Translational Bioinformatics 19
Series Editor: Xiangdong Wang, MD, PhD, Prof

Kang Ning *Editor*

# of Multi-Omics Data Integration and Data Mining

Techniques and Applications

Springer

Translational Bioinformatics

Volume 19

Translational bioinformatics is defined as the development of storage-related, analytic, and interpretive methods to optimize the transformation of increasingly voluminous biomedical data, and genomic data in particular, into proactive, predictive, preventive, and participatory health. Translational bioinformatics includes research on the development of novel techniques for the integration of biological and clinical data and the evolution of clinical informatics methodology to encompass biological observations. The end product of translational bioinformatics is the newly found knowledge from these integrative efforts that can be disseminated to a variety of stakeholders including biomedical scientists, clinicians, and patients. Issues related to database management, administration, or policy will be coordinated through the clinical research informatics domain. Analytic, storage-related, and interpretive methods should be used to improve predictions, early diagnostics, severity monitoring, therapeutic effects, and the prognosis of human diseases.

Kang Ning
Editor

Methodologies of Multi-Omics Data Integration and Data Mining

Techniques and Applications

Springer

Editor
Kang Ning
Department of Bioinformatics and Systems
Biology, Center of AI biology, College of
Life Science and Technology
Huazhong University of Science and
Technology
Wuhan, China

ISSN 2213-2775 ISSN 2213-2783 (electronic)
Translational Bioinformatics
ISBN 978-981-19-8209-5 ISBN 978-981-19-8210-1 (eBook)
https://doi.org/10.1007/978-981-19-8210-1

This Springer imprint is published by the registered company Springer Nature Singapore Pte Ltd.
The registered company address is: 152 Beach Road, #21-01/04 Gateway East, Singapore 189721, Singapore

Preface

Many biomedical and clinical questions can now be answered using the wealth of multi-omics data that has become available in the age of omics. In the process, however, it has also created hurdles in the integration, mining, and comprehensive understanding of omics data.

Several biomedical applications are of special interest among various applications. The first is about cancer omics, which is always at the forefront of omics data analysis. Previous cancer omics research has focused on genomics and transcriptomics, whereas current multi-omics analysis would undoubtedly be the focus of in-depth mining of cancer progression principles. The second is about inflammation disease omics, which has piqued the interest of the research community, in part due to the growing proportion of patients suffering from inflammatory diseases such as arthritis. Multi-omics research, particularly on the dynamics of multi-omics, would shed light on a better understanding of the development of inflammation disease. The third is about the microbiome, which is a current hot topic: microbial communities are now thought to be linked to a variety of diseases, including T2D, IBD, and others. And, as with so many other questions, the principle governing the regulation of the microbiome on these various diseases remains a mystery. As a result, metagenomic data mining and explanations would be extremely valuable in the omics field. The fourth topic is omics data integration, which is related not only to databases and online data analysis pipelines, as well as visualization tools, but also to the development of various methods for multi-omics data correlation analysis or even causal or dynamic pattern discovery in the data integration procedure. Only through such high-level data integration could a solid foundation for data mining be built. Finally, method development is critical for a better understanding of hidden principles that can only be recovered by novel creative artificial intelligence tools.

This book has covered not only multi-omics big-data integration and data-mining techniques, but also cutting-edge researches in the principles and applications of several omics, including cancer omics, inflammation disease omics, and microbiome research. (1) Multi-omics big-data integration and data-mining techniques: Data integration and data-mining techniques will be introduced, along with illustrative examples and figures, to provide a better understanding of the essence of the definitions of both multi-omics and data mining, as well as how they can be combined to gain the most insights from these omics data. (2) Advancement in concrete research on multi-omics big-data: readers will learn the fundamental procedures for conducting representative and concrete multi-omics studies: given a set of omics data, how data-mining techniques can be applied to meet the needs of specific biological questions of interest. (3) Cutting-edge research in applications such as cancer omics, inflammation disease, and microbiome research: three topics would be highlighted out of many applications, one on cancer omics data analysis and explanations, another on inflammation disease, and another on specifically featured microbiome applications such as those related to T2D and IBD. (4) Contemporary data resources, tools, and analytical platforms will also be featured for readers to gain hands-on experience.

Intended as a book on the biomedical big-data expedition in the omics age, this book focuses on data integration and data-mining methods for multi-omics researches, explaining the "What," "Why," and "How" of the topic in detail and with supporting examples. It is an attempt to bridge the gap between biomedical multi-omics big data and data-mining techniques to obtain optimal practices in contemporary bioinformatics and in-depth insights into biomedical and clinical questions.

Wuhan, China Kang Ning

About the Book

This book features multi-omics big-data integration and data-mining techniques. In the omics age, the paramount of multi-omics data from various sources is the new challenge we are facing, but it also provides clues for several biomedical or clinical applications. For multi-omics research, this book discusses in detail and with examples how to integrate data and performed data mining. This book focuses on data integration and data-mining methods for multi-omics research, which explains in detail and with supportive examples the "What," "Why," and "How" of the topic. The contents are organized into eight chapters, out of which one is for the introduction, followed by four chapters devoted to omics integration techniques and data-mining methods, and three chapters devoted to the applications of multi-omics analyses, where data-mining methods are used to demonstrate how multi-omics analyses can be used in practice. This book is an attempt to bridge the gap between biomedical multi-omics big data and the data-mining techniques, for the best practice of contemporary bioinformatics and the in-depth insights for the biomedical questions. It would be of interest to the researchers and practitioners who want to conduct multi-omics studies in cancer, inflammation disease, and microbiome researches.

Contents

1 Introduction to Multi-Omics 1
Kang Ning and Yuxue Li

Part I Omics Integration Techniques

2 Biomedical Applications: The Need for Multi-Omics 13
Yuxue Li and Kang Ning

3 -Omics Technologies and Big Data 33
Ansgar Poetsch and Yuxue Li

4 Multi-Omics Data Mining Techniques: Algorithms and Software . . . 55
Min Tang, Yi Liu, and Xun Gong

Part II Applications of Multi-omics Analyses

5 Multi-Omics Data Analysis for Cancer Research: Colorectal Cancer, Liver Cancer and Lung Cancer 77
Hantao Zhang, Xun Gong, and Min Tang

6 Multi-Omics Data Analysis for Inflammation Disease Research: Correlation Analysis, Causal Analysis and Network Analysis 101
Maozhen Han, Na Zhang, Zhangjie Peng, Yujie Mao, Qianqian Yang, Yiyang Chen, Mengfei Ren, and Weihua Jia

7 Microbiome Data Analysis and Interpretation: Correlation Inference and Dynamic Pattern Discovery 119
Kang Ning and Yuxue Li

8 Current Progress of Bioinformatics for Human Health 145
Jin Zhao, Shu Zhang, Shunyao Wu, Wenke Zhang, and Xiaoquan Su

Concluding Remarks 163

References 167

About the Editor

Kang Ning Professor, PI of Microbial Bioinformatics Group, Director of Department of Bioinformatics and Systems Biology, School of Life Science and Technology, Huazhong University of Science and Technology, China.

