit of potassium ion channel in humans. Besides curcumin and limonin, all the selected molecules were predicted to be mutagenic according to Ames test. Hence, computational analysis suggested that the herbal leads such as strictamin (*Alstonia scholaris*), ajmalicine (*Rauvolfia serpentina*), limonin (*Vitis vinifera*) and curcumin (*Curcuma longa*) qualified for drug likeliness and ADMET and can be further selected to study the interactions between the drug target by molecular docking and interaction studies. Similarly, the chemical structures of the antibiotics such as isoniazid and ethionamide are also used for the comparative analysis, and the 2D structures of the antibiotics and herbal-based leads are displayed in Fig. 12.3.

Lone et al. (2017b) have made an effort to identify potential inhibitors from herbal sources against the probable drug target, 3-dehydroquinate dehydratase (DHQase) of Mtb. The study constructed pharmacophore models and reported the probable interactions by molecular docking studies and performed in vitro assays to validate the findings (Lone et al. 2017b). Another study identified nine lead compounds against InhA by performing a pharmacophore-based virtual screening of the SPECS natural product database. Further, they have performed molecular dynamic simulations and quantum chemical studies of the nine leads to understand structural features essential for the activity (Lone et al. 2017a). Hence, computer-aided poly-pharmacological approach can be applied to screen several inhibitors against multiple drug targets of Mtb. A combination of pharmacophore and QSAR-based virtual screening strategy was undertaken to screen compounds from Asinex database (435,000 compounds) against three drug targets InhA, GlmU and DapB in Mtb. Further, these potential hits were studied in detail by molecular docking analysis (Janardhan et al. 2017).

12.10 Molecular Docking Analysis

In structure-based drug design, molecular docking analysis plays a pivotal role in the process of virtual screening for hit identification and in the drug discovery process for optimization of potential leads (Pagadala et al. 2017). Molecular docking studies help in predicting the orientation of the ligand when it binds to an enzyme or a receptor (Chaudhary and Mishra 2016). This methodology is widely employed to explore the behaviour of small molecules (ligand/inhibitors) in the active site of a receptor (de-Ruyck et al. 2016). It is increasingly being used as a tool in drug discovery and development process, as it can be used for both experimental structures and theoretical models (Meng et al. 2011).

Molecular docking procedures can be carried out in two ways either through flexible-body or rigid-body docking approach. In the first approach, both the ligand and receptor are conformationally flexible and allowed to rotate along multiple degrees of freedom. Secondly, in rigid-body docking, both the ligand and receptor are held static during the process (de-Ruyck et al. 2016). Similarly, the two major underlying principles involved in the docking studies are conformation search (by various algorithms such as point complementary, Monte Carlo, fragment-based genetic algorithms, systematic searches and distance geometry) and a scoring function (either empirical-based, force field-based, consensus-based or knowledge-based) to evaluate the binding efficiency of ligand towards a target (Dar and Mir 2017). Further, key interactions such as hydrogen bonds, hydrophobic interactions, van der Waals forces and electrostatic forces (charge-dipole, dipole-dipole and charge-charge) are taken into consideration during docking analysis (Chaudhary and Mishra 2016). Furthermore, the results obtained from molecular docking analysis include various parameters such as number of electrostatic forces, number of hydrogen bonds and the negative binding energy which is usually measured in terms of kcal/mol (Dar and Mir 2017). Similarly, the information obtained from these docking studies further help in understanding whether an inhibitor will be able to bind in the active site of key enzymes majorly responsible for pathogenesis (Meng et al. 2011). Likewise, it can be a powerful process for studying the specificity and binding of potential lead molecules against selected drug targets (Ferreira et al. 2015). The list of commonly used molecular docking softwares and tools has been displayed in the Table 12.6.

The authors have tried to elucidate the binding potential perdition of herbal-based molecules towards putative drug targets (GlyA) of Mtb identified by molecu-

Table 12.6 List of commonly used protein-ligand docking programs/software in computer-aided drug discovery and development

Docking program/software	Year of release	Type of license	References
DOCK	1988	Freeware	Ewing et al. (2001)
AutoDock	1990	Freeware	Morris et al. (1998)
SOFTDocking	1991	Academic	Jiang and Kim (1991)
DockVision	1992	Commercial	Hart and Read (1992)
LUDI	1992	Academic	Bohm (1992)
ADAM	1994	Commercial	Mizutani et al. (1994)
FLOG	1994	Academic	Miller et al. (1994)
DIVALI	1995	Freeware	Clark (1995)
GOLD	1995	Commercial	Jones et al. (1997)
Hammerhead	1996	Academic	Welch et al. (1996)
LIGIN	1996	Commercial	Sobolev et al. (1996)
FTDOCK	1997	Freeware	Gabb et al. (1997)
ICM-Dock	1997	Commercial	Totrov and Abagyan (1997)
QXP	1997	Academic	McMartin and Bohacek (1997)
SANDOCK	1998	Academic	Burkhard et al. (1998)
MCDOCK	1999	Academic	Liu and Wang (1999)
PRODOCK	1999	Academic	Trosset and Scheraga (1999)
DARWIN	2000	Freeware	Taylor and Burnett (2000)
EUDOC	2001	Academic	Pang et al. (2001)
PatchDock	2002	Freeware	Schneidman-Duhovny et al. (2005)
FDS	2003	Academic	Taylor et al. (2003)
FRED	2003	Academic	McGann et al. (2003)
HADDOCK	2003	Freeware	Dominguez et al. (2003)
LigandFit	2003	Commercial	Venkatachalam et al. (2003)
Surflex-Dock	2003	Commercial	Spitzer and Jain (2012)
iGEMDOCK	2004	Freeware	Yang and Chen (2004)
Glide	2004	Commercial	Halgren et al. (2004)
YUCCA	2005	Academic	Choi (2005)
eHiTS	2006	Commercial	Zsoldos et al. (2007)
MolDock	2006	Academic	Thomsen and Christensen (2006)
PLANTS	2006	Academic	Korb et al. (2006)
PSI-DOCK	2006	Academic	Pei et al. (2006)
EADock	2007	Freeware	Grosdidier et al. (2007)
FLIPDock	2007	Academic	Zhao and Sanner (2007)
MEDock	2007	Freeware	Chang et al. (2005)
ParDOCK	2007	Freeware	Gupta et al. (2007)
PSO@AUTODOCK	2007	Academic	Namasivayam and Gunther (2007)

(continued)

Table 12.6 (continued)

Docking program/software	Year of release	Type of license	References
SODOCK	2007	Academic	Chen et al. (2007)
Lead finder	2008	Commercial	Stroganov et al. (2008)
Molecular Operating Environment (MOE)	2008	Commercial	Vilar et al. (2008)
MS-DOCK	2008	Academic	Sauton et al. (2008)
PLATINUM	2008	Freeware	Pyrkov et al. (2009)
HomDock	2008	Freeware	Marialke et al. (2008)
Q-Dock	2009	Freeware	Brylinski and Skolnick (2008)
DOCK Blaster	2009	Freeware	Irwin et al. (2009)
DockingServer	2009	Commercial	Hazai et al. (2009)
AutoDock Vina	2010	Open source	Trott and Olson (2010)
FlexPepDock	2010	Freeware	London et al. (2011)
AADS	2011	Freeware	Singh et al. (2011)
BetaDock	2011	Freeware	Kim et al. (2011)
iScreen	2011	Freeware	Tsai et al. (2011)
LigDockCSA	2011	Academic	Shin et al. (2011)
PythDock	2011	Academic	Chung et al. (2011)
SwissDock	2011	Academic	Grosdidier et al. (2011)
VoteDock	2011	Academic	Plewczynski et al. (2011)
Pose & Rank	2011	Freeware	Fan et al. (2011)
BSP-SLIM	2012	Freeware	Lee and Zhang (2012)
idTarget	2012	Freeware	Wang et al. (2012)
Fleksy	2012	Freeware	Wagener et al. (2012)
ParaDockS	2012	Open source	Pippel et al. (2012)
rDock	2013	Open source	Ruiz-Carmona et al. (2014)
FlexAID	2015	Open source	Gaudreault and Najmanovich (2015)
POSIT	2015	Academic	Kelley et al. (2015a)
MOLS 2.0	2016	Open source	Paul and Gautham (2016)
Galaxy7TM	2016	Freeware	Lee and Seok (2016)
HybridDock	2016	Academic	Huang et al. (2016)
GalaxyDock BP2 score	2017	Freeware	Baek et al. (2017)

lar modelling study and compared to the binding of conventional antibiotics to their respective targets of Mtb. The molecular docking analysis was carried out for four herbal leads against the potential drug target serine hydroxymethyltransferase (GlyA) (PDB: 3H7F) and further compared with the binding interaction of two conventionally prescribed antibiotics such as isoniazid (pyridine-4-carbohydrazide) and ethionamide (2-ethylpyridine-4-carbothioamide) against their drug target enoyl-[acyl-carrier-protein] reductase NADH (InhA) (PDB: 4DRE). The binding site for each drug target was predicted by DEPTH web server (Tan et al. 2013). Some of the other commonly used tools for predication of binding cavities include

CASTp (Dundas et al. 2006), MetaPocket (Huang 2009), Q-SiteFinder (Laurie and Jackson 2005), MDpocket (Schmidtke et al. 2011) and SURFNET (Laskowski 1995). Further, a flexible-body docking was performed using AutoDock Vina v1.1.2, and the grid dimensions for the binding cavity of the receptors were performed as per standard protocols (Trott and Olson 2010). The best docked poses were selected on the basis of key interacting residues: cluster RMS, number of hydrogen bonds and minimum binding energy (kcal/mol).

The binding potential of four selected herbal leads towards GlyA and the antibiotics isoniazid and ethionamide against their drug target InhA has been displayed in Table 12.7 and Fig. 12.4. Molecular docking analysis suggested that limonin (7,16-Dioxo-7,16- dideoxylimondiol), commonly present in *Citrus* species demonstrated the best binding energy of −7.2 kcal/mol against GlyA with Tyr61 as the key interacting residue (Fig. 12.4e). Limonin is a vital component in the seeds of citrus fruits, and it has exhibited its pharmacological activity against several pathogenic Gram-positive and Gram-negative bacteria (Skariyachan et al. 2018; Ayaz et al. 2017). The docked complex of ajmalicine ((19α)-16,17-didehydro-19-methyloxayohimban-16-carboxylic acid methyl ester) and GlyA displayed a promising binding energy of −6.7 kcal/mol with Leu118, Ala119, Leu320 and Gly361 as key interacting residues (Fig. 12.4c). Ajmalicine is a naturally occurring alkaloid that is commonly present in *Rauwolfia serpentina*, *Mitragyna speciose* and *Catharanthus roseus* (Wink 2015; Nazzaro et al. 2013). This naturally occurring compound has exhibited broad-spectrum activity against both Gram-negative and Gram-positive bacteria (Wink 2015), and hence, it can be considered as a potential lead against various targets of Mtb. Further, strictamin (akuammilan-17-oic acid methyl ester) when docked with GlyA displayed the binding energy of −6.5 kcal/mol with Leu320, Arg363, Val321, Gly361 and Val310 as key interacting residues and one stabilizing hydrogen bond (Fig. 12.4f). Similarly, when the antibiotics isoniazid and ethionamide were docked with InhA, they revealed binding potential of −4.2 kcal/mol and −4.7 kcal/mol, respectively. The key interacting residues for ethionamide and InhA were observed to be Val175, Ala128 and Lys132 (Fig. 12.4b). Further, it was observed that the theoretical binding energy of herbal-based molecules and GlyA was found to be better that of the binding energy of the selected antibacterial and respective targets. Hence, from virtual screening and molecular docking analysis, it can be suggested that the herbal-based lead molecules displayed promising binding potential with minimum binding energy and stabilising interactions in comparison with the binding of two standard antibiotics towards their usual targets.

Rajendran and Sethumadhavan (2014) have analysed the role of bacterial enzyme pyrazinamidase (PncA) in pyrazinamide resistance by various computational analysis. They have studied the binding pocket analysis, solvent accessibility analysis, molecular docking and interaction analysis to understand the behaviour of mutant pyrazinamidase in MDR-Mtb. Further, the authors have also reported molecular dynamic simulations of this enzyme to understand the three-dimensional (3D) conformational behaviour during drug resistance and pathogenesis in Mtb (Rajendran and Sethumadhavan 2014). Similarly, Fakhar et al. (2016) have reported potential

Table 12.7 The binding of selected herbal-based leads and antibiotics (isoniazid and ethionamide) against serine hydroxymethyltransferase (GlyA) and enoyl-[acyl-carrier-protein] reductase NADH (InhA), respectively using AutoDock Vina

Antibiotics/ Herbal lead	IUPAC name	Target	Function	Interacting residues	No. of Hydrogen bonds	Binding energy (kcal/mol)
Isoniazid	Pyridine-4-carbohydrazide	Enoyl-[acyl-carrier-protein] reductase NADH (inhA) (PDB: 4DRE)	Drug target of the first line anti-TB drug isoniazid and second line drug ethionamide	Met155	0	−4.2
Ethionamide	2-Ethylpyridine-4-carbothioamide			Val175, Ala128, Lys132	0	−4.7
Ajmalicine	(19α)-16,17-didehydro-19-methyloxayohimban-16-carboxylic acid methyl ester	Serine hydroxymethyltransferase (glyA) (PDB: 3H7F)	Key enzyme in the biosynthesis of glycine, serine and threonine	Leu118, Ala119, Leu320, Gly361	0	−6.7
Limonin	7,16-Dioxo-7,16-dideoxylimondiol			Tyr61	0	−7.2
Curcumin	(1*E*,6*E*)-1,7-Bis(4-hydroxy-3-methoxyphenyl) hepta-1,6-diene-3,5-dione			Leu56, Arg59	0	−4.1
Strictamin	Akuammilan-17-oic acid methyl ester			Leu320, Arg363, Val321, Gly361, Val310	1	−6.5

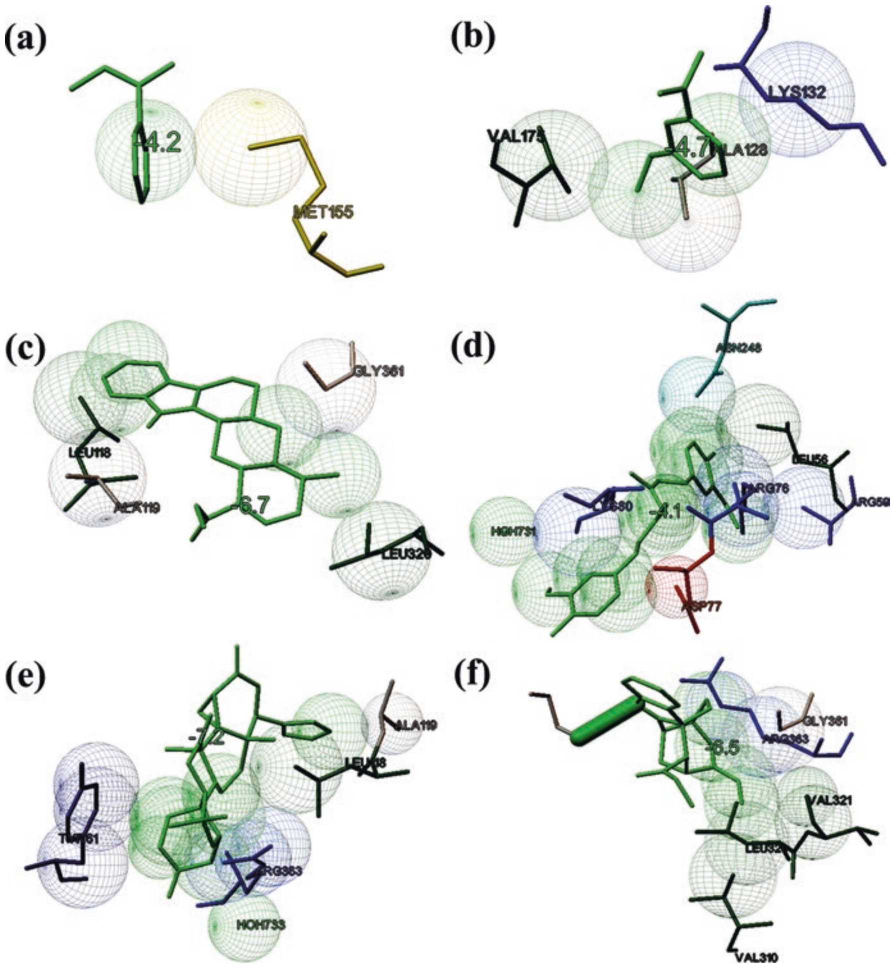


Fig. 12.4 The binding potential of currently prescribed antibiotics isoniazid and ethionamide with InhA and herbal leads ajmalicine, curcumin, limonin and strictamin with GlyA of *Mycobacterium tuberculosis*. (**a**) The binding energy for the best docked pose of isoniazid and InhA was observed to be −4.2 kcal/mol with Met155 as a key interacting residue. (**b**) The best docked pose of ethionamide and InhA with Val175, Ala128 and Lys132 as key interacting residues and a bind energy of −4.7 kcal/mol. (**c**) The binding energy for the best docked pose of ajmalicine and GlyA was predicted to be −6.7 kcal/mol. The key interacting residues were observed to be Leu118, Ala119, Leu320 and Gly361. (**d**) The binding energy for the best docked pose of curcumin and GlyA was predicted to be −4.1 kcal/mol. The key interacting residues were observed to be Leu56 and Arg59. (**e**) The binding energy for the best docked pose of limonin and GlyA was observed to be −7.2 kcal/mol with Tyr61 as a key interacting residue. (**f**) The best docked pose of strictamin and GlyA with a stabilizing hydrogen bond. The key interacting residues were observed to be Leu320, Arg363, Val321, Gly361 and Val310 with a binding energy of −6.5 kcal/mol

drug targets such as MurG, MurI, MraY, DapA, DapE, Ddl and Alr involved in the biosynthesis of peptidoglycan cell wall of MDR-Mtb. The 3D structures of these essential enzymes were predicted by homology modelling using Modeller 9v13. Further, the structural qualities of these models were validated by PDBsum, PROCHECK, ERRAT and QMEAN. The study further performed molecular docking and MD simulations to understand the interaction between the enzymes and their potential inhibitors (Fakhar et al. 2016). Similarly, in vitro anti-tubercular activity of five medicinal plants *Acalypha indica*, *Adhatoda vasica*, *Allium cepa*, *Allium sativum* and *Aloe vera* against MDR-Mtb has been reported (Gupta et al. 2007).