Kang obtained his BS in Computer Science from USTC, and PhD in Bioinformatics from NUS. He has had his Post-Doc training in Bioinformatics from the University of Michigan.

Kang has more than 20 years of experiences in bioinformatics for omics big-data integration, microbiome analyses, and single-cell analyses. His current research interests include AI method for multi-omics especially metagenomics data mining, as well as their applications. He is also interested in synthetic biology and high-performance computation.

Kang is the leading or corresponding author of over 100 papers and reviews on leading journals including PNAS, Gut, Genome Biology, Nucleic Acids Research, Briefings in Bioinformatics and Bioinformatics, which have more than 3,000 citations. He has been the committee member of several national bioinformatics and biology big-data committees in China. He serves as an editorial board member of several journals including Genomics Proteomics and Bioinformatics, Microbiology Spectrum and Scientific Reports, and served as reviewers for several international funding agencies including UK-BBSRC and UK-NERC. He has collaborations with biologists, doctors, and statisticians in many countries and has given talks on international conferences for more than hundred times. For details, please refer to his official website as: http://www.microbioinformatics.org/.

Chapter 1
Introduction to Multi-Omics

Kang Ning and Yuxue Li

The rapid development of technologies and informatics tools for producing and interpreting massive biological data sets (omics data) has resulted in a paradigm shift in how we approach biomedical challenges (Manzoni et al. 2018). Large data sets are typically generated during genomics, transcriptomics, proteomics, microbiomics, and metabolomics research (Osier et al. 2017). With the advancement of these omics investigations, multi-omics research has emerged as one of the most promising venues for a deeper understanding of biological problems (Sun and Hu 2016). As the name suggests, multi-omics encompasses all digital genetic resources relevant to the research objectives, and its related research will automatically generate more comprehensive information to achieve the purpose of the research.

Multi-omics studies typically include omics data from multiple sources, including genomics, transcriptomics, proteomics, epigenomics, and microbiomics (Chung and Kang 2019). Genomics refers to omics data derived from DNA materials (Manzoni et al. 2018). Transcriptomics is the study of omics data derived from RNA materials (Manzoni et al. 2018). Proteomics is the collection of omics data from protein materials (Manzoni et al. 2018). Epigenomics refers to omics data derived from the whole range of epigenetic alterations on genetic material (Casadesús and Noyer-Weidner 2013). Microbiomics refers to omics data derived from a microbial community's entire set of genetic materials (Kumar 2000). Each of these omics represents a different part of the research goal, and when combined, they could disclose the regulatory patterns and principles that govern how genetic materials regulate genotypes (Fig. 1.1). On a more generalized scope, the omics can also

K. Ning (✉) · Y. Li
Key Laboratory of Molecular Biophysics of the Ministry of Education, Hubei Key Laboratory of Bioinformatics and Molecular-imaging, Center of Artificial Intelligence Biology, Department of Bioinformatics and Systems Biology, College of Life Science and Technology, Huazhong University of Science and Technology, Wuhan, China
e-mail: ningkang@hust.edu.cn

K. Ning (ed.), *Methodologies of Multi-Omics Data Integration and Data Mining*,
Translational Bioinformatics 19, https://doi.org/10.1007/978-981-19-8210-1_1

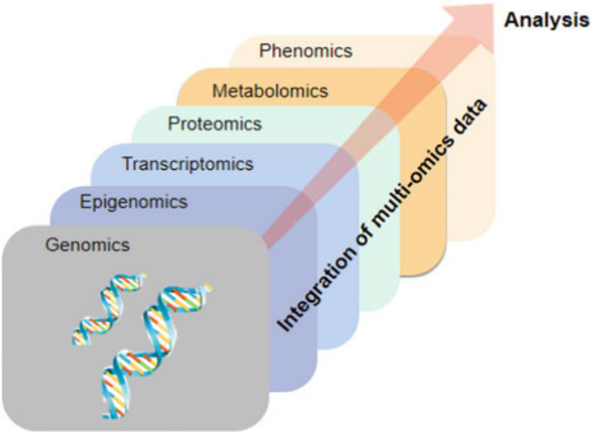

Fig. 1.1 The generalized definition of multi-omics. Large-scale acquisition of omics data from different molecular levels such as genome, transcriptome, proteome, epigenome, metabolome, microbiome, etc., and integrated analysis to achieve a deeper understanding of biological processes and molecular mechanisms

include those data from bioimaging, biosensors, and even social networks (Antonelli et al. 2019; Sriram and Subrahmanian 2020; Loizou 2016).

1.1 The History of Omics

The omics studies have quite a long history. Back in 1958, the first sequencing technique emerged, as Frederick Sanger has invented the protein sequencing methods, especially the amino acid sequence of insulin (Heather and Chain 2016). However, sequencing technology did not develop significantly during the next twenty to thirty years. DNA was originally extracted in 1869, it was not until more than a century later that the first genomes were sequenced, making genomics a relatively new field that truly began in 1970s.

1.1.1 1971–1910: Discovery of DNA

In 1871, Friedrich Miescher published a paper identifying the presence of nuclein and associated proteins in the nucleus. This is what we now call DNA, which is the foundation of the field of genomics.

Walter Sutton and Theodor Boveri discovered in 1904 that chromosomes appeared in pairs, with one inherited from each parent., which is known as the theory of chromosome inheritance. In 1910, Albert Kossel discovered the five nucleotide bases: adenine (A), cytosine (C), guanine (G), thymine (T), and uracil (U).

1.1.2 1950–1968: Development of Knowledge about DNA

Erwin Chargaff discovered the base pairing of adenosine, cytosine, guanine, and thymine nucleotides in 1950. He discovered that the concentrations of thymine and adenine or cytosine and guanine in DNA samples are always equal. As a result, he concluded that adenosine and thymine form a chromosome pair, while cytosine and guanine form a chromosome pair.

In 1952, Alfred Hershey and Martha Chase proved through a series of experiments that it was DNA, not protein, that carried inherited genetic features. The following year, the double helix structure of DNA was discovered by James Watson and Francis Crick (Portin 2014). A research team led by Marshall Nirenberg and Har Gobind Khorana discovered what is now known as DNA "codons" in 1961.

1.1.3 1977–Present: Sequencing of DNA Related Stories

Frederick Sanger developed a DNA sequencing technology in 1977 to sequence the first complete genome, known as the phiX174 virus, which opened the door to new possibilities in genomics. In 1983, Dr. Kary Mullis invented the polymerase chain reaction (PCR) technique for amplifying DNA (García-Quesada et al. 2021). The first bacterial genome sequence, Haemophilus influenza, was completed in 1995 (Fraser and Rappuoli 2004). The yeast genome was completed one year later (Zhang 1999). Dolly the sheep, the first cloned animal, was also born at this time (Elster 1999).

In 1990, the Human Genome Project was launched to sequence 3 billion letters of the human genome. As part of this project, chromosome 22 was sequenced as the first chromosome in 1999. The project was finished in 2003, and it confirmed that humans have between 20,000 and 25,000 genes.

In 2007, there was a breakthrough in DNA sequencing technology that increased the output of DNA sequencing by 70 times in 1 year. This prompted the launch of the 1000 Gene Project in 2008, intending to sequence the genomes of a large population of 2500 people.

In general, advances in DNA technology have aided the development of omics. Figure 1.2 depicts a brief timeline of recent developments in multi-omics research.

1.2 Omics: DNA, RNA, Protein, and Microbiome

There are now many hot omics studies such as transcriptomics for RNA research, proteomics for protein research, and microbiomics for microbiome research (Yu et al. 2018). Transcriptomics is the study of gene transcription in cells as well as transcriptional regulation in general (Dong and Chen 2013). Proteomics is the study of the composition of cells, tissues, or biological proteins and their changing

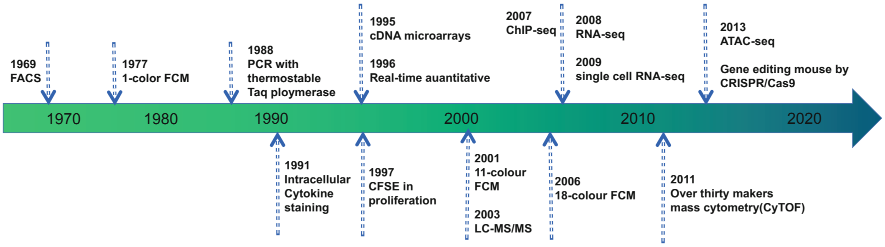

Fig. 1.2 The timeline for the development of multi-omics researches. The development history of multi-omics is actually the process of development and innovation of different omics data acquisition technologies. With the enrichment and cross-use of omics data, multi-omics analysis has gradually been applied

laws using protein as the research object. The term "microbiome" refers to the genomes of microorganisms (bacteria, archaea, lower or higher eukaryotes, and viruses) as well as their entire environment (Marchesi and Ravel 2015).

Multi-omics has emerged with the accumulation of various omics datas, and has become a research focus in recent years due to its importance in basic research and clinical application (Chakraborty et al. 2018). A series of disease-related differences are typically generated for omics data. These data can be used as disease process markers as well as insights into biological pathways or process differences between the disease and the control group. However, only one type of data analysis has limited relevance, primarily reflecting the reaction process rather than causality. The integration of various omics data types is typically used to clarify potential pathogenic changes or treatment targets that cause the disease, which can then be tested further. Multi-omics research, when compared to a single type of omics research, can better understand the basic information flow of diseases.

In recent years, sequencing technologies have generated a large amount of multi-omics data worldwide, but also brings many problems. First, because the growth rate of multivariate data was unimaginable ten years ago, large public databases have used cloud facilities to store these data. Secondly, the cost of generating multi-omics data has decreased rapidly, leading to a further increase in the amount of multi-omics data as well (NHGRI 2021).

1.3 Databases and Tools for Omics Studies

When confronted with multi-omics data, there is an increasing demand for computational methods that can rationally integrate and accurately analyze heterogeneous multi-omics data. To date, numerous databases and analytical tools have been developed to aid in the analysis of these multi-omics datasets (Tables 1.1 and 1.2).

Table 1.1 Representative databases for multi-omics research

Database	Functionality	Web link	Reference
ChEBI	Metabolomics database and ontology	http://bigd.big.ac.cn/databasecommons/database/id/364	Degtyarenko et al. (2007)
E.coli metabolome database (ECMDB)	Annotated metabolomics and metabolite pathway database	https://ecmdb.ca/	Guo et al. (2012)
FlyBase	Genes and RNA-seq data of different drosophila	https://flybase.org/	Thurmond et al. (2018)
GenBank (database)	Proteomics database open access annotated collection of all publically available nucleotide sequences and their protein transitions.	https://www.uniprot.org/database/DB-0028	Benson et al. (2017)
Human Metabolome Database (HMDB)	Human metabolite and pathway database	https://hmdb.ca/	Wishart et al. (2017)
KEGG	Collection of databases dealing with genomes biological pathways, disease, drugs and chemical substances	https://www.kegg.jp/	Kanehisa et al. (2016)

Table 1.2 Representative analytical tools for multi-omics research

Tool/Method	Tool/Method approach	Tool/Method link	Reference
PARADIGM	Probabilistic graphical models using directed factor graphs	http://paradigm.five3genomics.com/	Gluth et al. (2013)
iCluster	Joint latent variable model-based clustering method	https://cran.r-project.org/web/packages/iCluster/index.html	Shen et al. (2009)
iClusterPlus	Generalized linear regression for the formulation of the joint model	http://www.bioconductor.org/packages/release/bioc/html/iClusterPlus.html	Pierre-Jean et al. (2019)

1.4 Multi-Omics Applications

Multi-omics research has been successfully applied to many biological problems such as cancer omics, inflammatory disease omics and microbiome research.

The applications of multi-omics in disease (Hasin et al. 2017), including the integration of genome, epigenome, transcriptomics, proteomics, metabolomics and microbiome, as well as their interrelationships. Figure 1.3 depicts the various omics data types and disease research methods from an article. Each layer represents an omics data type. The omics data is gathered across the entire molecular pool, which is represented by circles. Except for the genome, all data layers reflect genetic regulation and the environment, and the environment's impact on each molecule may differ. In a recent work, cancer genomic profiling of 78 clinical tumor samples (Rusch et al. 2018) using three-platform sequencing of the whole genome, whole

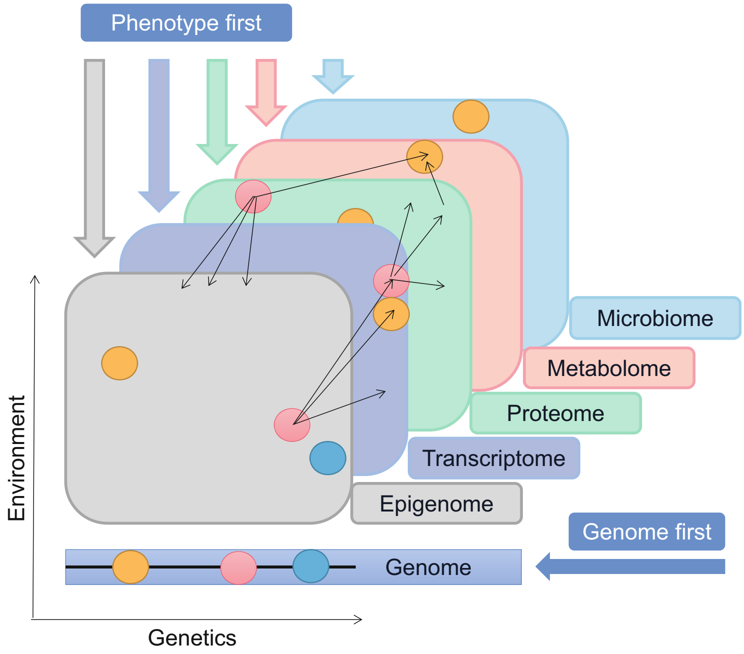

Fig. 1.3 Different omics data and corresponding analytical methods for disease research. Except for the genome, all data layers reflect both gene regulation and environment, which may affect each molecule to varying degrees. Potential interactions or correlations detected between molecules in different layers are represented by thin arrows

exome, and transcriptome to identify tumor-related structure variation (SV), somatic cell mutation, and pathogenic mutation, among other things. Sequencing, variant detection, variant classification, group review, and report generation are all covered in this the clinical three-platform sequencing design. In another research, multi-omics approaches to study secondary metabolites biosynthesis in microbes (Palazzotto and Weber 2018).