Ramesh et al. (2008) have reported a bio-computational study to understand the binding mode of anti-TB herbal ligands against the homology model of fatty acid synthase of Mtb H37Rv strain. The 3D structure of this protein was predicted using the Modeller package to study the ligand-receptor interactions. Further, molecular docking studies suggested that different herbal ligands such as aloe-emodin and nimbin are the best herbal candidates to replace the synthetic drugs thiolactomycin or cerulenin that are prescribed against Mtb (Ramesh et al. 2008). Hence, the receptor-ligand interactions can be easily studied by assessing thousands of potential conformations possessed by the process of molecular docking analysis. Although, molecular docking is a widely accepted approach in the process of structure-based drug discovery and development, there are several shortcomings which can be surmounted by molecular dynamic simulation studies (Pagadala et al. 2017).

12.11 Molecular Dynamic Simulations

A major limitation of molecular docking analysis is that the protein is held static during the process (Liu et al. 2018). The static models obtained by various experimental methods or through homology modelling provide vital information about the macromolecular structure. However, when a drug binds to its receptor in vivo, it does not encounter a frozen model, but rather a structure that is constantly in motion (Durrant and McCammon 2011). Molecular docking studies are not considered the dynamic motions of the receptor-ligand complex and can be overcome by another computational method known as molecular dynamic (MD) simulations (De-Vivo et al. 2016). MD simulations can be employed to identify allosteric binding sites, to understand the structure and functional association of receptor-ligand interactions, to study the mechanism of drug resistance and to provide accurate binding mode through optimization of lead compounds (Liu et al. 2018). Similarly, MD simulations can be applied to generate a set of reliable structures for analysis when a 3D structure for a particular target is unavailable or the binding sites are poorly defined (Ferreira et al. 2015). These studies allow both receptor and ligand(s) to alter their biological conformations in the receptor-ligand complex (Durrant and McCammon 2011). Molecular dynamic simulation studies make use of the most popular force fields such as Assisted Model Building with Energy Refinement (AMBER) (Case et al. 2017), GROningen MAchine for Chemical Simulations (GROMACS) (Abraham et al. 2015), Nanoscale Molecular Dynamics (NAMD) (Phillips et al. 2005) and Chemistry at HARvard Macromolecular Mechanics (CHARMM) (Brooks et al. 2009). These studies can be carried out using various programs such as Amsterdam Density Functional (ADF) and Abalone, Desmond and Large-scale Atomic/Molecular Massively Parallel Simulator (LAMMPS) (De-Vivo et al. 2016; Lindorff-Larsen et al. 2010).

In a recent study conducted by Naz et al. (2018), suggested that a novel benzamide inhibitor against *α*-subunit of tryptophan synthase (α-TRPS) was explored from Mtb by structure-based virtual screening, molecular docking and MD simulations (Naz et al. 2018). In another study conducted by Pandey et al. (2018), the authors have carried out structure-based molecular docking, molecular mechanics/generalized Born surface area prediction and MD simulations to study the mechanism behind fluoroquinolone resistance in MDR-Mtb. This study provides keys insights into the mechanism of drug resistance and identifying potential lead molecules against MDR-Mtb (Pandey et al. 2018). Multi-computer approaches such as grid computing, workstation clusters, personal computer clusters and massive parallel processors (MPP) facilitate CADD, yet, only large research groups or national research centres can afford these systems due to their high investment costs (Hung and Chen 2014). Nevertheless, MD simulations assist in several key drug discovery steps by undergoing continuous improvements in both computer power (increased GPU and cloud computing) and algorithm design (Liu et al. 2018).

12.12 Conclusions

XDR-Mtb has proven to be resistant against the majority of currently prescribed antibiotics, and hence, discovering compounds with antibacterial activity against potential drug targets is crucial in combating tuberculosis. The integration of databases and omics technologies helps in the rapid screening of potential drug targets and network-based novel multi-target drugs. Similarly, virtual screening has become an integral part of the drug discovery field in screening and optimization of lead molecules. This chapter illustrated that the herbal lead molecules possess better binding potential towards the putative targets of Mtb, which was identified by metabolic pathways analysis, in comparison with the binding of two conventional antibiotics and their respective targets. Thus, it can be suggested that herbal molecules such ajmalicine, curcumin, limonin and strictamin can be used as alternative lead molecules against the key enzyme serine hydroxymethyltransferase in Mtb. Furthermore, this chapter not only provides information about the latest developments in molecular medicine and computational drug discovery to combat tuberculosis but also opens a new paradigm towards the screening and development of novel leads against potential drug targets for XDR-Mtb.

References

Abraham MJ, Murtola T, Schulz R, Páll S, Smith JC, Hess B, Lindahl E (2015) GROMACS: high performance molecular simulations through multi-level parallelism from laptops to supercomputers. SoftwareX. https://doi.org/10.1016/j.softx.2015.06.001

Agyeman AA, Ofori-Asenso R (2016) Efficacy and safety profile of linezolid in the treatment of multidrug-resistant (MDR) and extensively drug-resistant (XDR) tuberculosis: a systematic review and meta-analysis. Ann Clin Microbiol Antimicrob 15(1):41

Alderwick LJ, Seidel M, Sahm H, Besra GS, Eggeling L (2006) Identification of a novel arabinofuranosyltransferase (AftA) involved in cell wall arabinan biosynthesis in *Mycobacterium tuberculosis*. J Biol Chem 281(23):15653–15661

Almeida-Da-Silva PE, Palomino JC (2011) Molecular basis and mechanisms of drug resistance in *Mycobacterium tuberculosis*: classical and new drugs. J Antimicrob Chemother 66(7):1417–1430

Amir A, Rana K, Arya A, Kapoor N, Kumar H, Siddiqui MA (2014) *Mycobacterium tuberculosis* H37Rv: *in silico* drug targets identification by metabolic pathways analysis. Int J Evol Biol 2014:284170

Anastasio TJ (2017) Editorial: computational and experimental approaches in multi-target pharmacology. Front Pharmacol 8:443

Averbukh I, Ben-Zvi D, Mishra S, Barkai N (2014) Scaling morphogen gradients during tissue growth by a cell division rule. Development 141(10):2150–2156

Ayaz F, Küçükboyacı N, Demirci B (2017) Chemical composition and antimicrobial activity of the essential oil of *Conyza canadensis* (L.) cronquist from Turkey. J Essent Oil Res 29(4):336–343

Baek M, Shin WH, Chung HW, Seok C (2017) GalaxyDock BP2 score: a hybrid scoring function for accurate protein-ligand docking. J Comput Aided Mol Des 31(7):653–666

Baldi A (2010) Computational approaches for drug design and discovery: an overview. Sys Rev Pharm 1(1):95–105

Bashiri G, Rehan AM, Sreebhavan S, Baker HM, Baker EN, Squire CJ (2016) Elongation of the poly-γ-glutamate tail of F420 requires both domains of the F420:γ-glutamyl ligase (FbiB) of *Mycobacterium tuberculosis*. J Biol Chem 291(13):6882–6894

Bates PA, Kelley LA, MacCallum RM, Sternberg MJ (2001) Enhancement of protein modeling by human intervention in applying the automatic programs 3D-JIGSAW and 3D-PSSM. Proteins 5:39–46

Baugh L, Phan I, Begley DW, Clifton MC, Armour B, Dranow DM, Taylor BM, Muruthi MM, Abendroth J, Fairman JW, Fox D 3rd, Dieterich SH, Staker BL, Gardberg AS, Choi R, Hewitt SN, Napuli AJ, Myers J, Barrett LK, Zhang Y, Ferrell M, Mundt E, Thompkins K, Tran N, Lyons-Abbott S, Abramov A, Sekar A, Serbzhinskiy D, Lorimer D, Buchko GW, Stacy R, Stewart LJ, Edwards TE, Van Voorhis WC, Myler PJ (2015) Increasing the structural coverage of tuberculosis drug targets. Tuberculosis (Edinb) 95(2):142–148

Belanger AE, Besra GS, Ford ME, Mikusová K, Belisle JT, Brennan PJ, Inamine JM (1996) The embAB genes of *Mycobacterium avium* encode an arabinosyl transferase involved in cell wall arabinan biosynthesis that is the target for the antimycobacterial drug ethambutol. Proc Natl Acad Sci U S A 93(21):11919–11924

Bell LCK, Noursadeghi M (2018) Pathogenesis of HIV-1 and *Mycobacterium tuberculosis* co-infection. Nat Rev Microbiol 16(2):80–90

Berrada ZL, Lin SY, Rodwell TC, Nguyen D, Schecter GF, Pham L, Janda JM, Elmaraachli W, Catanzaro A, Desmond E (2016) Rifabutin and rifampin resistance levels and associated rpoB mutations in clinical isolates of *Mycobacterium tuberculosis* complex. Diagn Microbiol Infect Dis 85(2):177–181

Bickerton GR, Paolini GV, Besnard J, Muresan S, Hopkins AL (2012) Quantifying the chemical beauty of drugs. Nat Chem 4(2):90–98

Blaszczyk M, Jamroz M, Kmiecik S, Kolinski A (2013) CABS-fold: server for the *de novo* and consensus-based prediction of protein structure. Nucleic Acids Res 41(Web Server issue):W406–W411

Bohm HJ (1992) The computer program LUDI: a new method for the *de novo* design of enzyme inhibitors. J Comput Aided Mol Des 6(1):61–78

Bradley P, Misura KM, Baker D (2005) Toward high-resolution *de novo* structure prediction for small proteins. Science 309(5742):1868–1871

Brooks BR, Brooks CL, Mackerell AD Jr, Nilsson L, Petrella RJ, Roux B, Won Y, Archontis G, Bartels C, Boresch S, Caflisch A, Caves L, Cui Q, Dinner AR, Feig M, Fischer S, Gao J, Hodoscek M, Im W, Kuczera K, Lazaridis T, Ma J, Ovchinnikov V, Paci E, Pastor RW, Post CB, Pu JZ, Schaefer M, Tidor B, Venable RM, Woodcock HL, Wu X, Yang W, York DM, Karplus M (2009) CHARMM: the biomolecular simulation program. J Comput Chem 30(10):1545–1614

Bruning JB, Murillo AC, Chacon O, Barletta RG, Sacchettini JC (2011) Structure of the *Mycobacterium tuberculosis* D-alanine:D-alanine ligase, a target of the antituberculosis drug D-cycloserine. Antimicrob Agents Chemother 55(1):291–301

Brylinski M, Skolnick J (2008) Q-Dock: low-resolution flexible ligand docking with pocket-specific threading restraints. J Biol Chem 29(10):1574–1588

Burkhard P, Taylor P, Walkinshaw MD (1998) An example of a protein ligand found by database mining: description of the docking method and its verification by a 2.3 A X-ray structure of a thrombinligand complex. J Mol Biol 277(2):449–466

Bushra E, Adem J (2016) Mycobacterial metabolic pathways as drug targets: a review. Int J Microbiol Res 7(3):74–87

Case DA, Cerutti DS, Cheatham TE, Darden TA, Duke RE, Giese TJ, Gohlke H, Goetz AW, Greene D, Homeyer N, Izadi S, Kovalenko A, Lee TS, LeGrand S, Li P, Lin C, Liu J, Luchko T, Luo R, Mermelstein D, Merz KM, Monard G, Nguyen H, Omelyan I, Onufriev A, Pan F, Qi R, Roe DR, Roitberg A, Sagui C, Simmerling CL, Botello-Smith WM, Swails J, Walker RC, Wang J, Wolf RM, Wu X, Xiao L, York DM, Kollman PA (2017) AMBER 2017. University of California, San Francisco

Centers for Disease Control and Prevention (CDC). (2018) https://www.cdc.gov/tb/topic/research/default.htm. Accessed 10 Apr 2018

Chambers HF, Turner J, Schecter GF, Kawamura M, Hopewell PC (2005) Imipenem for treatment of tuberculosis in mice and humans. Antimicrob Agents Chemother 49(7):2816–2821

Chandra N (2011) Computational approaches for drug target identification in pathogenic diseases. Expert Opin Drug Discov 6(10):975–979

Chang DT, Oyang YJ, Lin JH (2005) MEDock: a web server for efficient prediction of ligand binding sites based on a novel optimization algorithm. Nucleic Acids Res 33(Web Server issue):W233–W238

Chaudhary KK, Mishra N (2016) A review on molecular docking: novel tool for drug discovery. JSM Chem 4(3):1029

Chen HM, Liu BF, Huang HL, Hwang SF, Ho SY (2007) SODOCK: swarm optimization for highly flexible protein-ligand docking. J Biol Chem 28(2):612–623

Chen J, Zhang S, Cui P, Shi W, Zhang W, Zhang Y (2017) Identification of novel mutations associated with cycloserine resistance in *Mycobacterium tuberculosis*. J Antimicrob Chemother 72(12):3272–3276

Cheng T, Li Q, Zhou Z, Wang Y, Bryant SH (2012) Structure-based virtual screening for drug discovery: a problem-centric review. AAPS J 14(1):133–141

Chinsembu KC (2016) Tuberculosis and nature’s pharmacy of putative anti-tuberculosis agents. Acta Trop 153:46–56

Choi V (2005) YUCCA: an efficient algorithm for small-molecule docking. Chem Biodivers 2(11):1517–1524

Chung JY, Cho SJ, Hah JM (2011) A python-based docking program utilizing a receptor bound ligand shape: PythDock. Arch Pharm Res 34(9):1451–1458

Clark KP (1995) Flexible ligand docking without parameter adjustment across four ligand-receptor complexes. J Comput Chem 16:1210–1226

Clark DE (2003) *In-silico* prediction of blood–brain barrier permeation. Drug Discov Today 8(20):927–933

Cole ST, Brosch R, Parkhill J, Garnier T, Churcher C, Harris D, Gordon SV, Eiglmeier K, Gas S, Barry CE 3rd, Tekaia F, Badcock K, Basham D, Brown D, Chillingworth T, Connor R, Davies R, Devlin K, Feltwell T, Gentles S, Hamlin N, Holroyd S, Hornsby T, Jagels K, Krogh A, McLean J, Moule S, Murphy L, Oliver K, Osborne J, Quail MA, Rajandream MA, Rogers J, Rutter S, Seeger K, Skelton J, Squares R, Squares S, Sulston JE, Taylor K, Whitehead S, Barrell BG (1998) Deciphering the biology of *Mycobacterium tuberculosis* from the complete genome sequence. Nature 393(6685):537–544