In general, multi-omics data resources are rapidly growing, and their analysis tools and platforms are maturing. Multi-omics research has made remarkable achievements in cancer and biological problems. Some applications of multi-omics research are listed below.

1. **Multi-omics approaches to cancer** (Aure et al. 2013) tracked genetic associations caused by breast cancer using complete genome-wide copy number and expression data. The author proposed a method for analyzing in-cis correlated genes in biological processes that is not biased towards particular types or functional processes. The goal of this method is to find cis-regulated genes whose expression correlation with other genes supports the role of network interference in cancer. This method was used to examine the genome-wide

copy number and expression data of 100 primary breast cancer patients. A total of 6373 gene abnormalities were discovered, 578 of which were highly in-cis correlated to biological processes and 56 of which were in-trans correlated. Among these in-trans process associated and cis-correlated (iPAC) genes, 28 percent had previously been linked to breast cancer, and 64 percent had previously been linked to cancer. The proposed method identified several known and new cancer driver candidates by combining the statistical evidence from three independent sub-analyses focusing on copy number, gene expression, and the combination of the two. The validation of independent data sets backs up the conclusion that this method identifies cancer-related genes.

2. **Multi-omics approaches to chronic kidney disease** (Husi et al. 2014). To gain a comprehensive understanding of key molecular changes in the vascular system caused by diabetes, researchers performed integrated proteomics and bioinformatics analysis on the aortic blood vessel data of a low-dose streptozotocin-induced diabetic mouse model (10) The researchers discovered significant dysregulation of molecules involved in myogenesis, angiogenesis, hypertension, hypertrophy (associated with aortic wall thickening), and a significant decrease in fatty acid storage. Another novel discovery is the significant down-regulation of glycogen synthase kinase-3 (Gsk3) and the up-regulation of molecules associated with the tricarboxylic acid cycle (for example, aspartate aminotransferase [Got2] and hydroxy-oxy-acid transhydrogenase). Furthermore, the pathways involved in the breakdown of primary alcohols and amino acids have been altered, potentially leading to the formation of ketone bodies.
3. **The NASA twins study: a multidimensional analysis of a year-long human spaceflight** (Garrett-Bakelman et al. 2019; Jerrusalem 2015) investigated the biology that may change humans during long-term space travel. When NASA astronaut Scott Kelly launches on a year-long mission to the International Space Station, researchers will collect and compare genomic, molecular, and physiological data from Scott and his twin brother, former astronaut Mark Kelly. Data changes may reveal how the human body reacts to extreme environments. The multi-omics longitudinal analysis process for twins is depicted in Fig. 1.4. According to NASA, the study was divided into four parts. (1) Human physiology: How does the space environment change organs such as the heart, muscles and brain? (2) Behavioral health: The changes in astronauts' cognitive reasoning, decision-making and alertness in the space environment. (3) Microbiology: The study examines dietary differences and stressors in twins and how they affect their gut flora. (4) Molecular/omics: Research in this area will focus on possible changes in gene expression caused by the space environment, as well as the effects of radiation, claustrophobic conditions, and microgravity on proteins and metabolites.

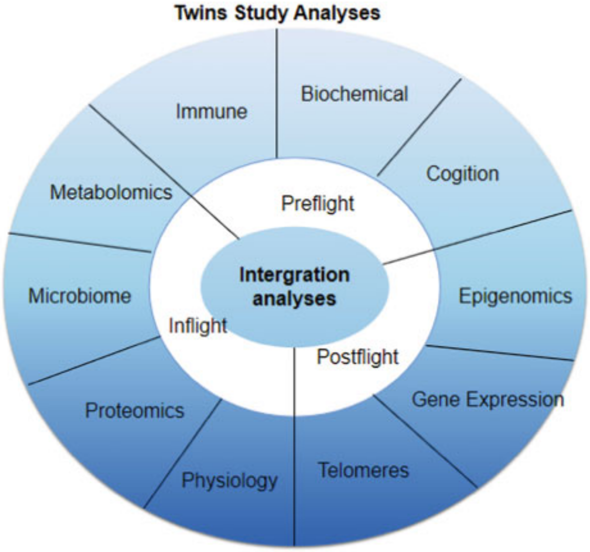

Fig. 1.4 Multidimensional, longitudinal assays of the NASA Twins Study. The twins collected and analyzed the characteristics of 10 generalized biomedical models before the flight (in flight), during flight (in flight), and after flight (after flight) for a total of 25 months

1.5 Future Perspectives

With more easily accessible data, as well as more mature databases and analytical tools for multi-omics data, multi-omics studies are expected to reveal more profound omics patterns that regulate the phenotypes of objectives of interest. Multi-omics studies have been widely applied to many biological problems in recent years. For example, many advances have been made in the epidemiology of chronic diseases (Pang et al. 2021). Zhou et al. conducted a longitudinal multi-omics study on the host-microbial dynamics in the pre-diabetes stage, gaining a better understanding of the multi-omics characteristics of this early stage (Zhou et al. 2019). Fiorenza Schussler Rose SM et al. used in-depth multi-omics measurement to identify clinically relevant T2D molecular pathways, investigated the ability of in-depth longitudinal analysis in health-related discoveries, identification of clinically relevant molecular pathways, and behavior influence, and employed omics method development. Using an insulin resistance predictive model, the findings show that in-depth longitudinal analysis can lead to actionable health findings and provide relevant information for accurate health (Schüssler-Fiorenza Rose et al. 2019). Liu et al. combined epidemiology, pharmacology, genetics, and gut microbiome data in the drug metabolite map, and the results demonstrated that the cross-sectional study of the effect of statins on metabolites is estimated and intervention and genetic observation research comparable. Proton pump inhibitors are linked to circulating metabolites, liver function, liver steatosis, and the gut microbiome in additional data integration. The research map serves as a tool for targeted experimental drug research and clinical trials aimed at improving drug efficacy, safety, and reuse (Liu et al. 2020). In general, single-omics research lacks multi-level integration, and the utility of inferring the etiology of complex diseases is limited. Multi-omics research significantly broadens the scope of etiology research. The previously difficult problem of obtaining pathological samples of multi-omics data has gradually been solved due to the rapid growth of multi-omics database resources. Multi-omics offers new ideas for traditional observational epidemiological research to infer the cause of chronic diseases, as well as valuable resources in the integration of systemic epidemiology to investigate disease mechanisms, and will serve as an important reference for subsequent further experimental verification studies (Pang et al. 2021).

However, the maintenance of long-term follow-up and laboratory testing is more expensive when obtaining multi-omics data related to the disease. Pathological samples of rare diseases are difficult to obtain in clinical practice, posing significant challenges to disease-related molecular biology research. Furthermore, the combined effects of multiple factors, as well as the high variability of a single data set, can lead to false discoveries, making it difficult to interpret the results of multi-omics analysis, particularly when identifying biologically related molecules (Pang et al. 2021). In general, while multi-omics research is progressing well, there is still a long way to go before we have a complete understanding of the holistic and dynamic pattern of how genetic materials regulate the phenotypes of objectives of interest.

References

Antonelli L, et al. Integrating imaging and omics data: a review. Biomed Signal Process Control. 2019;52:264–80.

Aure MR, et al. Identifying in-trans process associated genes in breast cancer by integrated analysis of copy number and expression data. PLoS One. 2013;8(1):e53014.

Benson DA, et al. GenBank. Nucleic Acids Res. 2017;46(D1):D41–7.

Casadesús J, Noyer-Weidner M. Epigenetics. In: Maloy S, Hughes K, editors. Brenner's encyclopedia of genetics (second edition). San Diego: Academic; 2013. p. 500–3.

Chakraborty S, et al. Onco-Multi-OMICS approach: a new frontier in cancer research. Biomed. Res. Int. 2018;2018:9836256.

Chung R-H, Kang C-Y. A multi-omics data simulator for complex disease studies and its application to evaluate multi-omics data analysis methods for disease classification. GigaScience. 2019;8:5.

Degtyarenko K, et al. ChEBI: a database and ontology for chemical entities of biological interest. Nucleic Acids Res. 2007;36(suppl_1):D344–50.

Dong Z, Chen Y. Transcriptomics: advances and approaches. Sci China Life Sci. 2013;56(10): 960–7.

Elster NR. Who is the parent in cloning? Hofstra. Law. Rev. 1999;27(3):533–55.

Fraser CM, Rappuoli R. Application of microbial genomic science to advanced therapeutics. Annu Rev Med. 2004;56(1):459–74.

García-Quesada A, et al. Seroprevalence and prevalence of Babesia vogeli in clinically healthy dogs and their ticks in Costa Rica. Parasit. Vectors. 2021;14(1):468.

Garrett-Bakelman FE, et al. The NASA twins study: a multidimensional analysis of a year-long human spaceflight. Science. 2019;364

Gluth S, Rieskamp J, Büchel C. Deciding not to decide: computational and neural evidence for hidden behavior in sequential choice. PLoS Comput. Biol. 2013;9(10):e1003309.

Guo AC, et al. ECMDB: the E. coli metabolome database. Nucleic Acids Res. 2012;41(D1): D625–30.

Hasin Y, Seldin M, Lusis A. Multi-omics approaches to disease. Genome. Biol. 2017;18(1):83.

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

Husi H, et al. Proteome-based systems biology analysis of the diabetic mouse aorta reveals major changes in fatty acid biosynthesis as potential hallmark in diabetes mellitus-associated vascular disease. Circ Cardiovasc Genet. 2014;7(2):161–70.

Jerrusalem. NASA, 2015. https://www.guokr.com/article/440100/.

Kanehisa M, et al. KEGG: new perspectives on genomes, pathways, diseases and drugs. Nucleic Acids Res. 2016;45(D1):D353–61.

Kumar, P.S., Microbiomics: Were we all wrong before? Periodontol. 2000, 2021, 85(1):8–11.

Liu J, et al. Integration of epidemiologic, pharmacologic, genetic and gut microbiome data in a drug–metabolite atlas. Nat Med. 2020;26(1):110–7.

Loizou GD. Animal-free chemical safety assessment. Front. Pharmacol. 2016;7:218.

Manzoni C, et al. Genome, transcriptome and proteome: the rise of omics data and their integration in biomedical sciences. Brief Bioinform. 2018;19(2):286–302.

Marchesi JR, Ravel J. The vocabulary of microbiome research: a proposal. Microbiome. 2015;3(1): 31.

(NHGRI), T.N.H.G.R.I. 2021. https://www.genome.gov/about-genomics/fact-sheets.

Osier ND, et al. Symptom science: repurposing existing omics data. Biol Res Nurs. 2017;19(1): 18–27.

Palazzotto E, Weber T. Omics and multi-omics approaches to study the biosynthesis of secondary metabolites in microorganisms. Curr. Opin. Microbiol. 2018;45:109–16.

Pang YJ, et al. A multi-omics approach to investigate the etiology of non-communicable diseases: recent advance and applications. Zhonghua Liu Xing Bing Xue Za Zhi. 2021;42(1):1–9.

Pierre-Jean M, et al. Clustering and variable selection evaluation of 13 unsupervised methods for multi-omics data integration. Brief Bioinform. 2019;21(6):2011–30.

Portin P. The birth and development of the DNA theory of inheritance: sixty years since the discovery of the structure of DNA. J. Genet. 2014;93(1):293–302.

Rusch M, et al. Clinical cancer genomic profiling by three-platform sequencing of whole genome, whole exome and transcriptome. Nat. Commun. 2018;9(1):3962.

Schüssler-Fiorenza Rose SM, et al. A longitudinal big data approach for precision health. Nat Med. 2019;25(5):792–804.

Shen R, Olshen AB, Ladanyi M. Integrative clustering of multiple genomic data types using a joint latent variable model with application to breast and lung cancer subtype analysis. Bioinformatics. 2009;25(22):2906–12.

Sriram RD, Subrahmanian E. Transforming Health Care through Digital Revolutions. J. Indian. Inst. Sci. 2020;100(4):753–72.

Sun YV, Hu Y-J. Chapter three—integrative analysis of multi-omics data for discovery and functional studies of complex human diseases. In: Friedmann T, Dunlap JC, Goodwin SF, editors. Advances in genetics. Academic; 2016. p. 147–90.

Thurmond J, et al. FlyBase 2.0: the next generation. Nucleic Acids Res. 2018;47(D1):D759–65.

Wishart DS, et al. HMDB 4.0: the human metabolome database for 2018. Nucleic Acids Res. 2017;46(D1):D608–17.

Yu G, Ibarra GH, Kaminski N. Fibrosis: lessons from OMICS analyses of the human lung. Matrix Biol. 2018;68-69:422–34.