Coll F, Phelan J, Hill-Cawthorne GA, Nair MB, Mallard K, Ali S, Abdallah AM, Alghamdi S, Alsomali M, Ahmed AO, Portelli S, Oppong Y, Alves A, Bessa TB, Campino S, Caws M, Chatterjee A, Crampin AC, Dheda K, Furnham N, Glynn JR, Grandjean L, Minh-Ha D, Hasan R, Hasan Z, Hibberd ML, Joloba M, Jones-López EC, Matsumoto T, Miranda A, Moore DJ, Mocillo N, Panaiotov S, Parkhill J, Penha C, Perdigão J, Portugal I, Rchiad Z, Robledo J, Sheen P, Shesha NT, Sirgel FA, Sola C, Oliveira Sousa E, Streicher EM, Helden PV, Viveiros M, Warren RM, McNerney R, Pain A, Clark TG (2018) Genome-wide analysis of multi- and extensively drug-resistant *Mycobacterium tuberculosis*. Nat Genet 50(2):307–316

D'Ambrosio L, Centis R, Tiberi S, Tadolini M, Dalcolmo M, Rendon A, Esposito S, Migliori GB (2017) Delamanid and bedaquiline to treat multidrug-resistant and extensively drug-resistant tuberculosis in children: a systematic review. J Thorac Dis 9(7):2093–2101

Dar AM, Mir S (2017) Molecular docking: approaches, types, applications and basic challenges. J Anal Bioanal Tech 8:356

de-Mendonça JD, Ely F, Palma MS, Frazzon J, Basso LA, Santos DS (2007) Functional characterization by genetic complementation of aroB-encoded dehydroquinate synthase from *Mycobacterium tuberculosis* H37Rv and its heterologous expression and purification. J Bacteriol 189(17):6246–6252

de-Ruyck J, Brysbaert G, Blossey R, Lensink MF (2016) Molecular docking as a popular tool in drug design, an *in silico* travel. Adv Appl Bioinform Chem 9:1–11

De-Vivo M, Masetti M, Bottegoni G, Cavalli A (2016) Role of molecular dynamics and related methods in drug discovery. J Med Chem 59(9):4035–4061

Dheda K, Chang KC, Guglielmetti L, Furin J, Schaaf HS, Chesov D, Esmail A, Lange C (2017) Clinical management of adults and children with multidrug-resistant and extensively drug-resistant tuberculosis. Clin Microbiol Infect 23(3):131–140

Dominguez C, Boelens R, Bonvin AM (2003) HADDOCK: a protein-protein docking approach based on biochemical or biophysical information. J Am Chem Soc 125(7):1731–1737

Dookie N, Rambaran S, Padayatchi N, Mahomed S, Naidoo K (2018) Evolution of drug resistance in *Mycobacterium tuberculosis*: a review on the molecular determinants of resistance and implications for personalized care. J Antimicrob Chemother. https://doi.org/10.1093/jac/dkx506

Dundas J, Ouyang Z, Tseng J, Binkowski A, Turpaz Y, Liang J (2006) CASTp: computed atlas of surface topography of proteins with structural and topographical mapping of functionally annotated residues. Nucleic Acids Res 34(Web Server issue):W116–W118

Durrant JD, McCammon JA (2011) Molecular dynamics simulations and drug discovery. BMC Biol 9:71

Engin HB, Gursoy A, Nussinov R, Keskin O (2014) Network-based strategies can help mono- and poly-pharmacology drug discovery: a systems biology view. Curr Pharm Des 20(8):1201–1207

Errey JC, Blanchard JS (2005) Functional characterization of a novel ArgA from *Mycobacterium tuberculosis*. J Bacteriol 187(9):3039–3044

European Center for Disease Prevention and Control (ECDC). (2018) https://ecdc.europa.eu/en/publications-data/tuberculosis-surveillance-and-monitoring-europe-2018. Accessed 10 Apr 2018

Ewing TJA, Makino S, Skillman AG, Kuntz ID (2001) DOCK 4.0: search strategies for automated molecular docking of flexible molecule databases. J Comput Aided Mol Des 15(5):411–428

Fakhar Z, Naiker S, Alves CN, Govender T, Maguire GE, Lameira J, Lamichhane G, Kruger HG, Honarparvar B (2016) A comparative modeling and molecular docking study on

Mycobacterium tuberculosis targets involved in peptidoglycan biosynthesis. J Biomol Struct Dyn 34(11):2399–2417

Fan H, Schneidman-Duhovny D, Irwin JJ, Dong G, Shoichet BK, Sali A (2011) Statistical potential for modeling and ranking of protein–ligand interactions. J Chem Inf Model 51(12):3078–3092

Ferraris DM, Spallek R, Oehlmann W, Singh M, Rizzi M (2015) Structures of citrate synthase and malate dehydrogenase of *Mycobacterium tuberculosis*. Proteins 83(2):389–394

Ferreira LG, Dos Santos RN, Oliva G, Andricopulo AD (2015) Molecular docking and structure-based drug design strategies. Molecules 20(7):13384–13421

Field SK (2015) Bedaquiline for the treatment of multidrug-resistant tuberculosis: great promise or disappointment? Ther Adv Chronic Dis 6(4):170–184

Fleischmann RD, Alland D, Eisen JA, Carpenter L, White O, Peterson J, DeBoy R, Dodson R, Gwinn M, Haft D, Hickey E, Kolonay JF, Nelson WC, Umayam LA, Ermolaeva M, Salzberg SL, Delcher A, Utterback T, Weidman J, Khouri H, Gill J, Mikula A, Bishai W, Jacobs WR Jr, Venter JC, Fraser CM (2002) Whole-genome comparison of *Mycobacterium tuberculosis* clinical and laboratory strains. J Bacteriol 184(19):5479–5490

Forrellad MA, Klepp LI, Gioffré A, Sabio-y-García J, Morbidoni HR, de la Paz Santangelo M, Cataldi AA, Bigi F (2013) Virulence factors of the *Mycobacterium tuberculosis* complex. Virulence 4(1):3–66

Gabb HA, Jackson RM, Sternberg MJ (1997) Modelling protein docking using shape complementarity, electrostatics and biochemical information. J Mol Biol 272(1):106–120

Gao Z, Li H, Zhang H, Liu X, Kang L, Luo X, Zhu W, Chen K, Wang X, Jiang H (2008) PDTD: a web-accessible protein database for drug target identification. BMC Bioinformatics 9:104

Gaudreault F, Najmanovich RJ (2015) FlexAID: revisiting docking on non-native-complex structures. J Chem Inf Model 55(7):1323–1336

Geromichalos GD (2012) Virtual screening strategies and application in drug designing. Drug Des 2:e109

Ghose AK, Herbertz T, Salvino JM, Mallamo JP (2006) Knowledge-based chemoinformatic approaches to drug discovery. Drug Discov Today 11(23–24):1107–1114

Gonzalo X, Drobniewski F (2013) Is there a place for β-lactams in the treatment of multidrug-resistant/extensively drug-resistant tuberculosis? Synergy between meropenem and amoxicillin/clavulanate. J Antimicrob Chemother 68(2):366–369

Graham DE, Xu H, White RH (2002) Identification of coenzyme M biosynthetic phosphosulfolactate synthase: a new family of sulfonate-biosynthesizing enzymes. J Biol Chem 277(16):13421–13429

Grochowski LL, Xu H, White RH (2008) Identification and characterization of the 2-phospho-L-lactate guanylyltransferase involved in coenzyme F420 biosynthesis. Biochemistry 47(9):3033–3037

Grosdidier A, Zoete V, Michielin O (2007) EADock: docking of small molecules into protein active sites with a multi objective evolutionary optimization. Proteins 67(4):1010–1025

Grosdidier A, Zoete V, Michielin O (2011) SwissDock, a protein-small molecule docking web service based on EADock DSS. Nucleic Acids Res 39.(Web Server issue:W270–W277

Gupta A, Gandhimathi A, Sharma P, Jayaram B (2007) ParDOCK: an all atom energy based Monte Carlo docking protocol for protein-ligand complexes. Protein Pept Lett 14(7):632–646

Halgren TA, Murphy RB, Friesner RA, Beard HS, Frye LL, Pollard WT, Banks JL (2004) Glide: a new approach for rapid, accurate docking and scoring. 2. Enrichment factors in database screening. J Med Chem 47(7):1750–1759

Hallam SJ, Putnam N, Preston CM, Detter JC, Rokhsar D, Richardson PM, DeLong EF (2004) Reverse methanogenesis: testing the hypothesis with environmental genomics. Science 305(5689):1457–1462

Hart TN, Read RJ (1992) A multiple-start Monte Carlo docking method. Proteins 13(3):206–222

Hazai E, Kovács S, Demkó L, Bikádi Z (2009) DockingServer: molecular docking calculations online. Acta Pharm Hung 79(1):17–21

Huang B (2009) MetaPocket: a meta approach to improve protein ligand binding site prediction. OMICS 13(4):325–330

Huang SY, Li M, Wang J, Pan Y (2016) HybridDock: A hybrid protein-ligand docking protocol integrating protein- and ligand-based approaches. J Chem Inf Model 56(6):1078–1087

Hung CL, Chen CC (2014) Computational approaches for drug discovery. Drug Dev Res 75(6):412–418

Irwin JJ, Shoichet BK, Mysinger MM, Huang N, Colizzi F, Wassam P, Cao Y (2009) Automated docking screens: a feasibility study. J Med Chem 52(18):5712–5720

Jabeen K, Shakoor S, Hasan R (2015) Fluoroquinolone-resistant tuberculosis: implications in settings with weak healthcare systems. Int J Infect Dis 32:118–123

Janardhan S, John L, Prasanthi M, Poroikov V, Narahari-Sastry G (2017) A QSAR and molecular modelling study towards new lead finding: polypharmacological approach to *Mycobacterium tuberculosis*. SAR QSAR Environ Res 28(10):815–832

Jiang F, Kim SH (1991) “Soft docking”: matching of molecular surface cubes. J Mol Biol 219(1):79–102

Jones DT (1999) GenTHREADER: an efficient and reliable protein folds recognition method for genomic sequences. J Mol Biol 287(4):797–815

Jones G, Willett P, Glen RC, Leach AR, Taylor R (1997) Development and validation of a genetic algorithm for flexible docking. J Mol Biol 267(3):727–748

Kanehisa M, Sato Y, Kawashima M, Furumichi M, Tanabe M (2016) KEGG as a reference resource for gene and protein annotation. Nucleic Acids Res 44(D1):D457–D462

Kar S, Roy K (2013) How far can virtual screening take us in drug discovery? Expert Opin Drug Discov 8(3):245–261

Katsila T, Spyroulias GA, Patrinos GP, Matsoukas MT (2016) Computational approaches in target identification and drug discovery. Comput Struct Biotechnol J 14:177–184

Kaur G, Pandey B, Grover A, Garewal N, Grover A, Kaur J (2018) Drug targeted virtual screening and molecular dynamics of LipU protein of *Mycobacterium tuberculosis* and *Mycobacterium leprae*. J Biomol Struct Dyn. https://doi.org/10.1080/07391102.2018.1454852

Kelley BP, Brown SP, Warren GL, Muchmore SW (2015a) POSIT: flexible shape-guided docking for pose prediction. J Chem Inf Model 55(8):1771–1780

Kelley LA, Mezulis S, Yates CM, Wass MN, Sternberg MJ (2015b) The Phyre2 web portal for protein modeling, prediction and analysis. Nat Protoc 10(6):845–858

Kim DS, Kim CM, Won CI, Kim JK, Ryu J, Cho Y, Bhak J (2011) BetaDock: shape-priority docking method based on beta-complex. J Biomol Struct Dyn 29(1):219–242

Kneidinger B, Marolda C, Graninger M, Zamyatina A, McArthur F, Kosma P, Valvano MA, Messner P (2002) Biosynthesis pathway of ADP-L-glycero-beta-D-manno-heptose in *Escherichia coli*. J Bacteriol 184(2):363–369

Ko Y, Choi I (2016) Putative 3D structure of QcrB from *Mycobacterium tuberculosis* cytochrome bc1 complex, a novel drug-target for new series of antituberculosis agent Q203. Bull Kor Chem Soc 37:725–731

Korb O, Stützle T, Exner TE (2006) PLANTS: application of ant colony optimization to structure-based drug design. In: Dorigo M, Gambardella LM, Birattari M, Martinoli A, Poli R, Stützle T (eds) Ant colony optimization and swarm intelligence, vol 4150. Springer, Berlin, Heidelberg, pp 247–258

Krüüner A, Jureen P, Levina K, Ghebremichael S, Hoffner S (2003) Discordant resistance to kanamycin and amikacin in drug-resistant *Mycobacterium tuberculosis*. Antimicrob Agents Chemother 47(9):2971–2973

Kumar P, Arora K, Lloyd JR, Lee IY, Nair V, Fischer E, Boshoff HI, Barry CE 3rd (2012) Meropenem inhibits D,D-carboxypeptidase activity in *Mycobacterium tuberculosis*. Mol Microbiol 86(2):367–381

Laskowski RA (1995) SURFNET: a program for visualizing molecular surfaces, cavities, and intermolecular interactions. J Mol Graph 13(5):323–330, 307–308

Laurie AT, Jackson RM (2005) Q-SiteFinder: an energy-based method for the prediction of protein-ligand binding sites. Bioinformatics 21(9):1908–1916

Lee GR, Seok C (2016) Galaxy7TM: flexible GPCR-ligand docking by structure refinement. Nucleic Acids Res 44(W1):W502–W506

Lee HS, Zhang Y (2012) BSP-SLIM: a blind low-resolution ligand-protein docking approach using predicted protein structures. Proteins 80(1):93–110

Leelananda SP, Lindert S (2016) Computational methods in drug discovery. Beilstein J Org Chem 12:2694–2718

Leeson PD, Davis AM, Steele J (2004) Drug-like properties: guiding principles for design–or chemical prejudice? Drug Discov Today Technol 1(3):189–195

LeMagueres P, Im H, Ebalunode J, Strych U, Benedik MJ, Briggs JM, Kohn H, Krause KL (2005) The 1.9 A crystal structure of alanine racemase from *Mycobacterium tuberculosis* contains a conserved entryway into the active site. Biochemistry 44(5):1471–1481

Li YH, Yu CY, Li XX, Zhang P, Tang J, Yang Q, Fu T, Zhang X, Cui X, Tu G, Zhang Y, Li S, Yang F, Sun Q, Qin C, Zeng X, Chen Z, Chen YZ, Zhu F (2018) Therapeutic target database update 2018: enriched resource for facilitating bench-to-clinic research of targeted therapeutics. Nucleic Acids Res 46(D1):D1121–D1127

Lindorff-Larsen K, Piana S, Palmo K, Maragakis P, Klepeis JL, Dror RO, Shaw DE (2010) Improved side-chain torsion potentials for the amber ff99SB protein force field. Proteins 78(8):1950–1958

Lionta E, Spyrou G, Vassilatis DK, Cournia Z (2014) Structure-based virtual screening for drug discovery: principles, applications and recent advances. Curr Top Med Chem 14(16):1923–1938

Lipinski CA (2004) Lead- and drug-like compounds: the rule-of-five revolution. Drug Discov Today Technol 1(4):337–341

Liu M, Wang S (1999) MCDOCK: a Monte Carlo simulation approach to the molecular docking problem. J Comput Aided Mol Des 13(5):435–451

Liu T, Tang GW, Capriotti E (2011) Comparative modeling: the state of the art and protein drug target structure prediction. Comb Chem High Throughput Screen 14(6):532–547

Liu X, Shi D, Zhou S, Liu H, Liu H, Yao X (2018) Molecular dynamics simulations and novel drug discovery. Expert Opin Drug Discov 13(1):23–37

London N, Raveh B, Cohen E, Fathi G, Schueler-Furman O (2011) Rosetta FlexPepDock web server--high resolution modeling of peptide-protein interactions. Nucleic Acids Res 39(Web Server issue):W249–W253

Lone MY, Athar M, Gupta VK, Jha PC (2017a) Identification of *Mycobacterium tuberculosis* enoyl-acyl carrier protein reductase inhibitors: a combined *in-silico* and *in-vitro* analysis. J Mol Graph Model 76:172–180

Lone MY, Athar M, Gupta VK, Jha PC (2017b) Prioritization of natural compounds against *Mycobacterium* tuberculosis 3-dehydroquinate dehydratase: A combined *in-silico* and *in-vitro* study. Biochem Biophys Res Commun 491(4):1105–1111

Lone MY, Manhas A, Athar M, Jha PC (2017c) Identification of InhA inhibitors: a combination of virtual screening, molecular dynamics simulations and quantum chemical studies. J Biomol Struct Dyn. https://doi.org/10.1080/07391102.2017.1372313

Maganti L, OSDD Consortium, Ghoshal N (2015) 3D-QSAR studies and shape based virtual screening for identification of novel hits to inhibit MbtA in *Mycobacterium tuberculosis*. J Biomol Struct Dyn 33(2):344–364

Maitre T, Aubry A, Jarlier V, Robert J, Veziris N, CNR-MyRMA (2017) Multidrug and extensively drug-resistant tuberculosis. Med Mal Infect 47(1):3–10

Manikandan K, Geerlof A, Zozulya AV, Svergun DI, Weiss MS (2011) Structural studies on the enzyme complex isopropylmalate isomerase (LeuCD) from *Mycobacterium tuberculosis*. Proteins 79(1):35–49

Mansuri R, Ansari MY, Singh J, Rana S, Sinha S, Sahoo GC, Dikhit MR, Das P (2016) Computational elucidation of structural basis for ligand binding with *Mycobacterium tuberculosis* glucose-1-phosphate thymidylyltransferase (RmlA). Curr Pharm Biotechnol 17(12):1089–1099

Mao C, Shukla M, Larrouy-Maumus G, Dix FL, Kelley LA, Sternberg MJ, Sobral BW, de-Carvalho LP (2013) Functional assignment of *Mycobacterium tuberculosis* proteome revealed by genome-scale fold-recognition. Tuberculosis (Edinb) 93(1):40–46

Marialke J, Tietze S, Apostolakis J (2008) Similarity based docking. J Chem Inf Model 48(1):186–196

Marks DS, Colwell LJ, Sheridan R, Hopf TA, Pagnani A, Zecchina R, Sander C (2011) Protein 3D structure computed from evolutionary sequence variation. PLoS One 6(12):e28766

Matteelli A, Roggi A, Carvalho AC (2014) Extensively drug-resistant tuberculosis: epidemiology and management. Clin Epidemiol 6:111–118

Maus CE, Plikaytis BB, Shinnick TM (2005a) Mutation of tlyA confers capreomycin resistance in *Mycobacterium tuberculosis*. Antimicrob Agents Chemother 49(2):571–577

Maus CE, Plikaytis BB, Shinnick TM (2005b) Molecular analysis of cross-resistance to capreomycin, kanamycin, amikacin, and viomycin in *Mycobacterium tuberculosis*. Antimicrob Agents Chemother 49(8):3192–3197

McGann MR, Almond HR, Nicholls A, Grant JA, Brown FK (2003) Gaussian docking functions. Biopolymers 68(1):76–90

McMartin C, Bohacek RS (1997) QXP: powerful, rapid computer algorithms for structure-based drug design. J Comput Aided Mol Des 11(4):333–344

Mehra R, Rani C, Mahajan P, Vishwakarma RA, Khan IA, Nargotra A (2016) Computationally guided identification of novel *Mycobacterium tuberculosis* GlmU lnhibitory leads, their optimization, and in vitro validation. ACS Comb Sci 18(2):100–116

Meng XY, Zhang HX, Mezei M, Cui M (2011) Molecular docking: a powerful approach for structure-based drug discovery. Curr Comput Aided Drug Des 7(2):146–157

Miller MD, Kearsley SK, Underwood DJ, Sheridan RP (1994) FLOG: a system to select 'quasi-flexible' ligands complementary to a receptor of known three-dimensional structure. J Comput Aided Mol Des 8(2):153–174

Mizutani MY, Tomioka N, Itai A (1994) Rational automatic search method for stable docking models of protein and ligand. J Mol Biol 243(2):310–326

Mohamad S, Ismail NN, Parumasivam T, Ibrahim P, Osman H, A Wahab H (2018) Antituberculosis activity, phytochemical identification of *Costus speciosus* (J. Koenig) Sm., *Cymbopogon citratus* (DC. Ex Nees) Stapf., and *Tabernaemontana coronaria* (L.) Willd. and their effects on the growth kinetics and cellular integrity of *Mycobacterium tuberculosis* H37Rv. BMC Complement Altern Med 18(1):5

Morris GM, Goodsell DS, Halliday RS, Huey R, Hart WE, Belew RK, Olson AJ (1998) Automated docking using a Lamarckian genetic algorithm and an empirical binding free energy function. J Comput Chem 19(14):1639–1662

Mukhopadhyay S, Nair S, Ghosh S (2012) Pathogenesis in tuberculosis: transcriptomic approaches to unraveling virulence mechanisms and finding new drug targets. FEMS Microbiol Rev 36(2):463–485

Namasivayam V, Gunther R (2007) PSO@AUTODOCK: a fast flexible molecular docking program based on swarm intelligence. Chem Biol Drug Des 70(6):475–484

Naz S, Farooq U, Ali S, Sarwar R, Khan S, Abagyan R (2018) Identification of new benzamide inhibitor against α-subunit of tryptophan synthase from *Mycobacterium tuberculosis* through structure-based virtual screening, anti-tuberculosis activity and molecular dynamics simulations. J Biomol Struct Dyn. https://doi.org/10.1080/07391102.2018.1448303

Nazzaro F, Fratianni F, De Martino L, Coppola R, De-Feo V (2013) Effect of essential oils on pathogenic bacteria. Pharmaceuticals (Basel) 6(12):1451–1474

Njire M, Tan Y, Mugweru J, Wang C, Guo J, Yew W, Tan S, Zhang T (2016) Pyrazinamide resistance in *Mycobacterium tuberculosis*: review and update. Adv Med Sci 61(1):63–71

Oprea TI (2000) Property distribution of drug-related chemical databases. J Comput Aided Mol Des 14(3):251–264

Pagadala NS, Syed K, Tuszynski J (2017) Software for molecular docking: a review. Biophys Rev 9(2):91–102

Pandey B, Grover S, Tyagi C, Goyal S, Jamal S, Singh A, Kaur J, Grover A (2018) Dynamics of fluoroquinolones induced resistance in DNA gyrase of *Mycobacterium tuberculosis*. J Biomol Struct Dyn 36(2):362–375

Pang YP, Perola E, Xu K, Prendergast FG (2001) EUDOC: a computer program for identification of drug interaction sites in macromolecules and drug leads from chemical databases. J Comput Chem 22(15):1750–1771

Paul DS, Gautham N (2016) MOLS 2.0: software package for peptide modeling and protein-ligand docking. J Mol Model 22(10):239

Pei JF, Wang Q, Liu ZM, Li QL, Yang K, Lai LH (2006) PSIDOCK: towards highly efficient and accurate flexible ligand docking. Proteins 62(4):934–946

Peng J, Xu J (2011) RaptorX: exploiting structure information for protein alignment by statistical inference. Proteins 10:161–171

Phillips JC, Braun R, Wang W, Gumbart J, Tajkhorshid E, Villa E, Chipot C, Skeel RD, Kalé L, Schulten K (2005) Scalable molecular dynamics with NAMD. J Comput Chem 26(16):1781–1802

Pippel M, Scharfe M, Meier R, Sippl W (2012) ParaDockS – an open source framework for molecular docking. J Cheminform. https://doi.org/10.1186/1758-2946-4-S1-F3

Plewczynski D, Łaźniewski M, von Grotthuss M, Rychlewski L, Ginalski K (2011) VoteDock: consensus docking method for prediction of protein-ligand interactions. J Comput Chem 32(4):568–581

Putri DU, Rintiswati N, Soesatyo MH, Haryana SM (2018) Immune modulation properties of herbal plant leaves: *Phyllanthus niruri* aqueous extract on immune cells of tuberculosis patient – *in vitro* study. Nat Prod Res 32(4):463–467

Pyrkov TV, Chugunov AO, Krylov NA, Nolde DE, Efremov RG (2009) PLATINUM: a web tool for analysis of hydrophobic/hydrophilic organization of biomolecular complexes. Bioinformatics 25(9):1201–1202

Qiu J, Zang S, Ma Y, Owusu L, Zhou L, Jiang T, Xin Y (2017) Homology modeling and identification of amino acids involved in the catalytic process of *Mycobacterium tuberculosis* serine acetyltransferase. Mol Med Rep 15(3):1343–1347

Quan D, Nagalingam G, Payne R, Triccas JA (2017) New tuberculosis drug leads from naturally occurring compounds. Int J Infect Dis 56:212–220

Rajendran V, Sethumadhavan R (2014) Drug resistance mechanism of PncA in *Mycobacterium tuberculosis*. J Biomol Struct Dyn 32(2):209–221

Raman K, Chandra N (2008) *Mycobacterium tuberculosis* interactome analysis unravels potential pathways to drug resistance. BMC Microbiol 8:234

Raman K, Yeturu K, Chandra N (2008) targetTB: a target identification pipeline for *Mycobacterium* tuberculosis through an interactome, reactome and genome-scale structural analysis. BMC Syst Biol 2:109

Ramesh KV, Purohit M, Mekhala K, Krishnan M, Wagle K, Deshmukh S (2008) Modeling the interactions of herbal drugs to β-ketoacyl ACP synthase of *Mycobacterium tuberculosis* H37Rv. J Biomol Struct Dyn 25(5):481–493

Reddy AS, Pati SP, Kumar PP, Pradeep HN, Sastry GN (2007) Virtual screening in drug discovery -- a computational perspective. Curr Protein Pept Sci 8(4):329–351

Ruiz-Carmona S, Alvarez-Garcia D, Foloppe N, Garmendia-Doval AB, Juhos S, Schmidtke P et al (2014) rDock: a fast, versatile and open source program for docking ligands to proteins and nucleic acids. PLoS Comput Biol 2014(10):e1003571

Saini DK, Tyagi JS (2005) High-throughput microplate phosphorylation assays based on DevR-DevS/Rv2027c 2-component signal transduction pathway to screen for novel antitubercular compounds. J Biomol Screen 10(3):215–224

Sambandamurthy VK, Wang X, Chen B, Russell RG, Derrick S, Collins FM, Morris SL, Jacobs WR Jr (2002) A pantothenate auxotroph of *Mycobacterium tuberculosis* is highly attenuated and protects mice against tuberculosis. Nat Med 8(10):1171–1174

Sanusi SB, Abu-Bakar MF, Mohamed M, Sabran SF, Mainasara MM (2017) Southeast Asian medicinal plants as a potential source of antituberculosis agent. Evid Based Complement Alternat Med 2017:7185649

Sauton N, Lagorce D, Villoutreix BO, Miteva MA (2008) MSDOCK: accurate multiple conformation generator and rigid docking protocol for multi-step virtual ligand screening. BMC Bioinformatics 2008:9

Schmidtke P, Bidon-Chanal A, Luque FJ, Barril X (2011) MDpocket: open-source cavity detection and characterization on molecular dynamics trajectories. Bioinformatics 27(23):3276–3285

Schneidman-Duhovny D, Inbar Y, Nussinov R, Wolfson HJ (2005) PatchDock and SymmDock: servers for rigid and symmetric docking. Nucleic Acids Res 33(Web Server issue):W363–W367

Schwede T, Kopp J, Guex N, Peitsch MC (2003) SWISS-MODEL: an automated protein homology-modeling server. Nucleic Acids Res 31(13):3381–3385

Seidel M, Alderwick LJ, Birch HL, Sahm H, Eggeling L, Besra GS (2007) Identification of a novel arabinofuranosyltransferase AftB involved in a terminal step of cell wall arabinan biosynthesis in *Corynebacterianeae*, such as *Corynebacterium glutamicum* and *Mycobacterium tuberculosis*. J Biol Chem 282(20):14729–14740

Seifert M, Catanzaro D, Catanzaro A, Rodwell TC (2015) Genetic mutations associated with isoniazid resistance in *Mycobacterium tuberculosis*: a systematic review. PLoS One 10(3):e0119628

Sengupta S, Roy D, Bandyopadhyay S (2015) Structural insight into *Mycobacterium tuberculosis* maltosyl transferase inhibitors: pharmacophore-based virtual screening, docking, and molecular dynamics simulations. J Biomol Struct Dyn 33(12):2655–2666

Shin WH, Heo L, Lee J, Ko J, Seok C, Lee J (2011) LigDock-CSA: protein-ligand docking using conformational space annealing. J Comput Chem 32(15):3226–3232

Shoichet BK (2004) Virtual screening of chemical libraries. Nature 432(7019):862–865

Shukla R, Shukla H, Sonkar A, Pandey T, Tripathi T (2017) Structure-based screening and molecular dynamics simulations offer novel natural compounds as potential inhibitors of *Mycobacterium tuberculosis* isocitrate lyase. J Biomol Struct Dyn. https://doi.org/10.1080/07391102.2017.1341337

Silva JRA, Bishai WR, Govender T, Lamichhane G, Maguire GEM, Kruger HG, Lameira J, Alves CN (2016) Targeting the cell wall of *Mycobacterium tuberculosis*: a molecular modeling investigation of the interaction of imipenem and meropenem with L,D-transpeptidase 2. J Biomol Struct Dyn 34(2):304–317

Singh RK, Kefala G, Janowski R, Mueller-Dieckmann C, von Kries JP, Weiss MS (2005) The high-resolution structure of LeuB (Rv2995c) from *Mycobacterium tuberculosis*. J Mol Biol 346(1):1–11

Singh T, Biswas D, Jayaram B (2011) AADS – an automated active site identification, docking, and scoring protocol for protein targets based on physicochemical descriptors. J Chem Inf Model 51(10):2515–2527

Skariyachan S, Manjunath M, Bachappanavar N (2018) Screening of potential lead molecules against prioritised targets of multi-drug-resistant-*Acinetobacter baumannii* – insights from molecular docking, molecular dynamic simulations and *in vitro* assays. J Biomol Struct Dyn. https://doi.org/10.1080/07391102.2018.1451387

Sneader W (1990) Chronology of drug introductions. Comp Med Chem 1:7–80

Sobolev V, Wade RC, Vriend G, Edelman M (1996) Molecular docking using surface complementarity. Proteins 25(1):120–129

Söding J, Biegert A, Lupas AN (2005) The HHpred interactive server for protein homology detection and structure prediction. Nucleic Acids Res 33(Web Server issue):W244–W248

Sopitthummakhun K, Thongpanchang C, Vilaivan T, Yuthavong Y, Chaiyen P, Leartsakulpanich U (2012) Plasmodium serine hydroxymethyltransferase as a potential anti-malarial target: inhibition studies using improved methods for enzyme production and assay. Malar J 11:194

Spitzer R, Jain AN (2012) Surflex-Dock: docking benchmarks and real-world application. J Comput Aided Mol Des 26(6):687–699

Stroganov OV, Novikov FN, Stroylov VS, Kulkov V, Chilov GG (2008) Lead finder: an approach to improve accuracy of protein-ligand docking, binding energy estimation, and virtual screening. J Chem Inf Model 48(12):2371–2385

Sun H, Zhang C, Xiang L, Pi R, Guo Z, Zheng C, Li S, Zhao Y, Tang K, Luo M, Rastogi N, Li Y, Sun Q (2016) Characterization of mutations in streptomycin-resistant *Mycobacterium tuberculosis* isolates in Sichuan, China and the association between Beijing-lineage and dual-mutation in gidB. Tuberculosis (Edinb) 96:102–106

Szklarczyk D, Morris JH, Cook H, Kuhn M, Wyder S, Simonovic M, Santos A, Doncheva NT, Roth A, Bork P, Jensen LJ, von-Mering C (2017) The STRING database in 2017: quality-controlled protein-protein association networks, made broadly accessible. Nucleic Acids Res 45(D1):D362–D368

Tan KP, Nguyen TB, Patel S, Varadarajan R, Madhusudhan MS (2013) Depth: a web server to compute depth, cavity sizes, detect potential small-molecule ligand-binding cavities and predict the pKa of ionizable residues in proteins. Nucleic Acids Res 41(Web Server issue):W314–W321

Tan Y, Su B, Zheng H, Song Y, Wang Y, Pang Y (2017) Molecular characterization of prothionamide-resistant *Mycobacterium tuberculosis* isolates in southern China. Front Microbiol 8:2358

Taylor JS, Burnett RM (2000) DARWIN: a program for docking flexible molecules. Proteins 41(2):173–191

Taylor RD, Jewsbury PJ, Essex JW (2003) FDS: flexible ligand and receptor docking with a continuum solvent model and soft-core energy function. J Comput Chem 24(13):1637–1656