Zhang MQ. Promoter analysis of co-regulated genes in the yeast genome. Comput Chem. 1999;23 (3):233–50.

Zhou W, et al. Longitudinal multi-omics of host–microbe dynamics in prediabetes. Nature. 2019;569(7758):663–71.

Part I
Omics Integration Techniques

Chapter 2
Biomedical Applications: The Need for Multi-Omics

Yuxue Li and Kang Ning

Multi-omics studies are urgently required for biomedical applications, not only because of the comprehensiveness of the omics that such multi-omics studies could consider but also because of the interconnections that different omics data would reveal towards a more in-depth understanding of the molecular biology behind living cells and/or communities of lives. The multi-omics strategy for biomedical applications is essentially a big data strategy. In the following sections, we will discuss (1) biomedical big data and challenges, (2) analytical techniques for biomedical multi-omics big data, particularly deep learning, (3) representative databases and tools for multi-omics data analysis, and (4) representative applications based on multi-omics data.

2.1 Biomedical Big Data and Challenges

Different from the traditional research paradigm, big data research is based on a statistical research, in which comparison, clustering, correlation analysis, classification analysis, and induction of a large number of data. For decades, the goal of experiments at the level of molecular biology has been to obtain conclusions or propose a new hypothesis in terms of biomedical big data and big data research is gradually shifting from hypothesis-driven to data-driven (Casanova et al. 2022). Now, we can explore the rules of massive data research, directly put forward hypotheses, and draw reliable conclusions based on massive biomedical big data.

Y. Li (✉) · K. Ning
Key Laboratory of Molecular Biophysics of the Ministry of Education, Hubei Key Laboratory of Bioinformatics and Molecular-imaging, Center of Artificial Intelligence Biology, Department of Bioinformatics and Systems Biology, College of Life Science and Technology, Huazhong University of Science and Technology, Wuhan, China
e-mail: liyuxue@hust.edu.cn

K. Ning (ed.), *Methodologies of Multi-Omics Data Integration and Data Mining*,
Translational Bioinformatics 19, https://doi.org/10.1007/978-981-19-8210-1_2

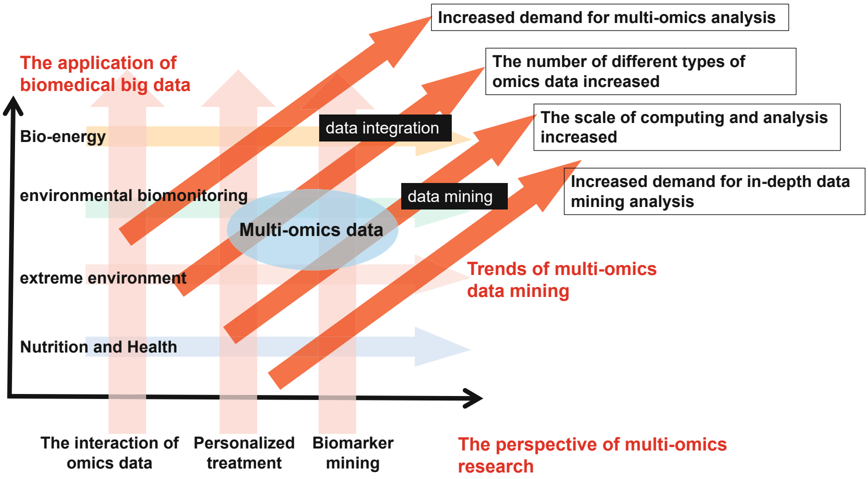
Increased demand for multi-omics analysis
The number of different types of omics data increased
The scale of computing and analysis increased
Increased demand for in-depth data mining analysis
The application of biomedical big data
Bio-energy
environmental biomonitoring
extreme environment
Nutrition and Health
data integration
data mining
Multi-omics data
Trends of multi-omics data mining
The interaction of omics data
Personalized treatment
Biomarker mining
The perspective of multi-omics research

Fig. 2.1 The properties of biomedical big data, and trends of biomedical big data mining. Biological big data analysis faces challenges: accuracy and speed, as the basic challenges inherited from big-data analytical tasks

However, the biomedical big data will be a double-edged sword. On the one hand, the difference of data will form a bottleneck in data integration at the same time of high-speed accumulation of big data. On the other hand, once this series of bottlenecks is broken, the profound biological laws contained in big data will greatly promote the understanding of human health.

Biomedical big data has unique properties, as represented by 3 V characteristics (Fig. 2.1). First, the volume of biomedical data is enormous. Second, the research has stringent requirements for the accuracy and speed with which results are processed. Third, the relevant source data are diverse and highly heterogeneous (Ning and Chen 2015). The needs for big data analysis is also high, which could be represented by 5H needs: High speed, High accuracy, High heterogeneity, High demand, High frequency of updates (Fig. 2.1). To be more specific, biomedical big data has high-dimensional characteristics in terms of multiple sample analysis angles, multiple omics data, and multiple sample sizes, necessitating the superposition, indexing, and learning of multi-dimensional data. What more, biomedical research's goals and processes are complex. Finally, differences in sample sources, processing methods, and storage formats in biomedical research result in a high degree of uncertainty and inconsistency of the research objects, necessitating intelligent data models for in-depth analysis (Ning and Chen 2015).

Among all omics data, microbiome big data has unique properties: heterogeneous biome from which samples are collected, highly incomplete reference, largely insufficient meta-data. Microbiome biological big-data analysis faces challenges:

dynamic pattern mining, and functional gene mining, as the bottleneck for data understanding.

2.2 Deep Learning for Biomedical Big Data

Deep learning techniques have been widely used in big data analysis (Zhang et al. 2017), including biomedical big data analysis (Krassowski et al. 2020). Since biological big data could be categorized into sequence data and feature-rich data in other forms (Pal et al. 2020), deep learning methods can be classified as sequence-based or feature-based.

Sequence-based deep learning methods typically transform sequences into feature space before data mining (Li et al. 2019). One-hot is a method of converting a biological sequence into a binary sequence based on a coding scheme. For example, A, C, G, and T could be coded as 0001, 0010, 0100, and 1000, respectively. However, one-hot is not only redundant in information but also introduces noise into data mining, it is not widely used today. The second method is to extract k-mer features directly from sequences, which is an efficient method but has poor interpretability. The third approach, which has been successful in a wide range of applications (Li et al. 2019), is based on feature extraction from sequences using domain knowledge.

Fetures-based deep learning methods are adaptable for a wide range of applications, including functional element mining (Nahmias et al. 2020), gene or species distribution pattern mining, and evolutionary relationship inference (Sadeqi Azer et al. 2020), among others. For example, in functional element mining, feature-based deep learning methods typically work as follows: feature extraction and alignment, modeling construction, and prediction. Another example is AlphaFold for protein 3D structure prediction, which considers residue-residue distances as features and trains the deep learning model to best fit for these features (Poplin et al. 2018).