Thomsen R, Christensen MH (2006) MolDock: a new technique for high-accuracy molecular docking. J Med Chem 49(11):3315–3321

Totrov M, Abagyan R (1997) Flexible protein-ligand docking by global energy optimization in internal coordinates. Proteins 1:215–220

Trosset JY, Scheraga HA (1999) Prodock: software package for protein modeling and docking. J Comput Chem 20(4):412–427

Trott O, Olson AJ (2010) AutoDock Vina: improving the speed and accuracy of docking with a new scoring function, efficient optimization, and multithreading. J Comput Chem 31(2):455–461

Tsai TY, Chang KW, Chen CY (2011) iScreen: world's first cloud-computing web server for virtual screening and de novo drug design based on TCM database@Taiwan. J Comput Aided Mol Des 25(6):525–531

Usha T, Shanmugarajan D, Goyal AK, Kumar CS, Middha SK (2017) Recent updates on computer-aided drug discovery: time for a paradigm shift. Curr Top Med Chem 17(30):3296–3307

Valvano MA, Messner P, Kosma P (2002) Novel pathways for biosynthesis of nucleotide-activated glycero-manno-heptose precursors of bacterial glycoproteins and cell surface polysaccharides. Microbiology 148(Pt 7):1979–1989

Veber DF, Johnson SR, Cheng HY, Smith BR, Ward KW, Kopple KD (2002) Molecular properties that influence the oral bioavailability of drug candidates. J Med Chem 45(12):2615–2623

Venkatachalam CM, Jiang X, Oldfield T, Waldman M (2003) LigandFit: a novel method for the shape directed rapid docking of ligands to protein active-sites. J Mol Graph Model 21(4):289–307

Vidyaraj CK, Chitra A, Smita S, Muthuraj M, Govindarajan S, Usharani B, Anbazhagi S (2017) Prevalence of rifampicin-resistant *Mycobacterium tuberculosis* among human-immunodeficiency-virus-seropositive patients and their treatment outcomes. J Epidemiol Glob Health 7(4):289–294

Vilar S, Cozza G, Moro S (2008) Medicinal chemistry and the molecular operating environment (MOE): application of QSAR and molecular docking to drug discovery. Curr Top Med Chem 8(18):1555–1572

Vilchèze C, Jacobs WR Jr (2014) Resistance to isoniazid and ethionamide in *Mycobacterium tuberculosis*: genes, mutations, and causalities. Microbiol Spectr 2(4):MGM2-0014-2013

Vyas V, Jain A, Jain A, Gupta A (2008) Virtual screening: a fast tool for drug design. Sci Pharm 76(3):333–360

Vyas VK, Ukawala RD, Ghate M, Chintha C (2012) Homology modeling a fast tool for drug discovery: current perspectives. Indian J Pharm Sci 74(1):1–17

Wagener M, Jd V, Nabuurs SB (2012) Flexible protein-ligand docking using the Fleksy protocol. J Comput Chem 33(12):1215–1217

Wang JC, Chu PY, Chen CM, Lin JH (2012) idTarget: a web server for identifying protein targets of small chemical molecules with robust scoring functions and a divide-and-conquer docking approach. Nucleic Acids Res 40(Web Server issue):W393–W399

Webb B, Sali A (2017) Protein structure modeling with MODELLER. Methods Mol Biol 1654:39–54

Welch W, Ruppert J, Jain AN (1996) Hammerhead: fast, fully automated docking of flexible ligands to protein binding sites. Chem Biol 3(6):449–462

Wink M (2015) Modes of action of herbal medicines and plant secondary metabolites. Medicines (Basel) 2(3):251–286

World Health Organization (WHO). (2018) http://www.who.int/tb/en/. Accessed 10 Apr 2018

Wu S, Zhang Y (2008) MUSTER: improving protein sequence profile-profile alignments by using multiple sources of structure information. Proteins 72(2):547–556

Xu D, Zhang Y (2012) Ab initio protein structure assembly using continuous structure fragments and optimized knowledge-based force field. Proteins 80(7):1715–1735

Yan RX, Si JN, Wang C, Zhang Z (2009) DescFold: a web server for protein fold recognition. BMC Bioinformatics 10:416

Yang JM, Chen CC (2004) GEMDOCK: a generic evolutionary method for molecular docking. Proteins 55(2):288–304

Yang J, Yan R, Roy A, Xu D, Poisson J, Zhang Y (2015) The I-TASSER suite: protein structure and function prediction. Nat Methods 12(1):7–8

Zhao Y, Sanner MF (2007) FLIPDock: docking flexible ligands into flexible receptors. Proteins 68(3):726–737

Zhao LL, Sun Q, Liu HC, Wu XC, Xiao TY, Zhao XQ, Li GL, Jiang Y, Zeng CY, Wan KL (2015) Analysis of embCAB mutations associated with ethambutol resistance in multidrug-resistant *Mycobacterium tuberculosis* isolates from China. Antimicrob Agents Chemother 59(4):2045–2050

Zheng J, Rubin EJ, Bifani P, Mathys V, Lim V, Au M, Jang J, Nam J, Dick T, Walker JR, Pethe K, Camacho LR (2013) Para-Aminosalicylic acid is a prodrug targeting dihydrofolate reductase in *Mycobacterium tuberculosis*. J Biol Chem 288(32):23447–23456

Zsoldos Z, Reid D, Simon A, Sadjad SB, Johnson AP (2007) eHiTS: a new fast, exhaustive flexible ligand docking system. J Mol Graph Model 26(1):198–212

Chapter 13
Understanding the Regulatory Features of Co-regulated Genes Using *Di*stant *R*egulatory *E*lements (DiRE) Genomic Tool in Health and Disease

Arif Mohammed, Othman A. Alghamdi, Mohd Rehan, Babajan Banaganapalli, Ramu Elango, and Noor Ahmad Shaik

Contents

Abbreviations

DEG	Differentially Expressed Gene
DiRE	Distant Regulatory Elements
ECR	Evolutionary Conserved Region
EI	Enhancer Identification

A. Mohammed (✉) · O. A. Alghamdi
Biology Department, Faculty of Sciences, University of Jeddah, Jeddah, Saudi Arabia
e-mail: amohammed1@uj.edu.sa; oalghamdi@uj.edu.sa

M. Rehan
King Fahd Medical Research Center, King Abdulaziz University, Jeddah, Saudi Arabia

Department of Medical Laboratory Technology, Faculty of Applied Medical Sciences, King Abdulaziz University, Jeddah, Saudi Arabia
e-mail: mrtahir@kau.edu.sa

B. Banaganapalli · R. Elango · N. A. Shaik (✉)
Princess Al-Jawhara Center of Excellence in Research of Hereditary Disorders, Department of Genetic Medicine, Faculty of Medicine, King Abdulaziz University, Jeddah, Saudi Arabia
e-mail: bbabajan@kau.edu.sa; relango@kau.edu.sa; nshaik@kau.edu.sa

N. A. Shaik et al. (eds.), *Essentials of Bioinformatics, Volume II*,
https://doi.org/10.1007/978-3-030-18375-2_13

GWAS	Genome-Wide Association Studies
RE	Regulatory Elements
TF	Transcription Factor
TFBS	Transcription Factor Binding Site
TSS	Transcriptional Start Site
UTR	Untranslated Regions

13.1 Introduction

Understanding the gene expression pattern and the identification of the specific genes expressed during different cellular and life processes are crucial for the understanding of various life processes and it also helps in defining the molecular pathology of different disease states (Baty et al. 2013). The complex and precise spatiotemporal gene expression often needs the presence of different cis-REs which are far placed from the promoter regions. Cell-lineage-specific TFs bind cis-REs distal to these promoters and also to those promoters which are more tightly regulated in spatiotemporal fashion and which needs external signals such as hormones, for example during cell growth and differentiation. Different cellular signals are integrated via promoter and cis-REs, which in turn regulate complex gene expression patterns in various cells and tissues in a coordinated manner (Sakabe et al. 2012).

Transcription in the higher eukaryotes, transcription is regulated by the interaction of enhancers and promoter regions which work in a coordinated manner. Several proteins like TFs, RNA polymerase, transcriptional cooactivators and histone modifying enzymes are needed for the expression of any gene at a given time. Both promoters and enhancers have some similar features such as TFBSs, but historically they are considered as two distinct classes of REs. The gene expression is initiated at transcriptional start sites (TSSs) when the promoter elements recruits the RNA polymerase II (Pol II) enzyme (Lenhard et al. 2012; Roy and Singer 2015; Schor et al. 2017; Vo Ngoc et al. 2017). Whereas, gene expression are promoted/enhanced by cis-regulatory DNA element known as enhancer elements. In general, enhancers are made up of clusters of TFBSs and it encompasses a few hundred base pairs (bps) to which various combinations of sequence-specific repressive and trans-activating factors binds. Enhancers has been found to be present in intergenic regions and exons. Interestingly, from their target genes, enhancers have been found to be present up to kilobases away and mediate their action via directly communicating with the promoter region (Lettice et al. 2003; Kleinjan and van Heyningen 2005). Unlike, promoters, enhancers can work in an orientation independent manner and can regulate transcription at another distal site using a different promoter. Interestingly, the binding of Pol II and general TFs to enhancers has also been observed (Koch et al. 2011). Recently, transcription has also been observed from enhancer elements (Tuan et al. 1992; De Santa et al. 2010; Kim et al. 2010; Lam et al. 2013).

In the process of development REs has been found to play a pivotal role. Any misregulation of these sequences may cause phenotypic consequences and can lead to diseases. Genome wide association studies (GWAS) and other similar studies (Stranger et al. 2011) have identified several disease- and trait-associated genetic variants of which a major chunk (~93%) of disease- and trait associated variants has been located within noncoding sequence which includes both promoter and enhancer elements (Cookson et al. 2009; Pomerantz et al. 2009; Musunuru et al. 2010; Harismendy et al. 2011). However, the impact of the mutations in the protein-coding regions can differ significantly from that of the mutations in cis-regulatory regions, even if they are regulating the same gene (Carroll 2008; Dimas et al. 2009). Mutations in the protein-coding regions are known to disrupt several aspects of gene regulation which include mRNA maturation, protein translation etc. and also protein folding and it's structure, whereas mutation related to the cis-regulatory regions like enhancers are mainly limited to cis effect on transcription (Sauna and Kimchi-Sarfaty 2011).

In modern biology, the genome wide study of TFBSs is one of the well and heavily researched area (Yáñez-Cuna et al. 2013). In general, prediction of the putative TFBSs are done in the upstream region of the gene TSS by searching specific short motifs. Identification of TFBS from a list of genes are of great importance as it can be helpful in interpreting gene expression data and comparing it with that of the TF function. Until recently, it was a tedious task to predict the promoter region associated TFBS from any gene list, which use to involve the gene promoter sequence extraction followed by search of pattern recognition via different motif databases like TRANSFAC (Wingender et al. 1996) or JASPAR (Bryne et al. 2008). Based on several experimental data it has been suggested that genes with similar expression patterns are either evolutionary or functionally correlated (Heyer et al. 1999; Spellman et al. 1998; Eisen et al. 1998). An important question in the field of gene expression is whether coexpressed genes are also co-regulated, that is, whether by sharing common cis-REs in their promoter regions these genes are most likely regulated by same TFs (Dottorini et al. 2013). Based on several experimental evidences it has currently been understood that proteins are not coded from most of the regions of human genome (Pennacchio et al. 2007). However, the function of these non-coding regions of genome is yet to be systematically categorize and understood. Moreover, several studies in human have suggested that complex gene regulation at the transcriptional level is functionally related to many discrete DNA elements which are often present hundred of kilobases (kb) far from their promoter regions (Lettice et al. 2003; Nobrega et al. 2003). Interestingly, various studies have suggested that the evolutionary sequence conservation as a good biological function indicator. Most of the tissue-specific enhancers that are functional during development are in the noncoding region of the genome and are highly evolutionarily conserved regions (ECRs) (Lettice et al. 2003; Waterston et al. 2002; Nobrega et al. 2003; Loots et al. 2000; Pennacchio et al. 2006).

To understand the pathology of a disease it is necessary to investigate the differences between the healthy and the diseased state and which in turn help in the treatment of the disease. Gene expression studies is a very useful tool to study the differences between healthy and a diseased state. Study related to the differentially-expressed genes (DEGs) are of importance as it is helpful in identification of DEGs in health and disease. Study of DEGs are of great importance in the field of clinical and pharmaceutical research as it can lead to identification of therapeutic targets, candidate biomarkers and to pinpoint the gene signature of diagnostic values. It has to be noted that sometime the individual DEGs studies may not provide a significant biological meaning on it's own but still will be useful when it is analysed along with other similar studies to perform integrated analysis related to a particular disease (Rodriguez-Esteban and Jiang 2017; Loging et al. 2007; Campbell et al. 2010).

Here we present a method named DiRE, which is a web server and a bioinformatics tool. It is a user friendly and easy to use online tool by means of which regulatory features can be investigated on the dataset of genes submitted by the users. Based on the user provided input genes and it's co-expression pattern (e.g. down- or up-regulation), function-specific (e.g. tissue, time) REs can be predicted by the DiRE server that can work as repressors or enhancers. DiRE will also give information about the important regulatory TFs which essentially bring about their effects (Gotea and Ovcharenko 2008). REs can be detected by DiRE which are located outside the proximal promoter regions as DiRE conduct the search of full gene locus. Function specific REs which consists of conserved and specifically associated TFBSs are predicted by the software. DiRE also scores the individual association of TFs shared by the input genes group with the biological function (Gotea and Ovcharenko 2008). Candidate REs are selected by the DiRE software from the gene loci which are based on pattern of precomputed alignments of inter-species conservation of genomic sequence from human, rodent, fish and other vertebrates (Aid-Pavlidis et al. 2009). Such alignment allows the DiRE software to detect phylogenetically conserved REs present in different species at the same genomic positions.

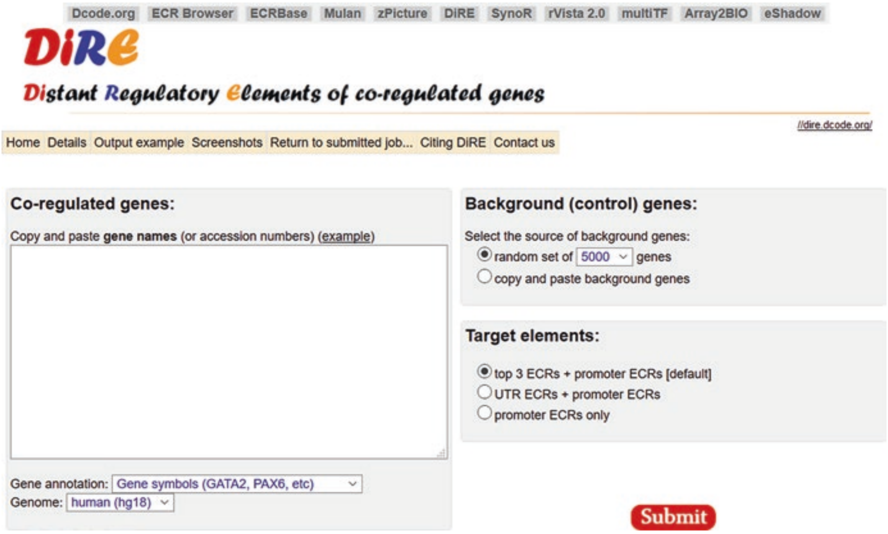
TRANSFAC Professional database (version 10.2) which works on position weight matrices (PWM) is used by DiRE tool (Ovcharenko et al. 2004). Around 7500 background genes are used by the DiRE. TFBSs that are extracted occur in less frequently in the 95 percent of permutation tests than in the original distribution (corresponding to p-value < 0.05 to observe the original distribution by chance) and which corresponds to at least a double increase in the original distribution density compared to an average pair density in permutation tests. In the DiRE tool the correction for the multiple hypothesis testing is done by using hypergeometric distribution with Bonferroni correction (Waterston et al. 2002). DiRE describe the 'importance score' as the TF occurrence product (% of tissue-specific TF with particular TFBS) and its weight candidate TF for each found TFBS. The importance score therefore is based on the specificity of the TF containing the specific TFBS and the TFBS abundance in tissue-specific TF (Wingender et al. 2000).

In this chapter we present a genomic tool called DiRE, which is a freely available web server. This tool can predict distant (outside of proximal promoter regions) REs of co-regulated genes in a user-friendly manner.