Deep learning algorithms have emerged in many fields, including text, medical care, and finance, and have shown great promise. In the field of gene sequencing, there have also been numerous application breakthroughs. DeepVariant (Poplin et al. 2018), a deep learning-based mutation detection software developed by Google, was the first to use artificial intelligence technology to detect mutations and decode genetic data. As we known, sequencing errors are unavoidable in the process of high-throughput sequencing technology. Mutation detection, as a link in the gene sequencing process, is designed to prevent various errors accumulated during the sequencing process from generating false negative and false positive information in the final mutation information, thereby affecting the accuracy of the sequencing report. The standard mutation detection tool at the moment is GATK's Haplotype Caller (Happ et al. 2019). The detection accuracy of single nucleotide polymorphism variation on commonly used 30X deep whole genome (WGS) sequencing data can reach around 99.7% (Poplin et al. 2018).

Google as a pioneer in the field of artificial intelligence, recently launched DeepVarian, a mutation detection software based on a deep learning algorithm, which aims to improve the accuracy of mutation detection by using the rapidly developing artificial intelligence technology. Advantages of DeepVariant include more complex classification features and the ability to propose effective measures for specific problems.

The Process of DeepVariant can be divided into three steps:

1. The input data required by DeepVariant can be used to obtain BAM data by GATK or other mature processes to conduct quality control, comparison, sorting, deduplication and other operations on sequencing data.
2. The potential mutation sites are retrieved and screened from the BAM data, and the sequencing data near these points are expanded and spliced into a pileup image.
3. The high-depth convolutional neural network with pre-training is used for image recognition and classification of Pileup images, and the variation information of target sites, such as various types and qualities, is provided.

DeepVariant has mutation detection accuracy far outperforms state-of-the-art methods. Various deep learning algorithms will redefine mutation detection and even the standards of each link in the bioinformatics analysis process in the future, catapulting the bioinformatics industry to new heights (Poplin et al. 2018).

2.2.1 Application of Functional Gene Mining

Gene annotation is based on "homology equals functional similarity", using bioinformatics methods to search and compare the similarity of unknown gene sequences in public databases. Homology with annotated genes in the database can be used to infer the function of unknown genes. GenBank (NCBI), EMBL, and DDBJ are the most well-known annotated nucleic acid databases. SwissProt and TrEMBL are two of the most popular protein databases. BLAST, FASTA, and other search and comparison software are commonly used.

2.2.1.1 Using Comparisons and Annotations to Discover Genes

Unigene is created by processing and splicing a large number of EST sequences obtained through sequencing. Candidate genes with reference annotation function can be obtained or new genes discovered through comparison and annotation with multiple public databases using software such as BLAST. This method is most commonly used to find genes in species with known genomic information or clear metabolic pathways (Yang et al. 2020).

2.2.1.2 Using Expression Differences to Discover Genes

Transcriptome sequencing can be used to compare differentially expressed genes in non-model species that lack reference genomes and have weaker molecular basic research. Conduct cluster analysis on these differentially expressed genes, group genes with similar functions together, and determine the functions of unknown genes grouped into one class via known functional genes.

High-throughput transcriptome sequencing technology is still in its infancy, but it provides a new platform and a huge development opportunity for molecular biology and transcriptome research, with obvious benefits. Exploration of genes using transcriptome sequencing technology will greatly enrich the gene resources of eukaryotes, particularly non-model organisms, and promote the development of molecular breeding.

Machine learning methods, such as deep learning, have very promising prospects for discovering hidden structures and making accurate predictions based on large amounts of data. The author introduced the background of deep learning and its successful application in biological problems in a review. In addition to presenting specific applications and providing practical tips (Angermueller et al. 2016). Figure 2.2 is a typical learning case process in machine learning.

2.2.2 Protein Structure Prediction

A protein's function frequently depends on its unique three-dimensional structure, and predicting the structure of proteins is critical to understanding its role in the body as well as diagnosing and treating diseases thought to be caused by misfolded proteins. Alzheimer's, Parkinson's, Huntington's, and cystic fibrosis are a few examples (Callaway 2020). However, predicting a protein's three-dimensional structure from a limited gene sequence is more difficult because it is difficult to predict how long a chain of amino acid residues will fold into a protein with a complex three-dimensional structure. This is referred to as the "protein folding problem" (Strodel 2021).

Scientists have developed experimental techniques to determine the structure of proteins over the last 50 years, such as cryo-electron microscopy, nuclear magnetic resonance, and X-ray crystallography, but these methods rely heavily on trial and error and cost a lot of money and time to identify a structure. So, biologists are turning to artificial intelligence methods to replace this time-consuming and laborious protein processing process (Callaway 2020).

Protein structure prediction methods:

1. Comparative homology modeling: to construct protein model according to the alignment information between protein sequence and already structured proteins (Webb and Sali et al. 2016).