The tutorial described below is for the set of genes the users have:

Step 1: Open the server DiRE (https://dire.dcode.org/). Users will see the webpage as in Fig. 13.1 (see below for details).

Step 2: Copy and paste gene names (or accession numbers) of the co-regulated genes in the text box (Fig. 13.2a). List of records should be pasted with one record per line by the users in the main DiRE server window (Fig. 13.2a).



Co-regulated genes:

Copy and paste **gene names** (or accession numbers) (example)

Gene annotation: Gene symbols (GATA2, PAX6, etc)

Genome: human (hg18)

Background (control) genes:

Select the source of background genes:

● random set of 5000 genes

○ copy and paste background genes

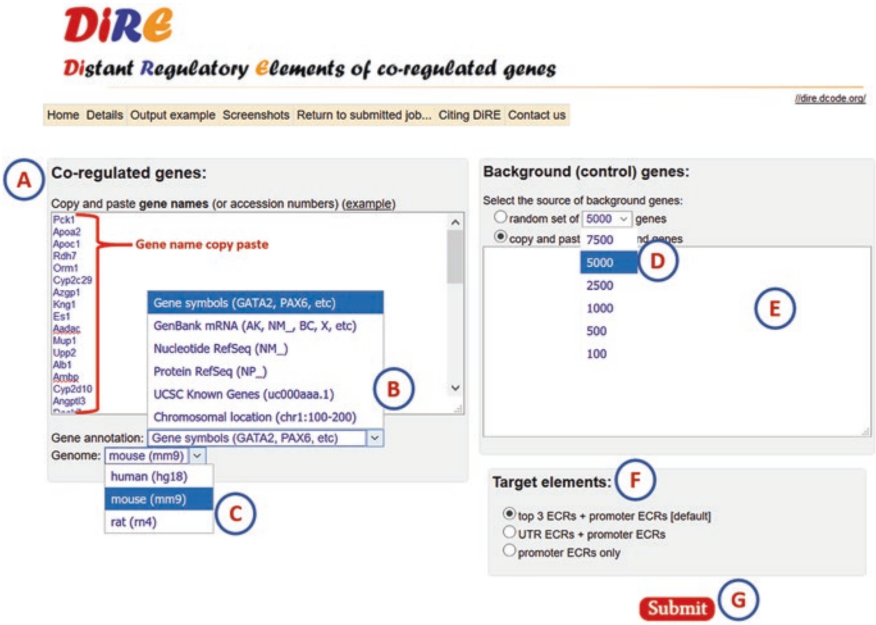
Target elements:

● top 3 ECRs + promoter ECRs [default]

○ UTR ECRs + promoter ECRs

○ promoter ECRs only

Fig. 13.1 Snapshot of the screen of the DiRE main webpage



Co-regulated genes:

Copy and paste **gene names** (or accession numbers) (example)

Pck1
Apoa2
Apoc1
Rdh7
Orm1
Cyp2c29
Azgp1
Kng1
Es1
Aadac
Mup1
Upp2
Alb1
Ambp
Cyp2d10
Angptl3

Gene annotation: Gene symbols (GATA2, PAX6, etc)

- Gene symbols (GATA2, PAX6, etc)
- GenBank mRNA (AK, NM_, BC, X, etc)
- Nucleotide RefSeq (NM_)
- Protein RefSeq (NP_)
- UCSC Known Genes (uc000aaa.1)
- Chromosomal location (chr1:100-200)

Genome: mouse (mm9)

- human (hg18)
- mouse (mm9)
- rat (rn4)

Background (control) genes:

Select the source of background genes:

○ random set of 5000 genes

● copy and past[e background ge]nd genes

- 7500
- 5000
- 2500
- 1000
- 500
- 100

Target elements:

● top 3 ECRs + promoter ECRs [default]

○ UTR ECRs + promoter ECRs

○ promoter ECRs only

Fig. 13.2 An example of the analysis session as shown in the main window of the DiRE server. Different panels suggest the available options for the analysis

Step 3: Users should make sure that the supplied gene list matches one of the following accepted data types like gene symbols, GenBank nucleotide or protein accession numbers, Protein RefSeq number, chromosomal location (or chromosome coordinates) or accession numbers from the UCSC known genes annotation. From the pull-down menu this can be selected (Fig. 13.2b).

Step 4: Further, users should choose the species from a pull-down menu (Fig. 13.2c) for which TFBS content and precomputed alignments presently for rat, human and mouse exist. Users should check that the coordinates should match the corresponding genome assembly of that species, in case the users choose the genes provided in the form of genomic coordinates (Fig. 13.2c).

Step 5: Users should choose the background (control) genes (Fig. 13.2d) which will serve as the background distribution of TFBS clusters. It has two option: (i) choose from the random set of genes (various static lists of 7500 background genes) chosen either from rat, human or the mouse genome in a random way. (Fig. 13.2d) or (ii) users can copy and paste their own list of background genes (Fig. 13.2e). Opting for option (i) benefits the users in that the list will remain same so that across different runs comparison of the results can be made and which can be reproduced. However, if there are some genes in both the gene background lists and the signal, they will be removed from the background set. In such a case the users should opt for option (ii) and provide background genes list of their choice (Fig. 13.2e), that could be very helpful if contrasting expression data exists as in the case of data generated from microarray gene experiments. Similar to co-regulated gene list, the user supplied list of genes should be formatted if the user chose for option (ii).

Note: Users should use at least a few thousand genes as the background gene to avoid the biased representation of random expectations.

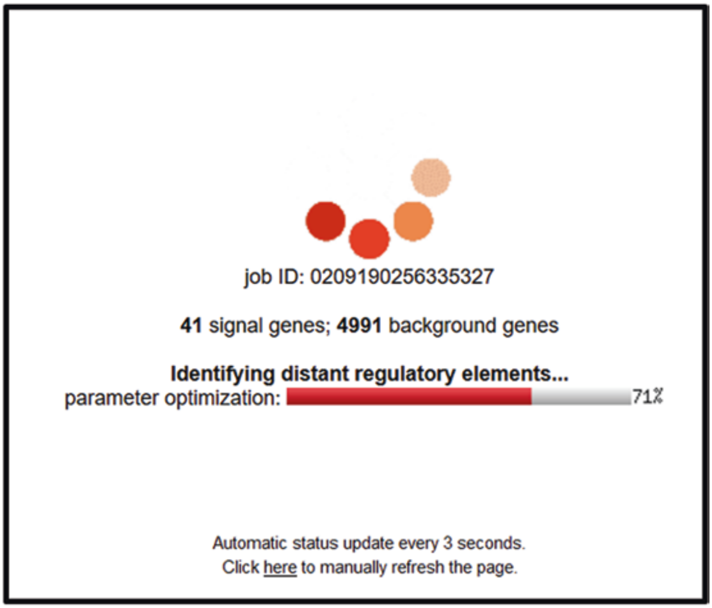
Step 6: Select the target element (Fig. 13.2f). Users may choose from the given options with different target elements. Options are available for the different set of target elements with evolutionary conserved regions (ECR). If users do not specify the "target element" field, then the target element "top 3 ECRs + promoter ECRs" as default will work.

Step 7: Finally, click "Submit" (Fig. 13.2g).

Step 8: After the job is submitted, while DiRE is running, the users will see the screen (Fig. 13.3). For user to return to the query page later, job ID should be noted. Depending on the user provided background and signal gene numbers this job may take up to thirty minutes.

Step 9: Upon the job completion, users will be taken to the result page (Fig. 13.4). This page displays the following sections as "Request ID" (Fig. 13.4a), "Potential Regulatory Elements" (Fig. 13.4b), "Candidate Transcription Factors" (Fig. 13.4c) and "Extra Data" (Fig. 13.4d) sections, in order from top to bottom (Fig. 13.4).

Step 10: As shown in (Fig. 13.5), "Request ID" is provided (Fig. 13.5a), to the users which can be used for retrieval of data in future. A permanent link is also provided to the users for future data retrieval (Fig. 13.5b).



job ID: 0209190256335327

41 signal genes; **4991** background genes

Identifying distant regulatory elements...

parameter optimization: 71%

Automatic status update every 3 seconds.

Click here to manually refresh the page.

Fig. 13.3 Screenshot while the DiRE is running after a job submission. Job ID shown at the top can be used later to return to the user query

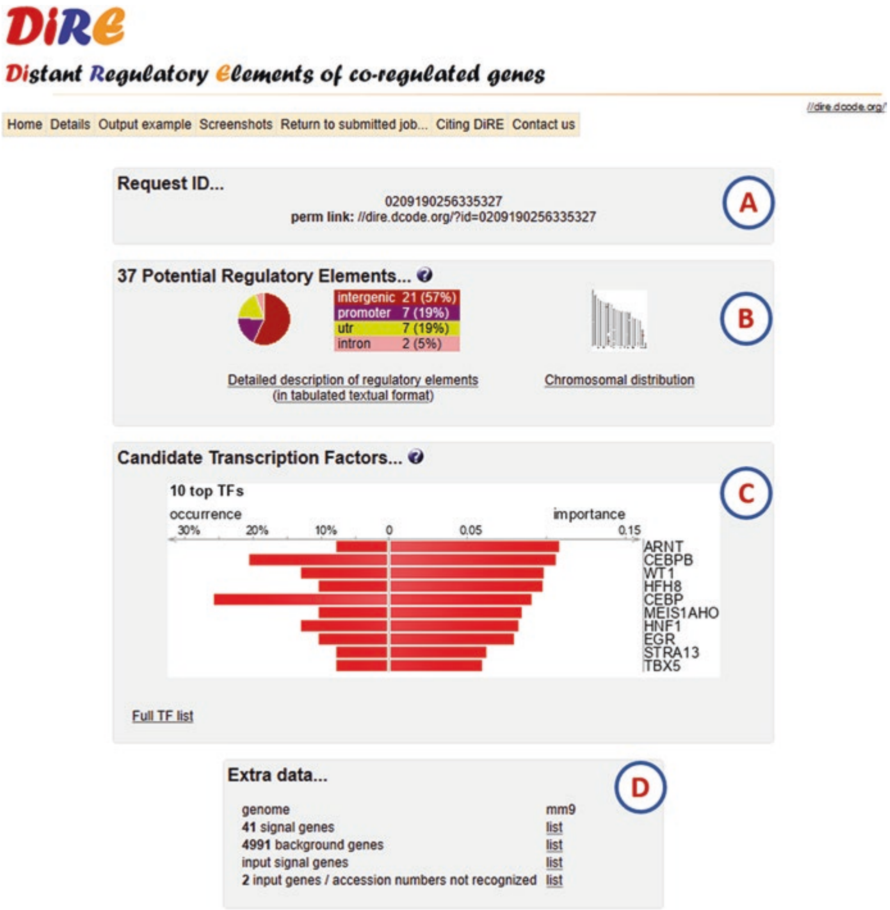
Step 11: Users can find the summary of the detected "Potential Regulatory elements" (Fig. 13.6a), categorized as intergenic, promoter, UTR and intron) showing the number and percentage of REs.

Step 12: Users can click on (Fig. 13.6b) and a new widow will appear allowing the users to see the genomic distribution of the predicted RE present relative to the genes they probably control which is shown by the red bars on the chromosomal representations (Fig. 13.7).

Step 13: Users can find the detail "Description of REs" by clicking tab as shown in (Fig. 13.6c) which will take the users to a new page (Fig. 13.8). For detail see below.

Step 14: Users can further click on icon "in tabulated textual format" (Fig. 13.6d) and go to a new page showing details of the REs in the tabular form.

Step 15: As mentioned in *Step 13*, users can see the details of the "Description of regulatory elements" (Fig. 13.8). Users may click on icon "Description of regulatory elements" (Fig. 13.6c) and go to a new page showing details of the REs in the tabular form (Fig. 13.8). Users may also see the description of candidate RE containing an annotation (Fig. 13.8a) based on the element location relative to the characteristics of the locus of gene (intron, intergenic, UTR) (Fig. 13.8b),



DiRE

Distant Regulatory Elements of co-regulated genes

//dire.dcode.org/

Home Details Output example Screenshots Return to submitted job... Citing DiRE Contact us

Request ID...

0209190256335327
perm link: //dire.dcode.org/?id=0209190256335327

37 Potential Regulatory Elements...

intergenic	21	(57%)
promoter	7	(19%)
utr	7	(19%)
intron	2	(5%)

Detailed description of regulatory elements (in tabulated textual format)

Chromosomal distribution

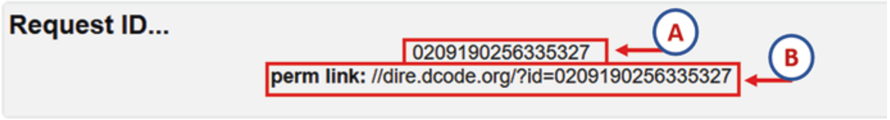
Candidate Transcription Factors...

Full TF list

Extra data...

genome	mm9
41 signal genes	list
4991 background genes	list
input signal genes	list
2 input genes / accession numbers not recognized	list

Fig. 13.4 Screenshot of the DiRE server upon job completion. This page displays various sections as shown



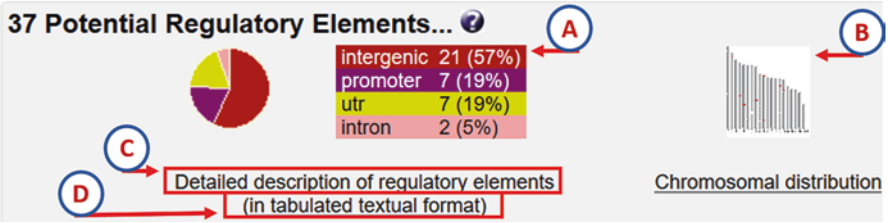
Request ID...

0209190256335327
perm link: //dire.dcode.org/?id=0209190256335327

Fig. 13.5 Screenshot of "Request ID" panel as shown after job completion is provided along with a permanent link to the users for future data retrieval

score (Fig. 13.8c), the gene locus coordinates (Fig. 13.8d), the gene official symbol(s) (Fig. 13.8e) and a list of TFBSs that has scored positively in that element (Fig. 13.8f).

Step 16: Furthermore, the users may resort the list by clicking in the column headers (Fig. 13.8).



37 Potential Regulatory Elements...

intergenic	21 (57%)
promoter	7 (19%)
utr	7 (19%)
intron	2 (5%)

A
B
C
D
Detailed description of regulatory elements
(in tabulated textual format)
Chromosomal distribution

Fig. 13.6 Screenshot of the "Potential regulatory elements" panel as shown after the job completion. It shows the summary of the detected REs

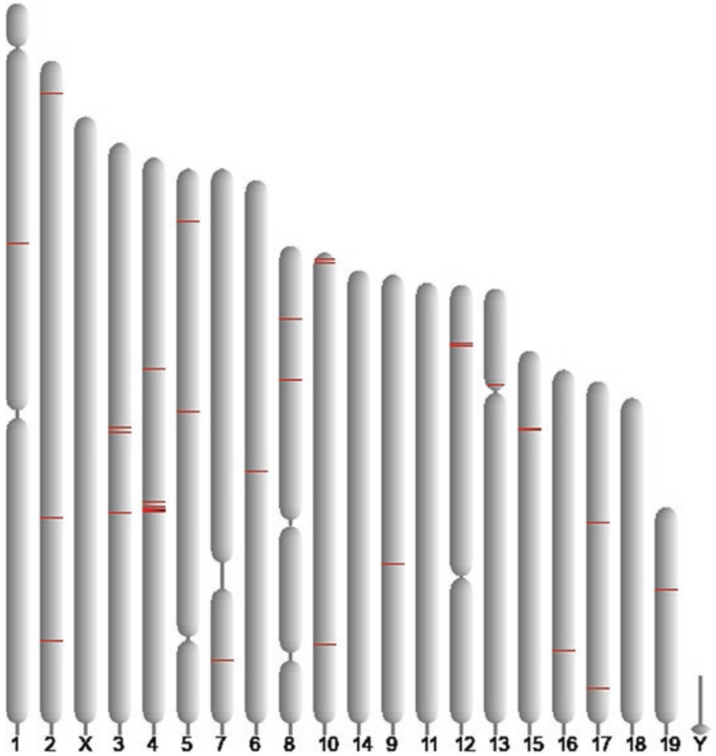
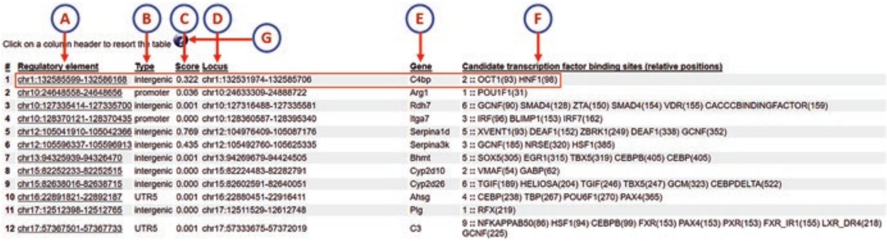


Fig. 13.7 Enlarged view of the "Chromosomal distribution" panel from the result section. Red bars on the chromosomal representations shows the predicted REs

Step 17: The users may also find the detail description of different column headers (Fig. 13.8) by clicking (?) tab as shown in (Fig. 13.8g) which will take the user to a new page (Fig. 13.9).

Step 18: Users may click on any RE (Fig. 13.8a) and it will be redirected to the ECR Browser (Aid-Pavlidis et al. 2009) (Fig. 13.10), where one can get a more comprehensive picture of the locus. In ECR Browser users can explore the genomic landscape and the conservation of individual candidate RE.



Click on a column header to resort the table

#	Regulatory element	Type	Score	Locus	Gene	Candidate transcription factor binding sites (relative positions)
1	chr1:132585599-132586168	intergenic	0.322	chr1:132531974-132585706	C4bp	2 :: OCT1(93) HNF1(98)
2	chr10:24648558-24648656	promoter	0.036	chr10:24633309-24888722	Arg1	1 :: POU1F1(31)
3	chr10:127335414-127335700	intergenic	0.001	chr10:127316488-127335581	Rdh7	6 :: GCNF(90) SMAD4(128) ZTA(150) SMAD4(154) VDR(155) CACCCBINDINGFACTOR(159)
4	chr10:128370121-128370435	promoter	0.000	chr10:128360587-128395340	Itga7	3 :: IRF(96) BLIMP1(153) IRF7(162)
5	chr12:105041910-105042366	intergenic	0.769	chr12:104976409-105087176	Serpina1d	5 :: XVENT1(93) DEAF1(152) ZBRK1(249) DEAF1(338) GCNF(352)
6	chr12:105596337-105596913	intergenic	0.435	chr12:105492760-105625335	Serpina3k	3 :: GCNF(185) NRSE(320) HSF1(385)
7	chr13:94325939-94326470	intergenic	0.001	chr13:94269679-94424505	Bhmt	5 :: SOX5(305) EGR1(315) TBX5(319) CEBPB(405) CEBP(405)
8	chr15:82252233-82252515	intergenic	0.000	chr15:82224483-82282791	Cyp2d10	2 :: VMAF(54) GABP(62)
9	chr15:82638016-82638715	intergenic	0.000	chr15:82602591-82640051	Cyp2d26	6 :: TGIF(189) HELIOSA(204) TGIF(246) TBX5(247) GCM(323) CEBPDELTA(522)
10	chr16:22891821-22892187	UTR5	0.001	chr16:22880451-22916411	Ahsg	4 :: CEBP(238) TBP(267) POU6F1(270) PAX4(365)
11	chr17:12512398-12512765	intergenic	0.000	chr17:12511529-12612748	Plg	1 :: RFX(219)
12	chr17:57367501-57367733	UTR5	0.001	chr17:57333675-57372019	C3	9 :: NFKAPPAB50(86) HSF1(94) CEBPB(99) FXR(153) PAX4(153) PXR(153) FXR_IR1(155) LXR_DR4(218) GCNF(225)

Fig. 13.8 View of the “Detailed description of regulatory elements” panel. A new page showing details of the REs in the tabular form

Step 19: Fig. 13.11 shows TFs found in candidate REs and the top 10 are displayed in this section (Fig. 13.11a). The occurrence and importance measures for each TF can be seen.

Step 20: Furthermore, the users may click on “Full TF list” (Fig. 13.11b) and a new window will appear with the TFs complete list that are positively linked with the co-regulated gene (Fig. 13.12). For each “TF” (Fig. 13.12a) users will discover the TF “Occurrence” in REs (Fig. 13.12b), and the “Importance” of TF (Fig. 13.12c). Users also may also resort the list by clicking in the column headers.

Step 21: The users may also find the detail description of different terms by clicking (?) tab as shown in (Fig. 13.12d) (Waterston et al. 2002) which will take the user to a new page (Fig. 13.13).

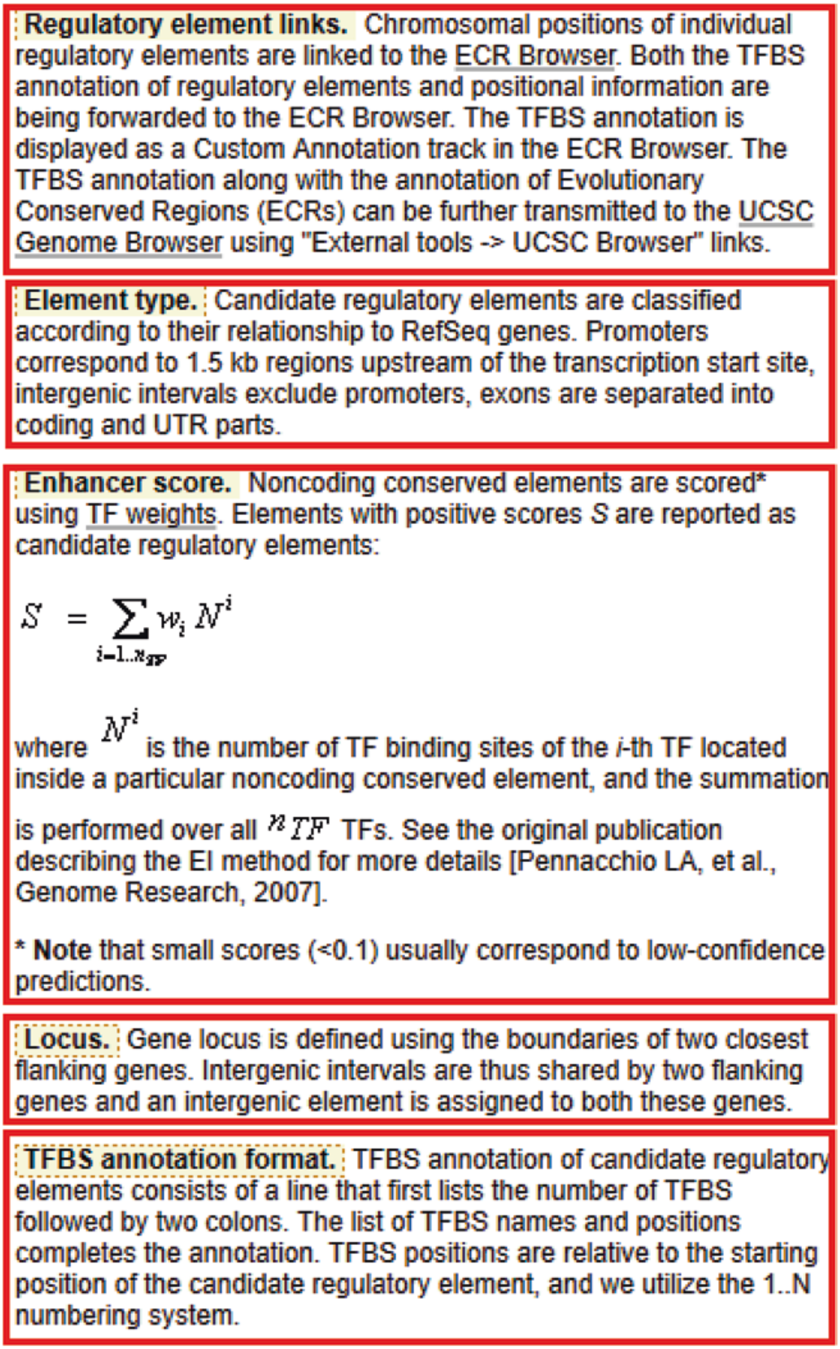
Step 22: For convenience of the users, original data is available through links (Fig. 13.14). The initial gene list used in the computation and their mapped position are given on the target genome.

Step 23: Finally, the users may return to the submitted job by clicking the tab as shown in (Fig. 13.15a). Users may enter a 16-digit request ID (Fig. 13.4a) to the box as shown in (Fig. 13.15c) and click the “Submit” button. Users should note that the 16-digit request ID to be pasted in Fig. 13.15c is from Fig. 13.4a.

Following are the advantages/use and limitations of the DiRE genomic tool:

A. *Use/advantages of DiRE tool*

1. It enables scientists to predict prevalent regulatory characteristics of co-regulated genes computationally.
2. In vertebrate genomes, DiRE can predict remote REs regardless of their relative location on the gene they control.
3. It can predict either repressor or enhancer elements, based on whether the genes of interest are down- or up-regulated, or general REs of any kind if the input data originates from a specific biological group that does not necessarily involve expression data (such as a Gene Ontology (Ashburner et al. 2000) or KEGG category (Altermann and Klaenhammer 2005)).



Regulatory element links. Chromosomal positions of individual regulatory elements are linked to the ECR Browser. Both the TFBS annotation of regulatory elements and positional information are being forwarded to the ECR Browser. The TFBS annotation is displayed as a Custom Annotation track in the ECR Browser. The TFBS annotation along with the annotation of Evolutionary Conserved Regions (ECRs) can be further transmitted to the UCSC Genome Browser using "External tools -> UCSC Browser" links.

Element type. Candidate regulatory elements are classified according to their relationship to RefSeq genes. Promoters correspond to 1.5 kb regions upstream of the transcription start site, intergenic intervals exclude promoters, exons are separated into coding and UTR parts.

Enhancer score. Noncoding conserved elements are scored* using TF weights. Elements with positive scores *S* are reported as candidate regulatory elements:

$$S = \sum_{i=1..n_{TF}} w_i N^i$$

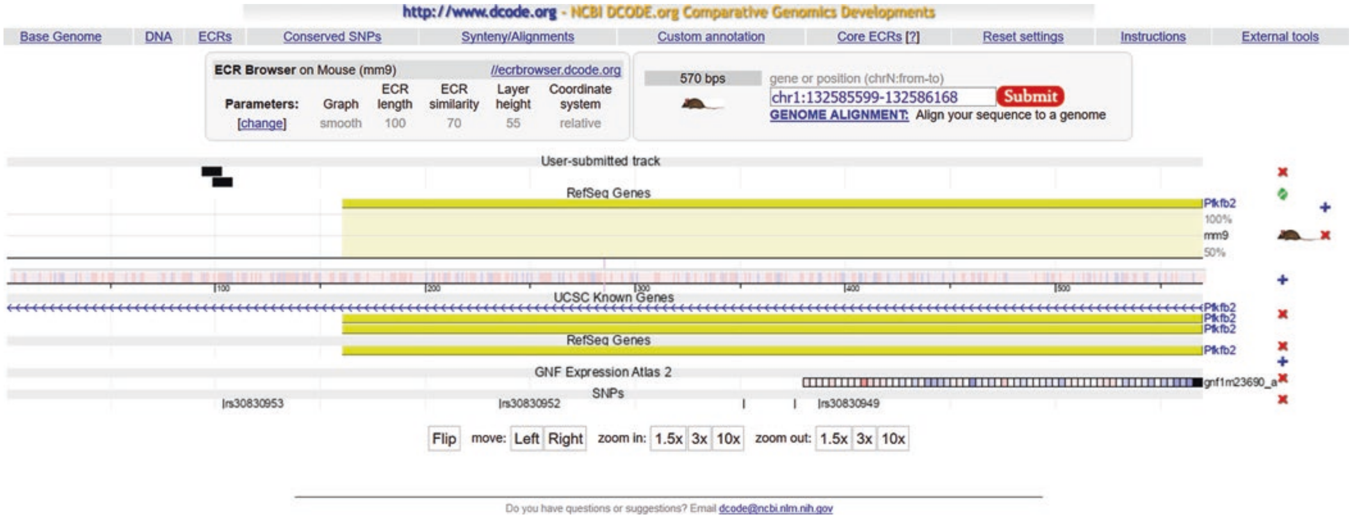
where N^i is the number of TF binding sites of the *i*-th TF located inside a particular noncoding conserved element, and the summation is performed over all n_{TF} TFs. See the original publication describing the EI method for more details [Pennacchio LA, et al., Genome Research, 2007].

* **Note** that small scores (<0.1) usually correspond to low-confidence predictions.

Locus. Gene locus is defined using the boundaries of two closest flanking genes. Intergenic intervals are thus shared by two flanking genes and an intergenic element is assigned to both these genes.

TFBS annotation format. TFBS annotation of candidate regulatory elements consists of a line that first lists the number of TFBS followed by two colons. The list of TFBS names and positions completes the annotation. TFBS positions are relative to the starting position of the candidate regulatory element, and we utilize the 1..N numbering system.

Fig. 13.9 Screenshot showing the detail description of different column headers as present in the "Detailed description of regulatory elements" panel



http://www.dcode.org - NCBI DCODE.org Comparative Genomics Developments

ECR Browser on Mouse (mm9) //ecrbrowser.dcode.org

Parameters:	Graph	ECR length	ECR similarity	Layer height	Coordinate system
[change]	smooth	100	70	55	relative

570 bps

gene or position (chrN:from-to): chr1:132585599-132586168

GENOME ALIGNMENT: Align your sequence to a genome

Do you have questions or suggestions? Email dcode@ncbi.nlm.nih.gov

Fig. 13.10 Screenshot of the ECR Browser showing a detail picture of the RE locus. Users can explore the conservation and the genomic landscape of each candidate RE

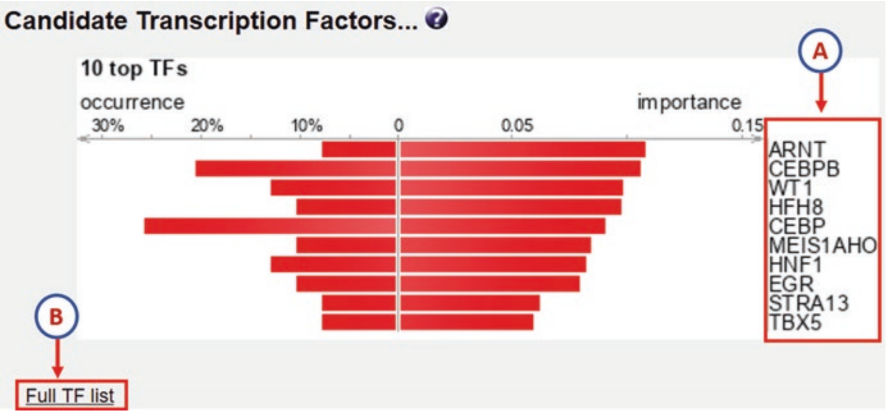


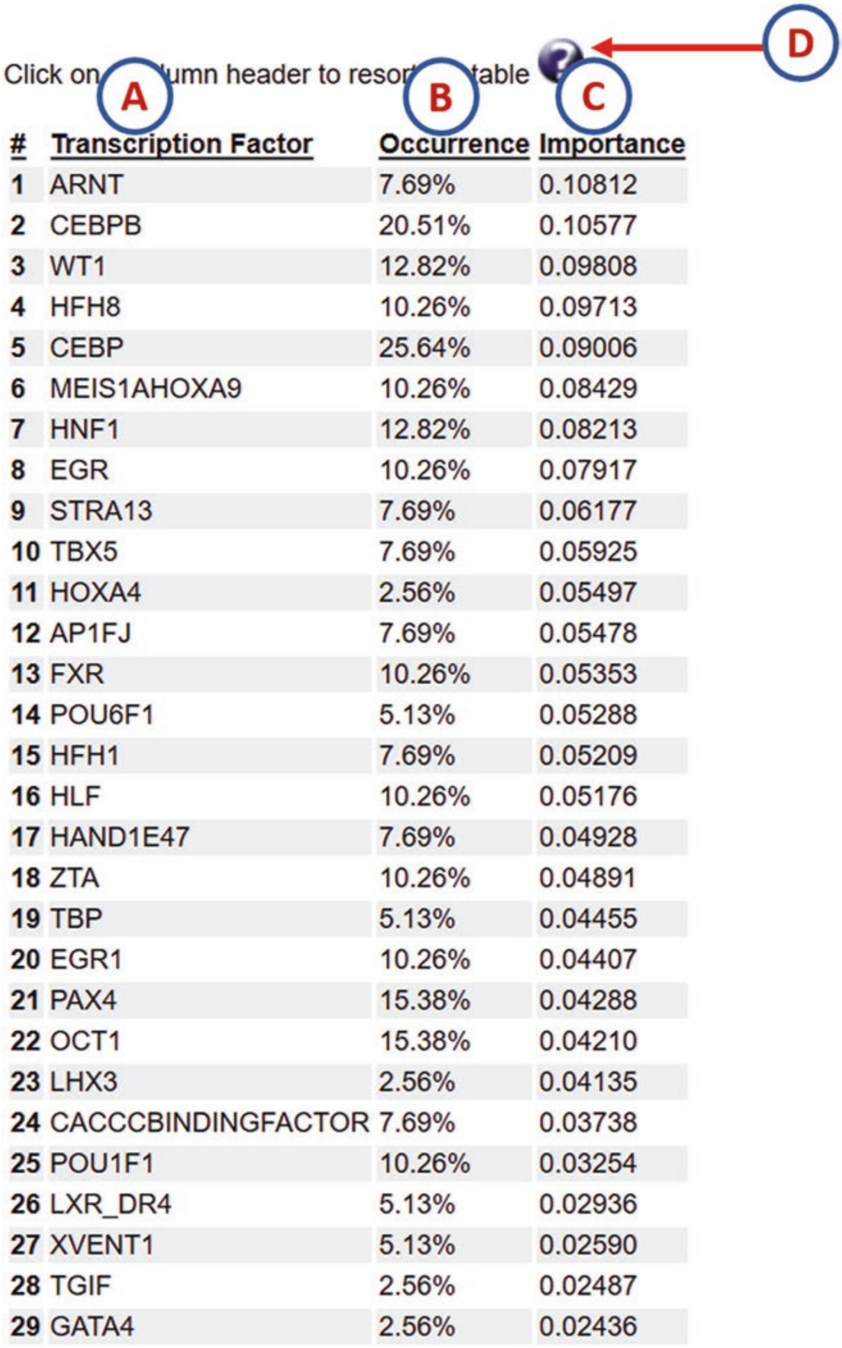
Fig. 13.11 Screenshot of the "Candidate Transcription Factors" panel as shown after the job completion. It shows TFs found in candidate REs and the ten most important ones are highlighted

4. This genomic tool can also be used to investigate for statistically over-represented TFBSs among all the conserved genes and in the clusters.
5. The TFs of DEGs enriched in KEGG pathways (He et al. 2017) can also be predicted from this database.
6. It can be used to construct the gene-TF regulatory network based on the predicted TF–DEGs pairs (Pennacchio et al. 2006).
7. DiRE may also be used to identify phylogenetically conserved REs that are present at the same genomic locations in various species.

All the above valuable points can lead to the discovery of therapeutic targets, gene signatures and candidate biomarkers, which will be useful for several disease diagnostics, including cancer.

B. *Limitations of DiRE tool*

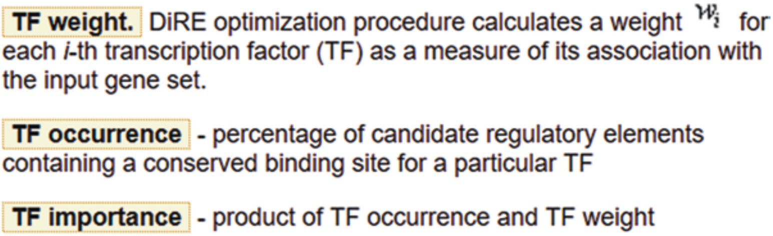
1. It should be remembered that the outcomes are based on a series of datasets which are precomputed.
2. Draft quality of distinct genomes could jeopardize the precomputed ECR Browser (Aid-Pavlidis et al. 2009) alignments.
3. Since DiRE defines TFBS based on the TRANSFAC database (Ovcharenko et al. 2004), therefore a poorly defined TF binding specificity or different TFs with very identical binding specificities or a missing TF may adversely impact the quality of DiRE predictions.



Click on column header to resort table

#	Transcription Factor	Occurrence	Importance
1	ARNT	7.69%	0.10812
2	CEBPB	20.51%	0.10577
3	WT1	12.82%	0.09808
4	HFH8	10.26%	0.09713
5	CEBP	25.64%	0.09006
6	MEIS1AHOXA9	10.26%	0.08429
7	HNF1	12.82%	0.08213
8	EGR	10.26%	0.07917
9	STRA13	7.69%	0.06177
10	TBX5	7.69%	0.05925
11	HOXA4	2.56%	0.05497
12	AP1FJ	7.69%	0.05478
13	FXR	10.26%	0.05353
14	POU6F1	5.13%	0.05288
15	HFH1	7.69%	0.05209
16	HLF	10.26%	0.05176
17	HAND1E47	7.69%	0.04928
18	ZTA	10.26%	0.04891
19	TBP	5.13%	0.04455
20	EGR1	10.26%	0.04407
21	PAX4	15.38%	0.04288
22	OCT1	15.38%	0.04210
23	LHX3	2.56%	0.04135
24	CACCCBINDINGFACTOR	7.69%	0.03738
25	POU1F1	10.26%	0.03254
26	LXR_DR4	5.13%	0.02936
27	XVENT1	5.13%	0.02590
28	TGIF	2.56%	0.02487
29	GATA4	2.56%	0.02436

Fig. 13.12 Screenshot of the list of TFs that are positively associated with the co-regulated genes. For each TF, its “Occurrence” in REs and its “Importance” are shown

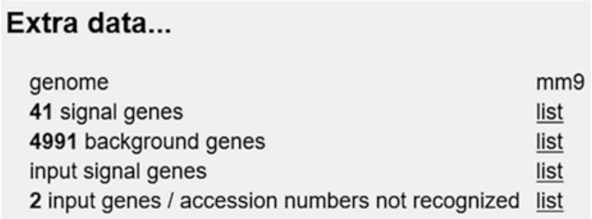


TF weight. DiRE optimization procedure calculates a weight w_i for each *i*-th transcription factor (TF) as a measure of its association with the input gene set.

TF occurrence - percentage of candidate regulatory elements containing a conserved binding site for a particular TF

TF importance - product of TF occurrence and TF weight

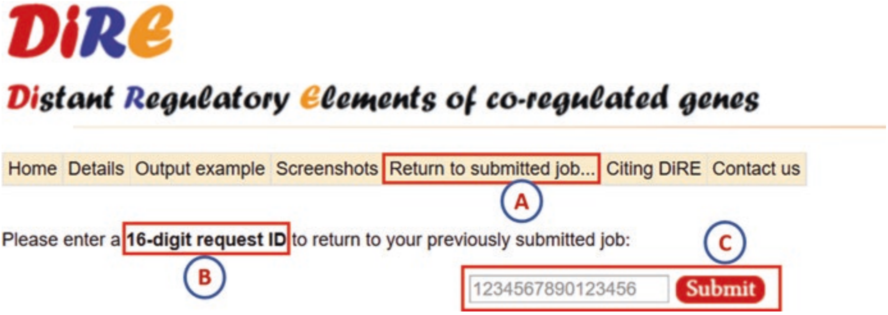
Fig. 13.13 Screenshot showing the detail description of different terms as shown in Fig. 13.12



Extra data...

genome	mm9
41 signal genes	list
4991 background genes	list
input signal genes	list
2 input genes / accession numbers not recognized	list

Fig. 13.14 Screenshot of the "Extra Data" panel as shown after the job completion. For users, original data is available through the links provided. Original gene list used in the computation and their mapped location on the target genome are also provided



Please enter a **16-digit request ID** to return to your previously submitted job:

Fig. 13.15 As shown in the figure user can use the unique 16-digit request ID to return to the submitted job

13.2 Conclusion

This chapter would enable investigators to predict computationally the prevalent regulatory features of co-regulated genes. The above described online web server is a freely available and easy-to-use genomic tool. We believe, the step by step method described in this chapter will allow biologist with little or no experience in bioinformatics to use such an important genomic tool. The above described method will provide the molecular biologist, clinician etc an easy access to study DEGs in health and disease conditions. Using the DiRE tool may allow researcher in isolating biomarkers specific for disease monitoring and it's progression and development.

References

Aid-Pavlidis T, Pavlidis P, Timmusk T (2009) Meta-coexpression conservation analysis of microarray data: a “subset” approach provides insight into brain-derived neurotrophic factor regulation. BMC Genomics 10:420

Altermann E, Klaenhammer TR (2005) PathwayVoyager: pathway mapping using the Kyoto Encyclopedia of Genes and Genomes (KEGG) database. BMC Genomics 6:60

Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM et al (2000) Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. Nat Genet 25(1):25–29

Baty F, Rüdiger J, Miglino N, Kern L, Borger P, Brutsche M (2013) Exploring the transcription factor activity in high-throughput gene expression data using RLQ analysis. BMC Bioinformatics 14:178

Bryne JC, Valen E, Tang MH, Marstrand T, Winther O, da Piedade I et al (2008) JASPAR, the open access database of transcription factor-binding profiles: new content and tools in the 2008 update. Nucleic Acids Res 36(Database issue):D102–D106

Campbell SJ, Gaulton A, Marshall J, Bichko D, Martin S, Brouwer C et al (2010) Visualizing the drug target landscape. Drug Discov Today 15(1–2):3–15

Carroll SB (2008) Evo-devo and an expanding evolutionary synthesis: a genetic theory of morphological evolution. Cell 134(1):25–36

Cookson W, Liang L, Abecasis G, Moffatt M, Lathrop M (2009) Mapping complex disease traits with global gene expression. Nat Rev Genet 10(3):184–194

De Santa F, Barozzi I, Mietton F, Ghisletti S, Polletti S, Tusi BK et al (2010) A large fraction of extragenic RNA pol II transcription sites overlap enhancers. PLoS Biol 8(5):e1000384

Dimas AS, Deutsch S, Stranger BE, Montgomery SB, Borel C, Attar-Cohen H et al (2009) Common regulatory variation impacts gene expression in a cell type-dependent manner. Science 325(5945):1246–1250

Dottorini T, Palladino P, Senin N, Persampieri T, Spaccapelo R, Crisanti A (2013) CluGene: a bioinformatics framework for the identification of co-localized, co-expressed and co-regulated genes aimed at the investigation of transcriptional regulatory networks from high-throughput expression data. PLoS One 8(6):e66196

Eisen MB, Spellman PT, Brown PO, Botstein D (1998) Cluster analysis and display of genome-wide expression patterns. Proc Natl Acad Sci U S A 95(25):14863–14868

Gotea V, Ovcharenko I (2008) DiRE: identifying distant regulatory elements of co-expressed genes. Nucleic Acids Res 36(Web Server issue):W133–W139

Harismendy O, Notani D, Song X, Rahim NG, Tanasa B, Heintzman N et al (2011) 9p21 DNA variants associated with coronary artery disease impair interferon-γ signalling response. Nature 470(7333):264–268

He Y, Liu J, Zhao Z, Zhao H (2017) Bioinformatics analysis of gene expression profiles of esophageal squamous cell carcinoma. Dis Esophagus 30(5):1–8

Heyer LJ, Kruglyak S, Yooseph S (1999) Exploring expression data: identification and analysis of coexpressed genes. Genome Res 9(11):1106–1115

Kim TK, Hemberg M, Gray JM, Costa AM, Bear DM, Wu J et al (2010) Widespread transcription at neuronal activity-regulated enhancers. Nature 465(7295):182–187

Kleinjan DA, van Heyningen V (2005) Long-range control of gene expression: emerging mechanisms and disruption in disease. Am J Hum Genet 76(1):8–32

Koch F, Fenouil R, Gut M, Cauchy P, Albert TK, Zacarias-Cabeza J et al (2011) Transcription initiation platforms and GTF recruitment at tissue-specific enhancers and promoters. Nat Struct Mol Biol 18(8):956–963

Lam MT, Cho H, Lesch HP, Gosselin D, Heinz S, Tanaka-Oishi Y et al (2013) Rev-Erbs repress macrophage gene expression by inhibiting enhancer-directed transcription. Nature 498(7455):511–515

Lenhard B, Sandelin A, Carninci P (2012) Metazoan promoters: emerging characteristics and insights into transcriptional regulation. Nat Rev Genet 13(4):233–245

Lettice LA, Heaney SJ, Purdie LA, Li L, de Beer P, Oostra BA et al (2003) A long-range Shh enhancer regulates expression in the developing limb and fin and is associated with preaxial polydactyly. Hum Mol Genet 12(14):1725–1735

Loging W, Harland L, Williams-Jones B (2007) High-throughput electronic biology: mining information for drug discovery. Nat Rev Drug Discov 6(3):220–230

Loots GG, Locksley RM, Blankespoor CM, Wang ZE, Miller W, Rubin EM et al (2000) Identification of a coordinate regulator of interleukins 4, 13, and 5 by cross-species sequence comparisons. Science 288(5463):136–140

Musunuru K, Strong A, Frank-Kamenetsky M, Lee NE, Ahfeldt T, Sachs KV et al (2010) From noncoding variant to phenotype via SORT1 at the 1p13 cholesterol locus. Nature 466(7307):714–719

Nobrega MA, Ovcharenko I, Afzal V, Rubin EM (2003) Scanning human gene deserts for long-range enhancers. Science 302(5644):413

Ovcharenko I, Nobrega MA, Loots GG, Stubbs L (2004) ECR browser: a tool for visualizing and accessing data from comparisons of multiple vertebrate genomes. Nucleic Acids Res 32(Web Server issue):W280–W286

Pennacchio LA, Ahituv N, Moses AM, Prabhakar S, Nobrega MA, Shoukry M et al (2006) In vivo enhancer analysis of human conserved non-coding sequences. Nature 444(7118):499–502

Pennacchio LA, Loots GG, Nobrega MA, Ovcharenko I (2007) Predicting tissue-specific enhancers in the human genome. Genome Res 17(2):201–211

Pomerantz MM, Ahmadiyeh N, Jia L, Herman P, Verzi MP, Doddapaneni H et al (2009) The 8q24 cancer risk variant rs6983267 shows long-range interaction with MYC in colorectal cancer. Nat Genet 41(8):882–884

Rodriguez-Esteban R, Jiang X (2017) Differential gene expression in disease: a comparison between high-throughput studies and the literature. BMC Med Genet 10(1):59

Roy AL, Singer DS (2015) Core promoters in transcription: old problem, new insights. Trends Biochem Sci 40(3):165–171

Sakabe NJ, Savic D, Nobrega MA (2012) Transcriptional enhancers in development and disease. Genome Biol 13(1):238

Sauna ZE, Kimchi-Sarfaty C (2011) Understanding the contribution of synonymous mutations to human disease. Nat Rev Genet 12(10):683–691

Schor IE, Degner JF, Harnett D, Cannavò E, Casale FP, Shim H et al (2017) Promoter shape varies across populations and affects promoter evolution and expression noise. Nat Genet 49(4):550–558

Spellman PT, Sherlock G, Zhang MQ, Iyer VR, Anders K, Eisen MB et al (1998) Comprehensive identification of cell cycle-regulated genes of the yeast Saccharomyces cerevisiae by microarray hybridization. Mol Biol Cell 9(12):3273–3297

Stranger BE, Stahl EA, Raj T (2011) Progress and promise of genome-wide association studies for human complex trait genetics. Genetics 187(2):367–383

Tuan D, Kong S, Hu K (1992) Transcription of the hypersensitive site HS2 enhancer in erythroid cells. Proc Natl Acad Sci U S A 89(23):11219–11223

Vo Ngoc L, Wang YL, Kassavetis GA, Kadonaga JT (2017) The punctilious RNA polymerase II core promoter. Genes Dev 31(13):1289–1301

Waterston RH, Lindblad-Toh K, Birney E, Rogers J, Abril JF, Agarwal P et al (2002) Initial sequencing and comparative analysis of the mouse genome. Nature 420(6915):520–562

Wingender E, Dietze P, Karas H, Knüppel R (1996) TRANSFAC: a database on transcription factors and their DNA binding sites. Nucleic Acids Res 24(1):238–241

Wingender E, Chen X, Hehl R, Karas H, Liebich I, Matys V et al (2000) TRANSFAC: an integrated system for gene expression regulation. Nucleic Acids Res 28(1):316–319

Yáñez-Cuna JO, Kvon EZ, Stark A (2013) Deciphering the transcriptional cis-regulatory code. Trends Genet 29(1):11–22

Index

N. A. Shaik et al. (eds.), *Essentials of Bioinformatics, Volume II*,
https://doi.org/10.1007/978-3-030-18375-2