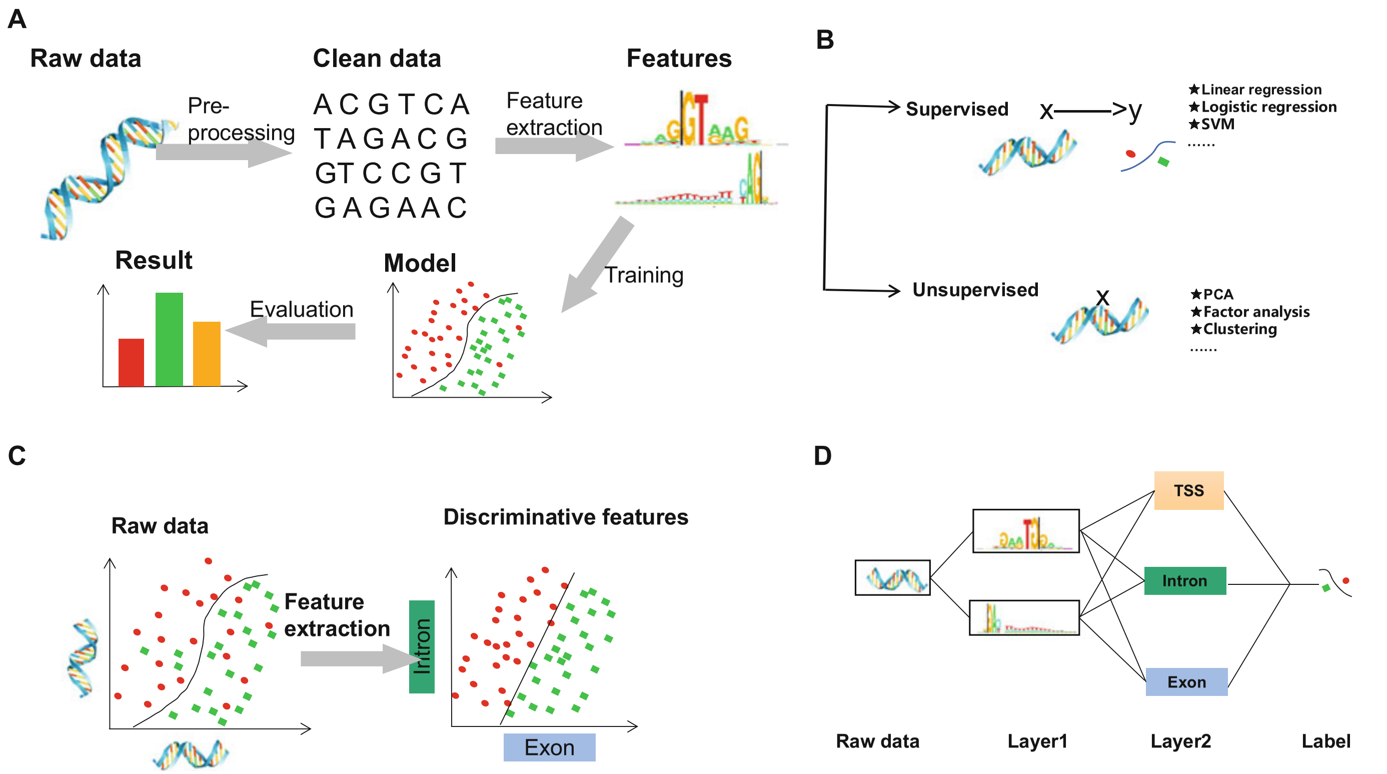

Fig. 2.2 Machine learning and typical learning contexts. (**a**) The traditional machine learning workflow consists of four steps: data preprocessing, feature extraction, model learning, and model evaluation. (**b**) The supervised machine learning method associates the input feature x with the output label y, whereas the unsupervised machine learning method learns factors about x without looking at the label. (**c**) Extract features from high-dimensional data for classification. (**d**) Deep networks learn increasingly abstract feature representations from raw data based on the hierarchical structure

2. Threading fold recognition: to find the most suitable template for the unknown protein, sequence and structure comparison is carried. Finally, the structure model is established (Buchan and Jones 2017).
3. Ab initio /de novo methods: to predict protein structure from scratch from the sequence itself (Guo et al. 2008).

Deep learning methods that rely on genetic data to predict problems have grown in popularity in recent years. With the rapid reduction in the cost of gene sequencing, a large amount of genomics data has been accumulated. Deep learning methods that rely on genetic data to predict problems have grown in popularity. DeepMind's efforts on this front resulted in the AlphaFold (Serpell et al. 2021), which CASP organizers hailed as an "unprecedented advance in the ability of computational methods to predict protein structure." AlphaFold solves the problem of creating target structures from scratch without the use of previously solved proteins as templates (Callaway 2020). The basic principle and process of protein structure prediction are depicted in Fig. 2.3.

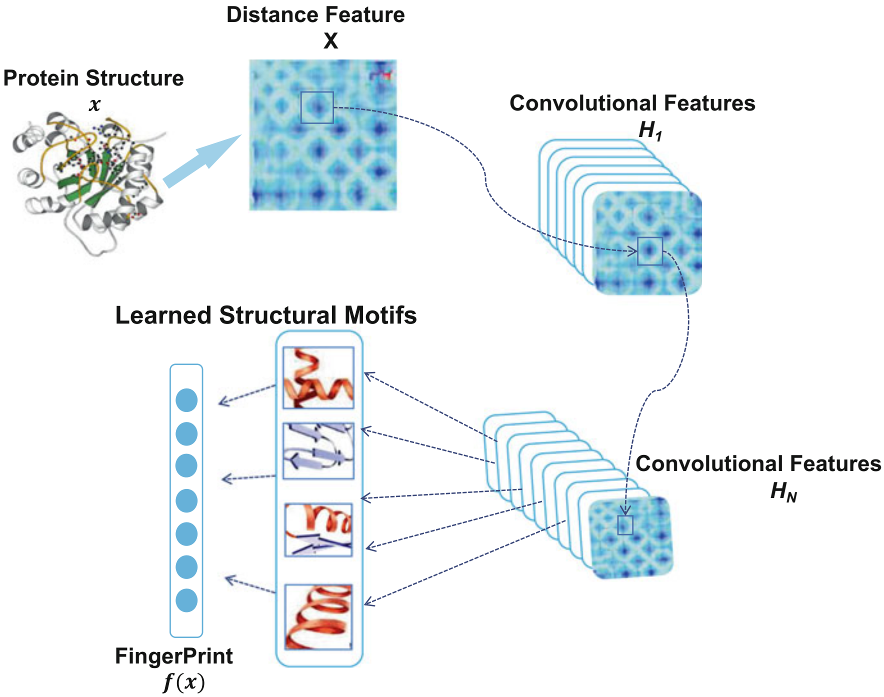

Fig. 2.3 The process of protein structure prediction. Protein sequence features are first transformed to residue-residue distance matrix, and then processed through convolution and feature selection processes, before the final structural motif and protein structures are obtained

2.3 Representative Databases and Analytical Tools

Faced with multi-omics data, there is an increasing demand for computational methods that can rationally integrate and accurately analyze heterogeneous multi-omics data. Until now, numerous databases have been created to aid in the analysis of these multi-omics datasets (Table 2.1).

One of the most prominent databases for multi-omics studies is the TCGA database and platform (Li et al. 2020). Based on paramount multi-omics data, various analytical tools have also been developed (Table 2.2).

Table 2.1 Representative databases for multi-omics researches

Tools	Types	Web Link	Reference
The Cancer Genome Atlas (TCGA)	RNA-Seq, DNA-Seq. miRNA-Seq, SNV, CNV, DNA methylation, RPPA	https://cancergenome.nih.gov/	Wang et al. (2016); Danaher et al. (2018)
Clinical Proteomic Tumor Analysis Consortium (CPTAC)	Proteomics data corresponding to TCGA cohorts	https://cptac-data-portal.georgetown.edu/cptaPublic/	Rudnick et al. (2016)
International Cancer Genomics Consortium (ICGC)	Whole-genome sequencing, genomics variation data (somatic and germline mutation)	https://icgc.org/	Li et al. (2020)
Cancer Cell Line Encyclopedia (CCLE)	Gene expression, copy number, and sequencing data; pharmacological profiles of 24 anticancer drugs	https://portals.broadinstitute.org/ccle	Ghandi et al. (2019); Nusinow et al. (2020)
Molecular Taxonomy of Breast Cancer International Consortium (METABRIC)	Clinical traits, gene expression, SNP and CNV	https://molonc.bccrc.ca/aparicio-lab/research/metabric/	Heng et al. (2017)

Table 2.2 Representative analytical tools for multi-omics researches

Tool/ Method	Tool/Method approach	Tool/Method Link	Reference
PARADIGM	Probabilistic graphical models using directed factor graphs	http://paradigm.five3genomics.com/	Gluth et al. (2013)
iCluster	Joint latent variable model-based clustering method	https://cran.r-project.org/web/packages/iCluster/index.html	Shen et al. (2009)
iClusterPlus	Generalized linear regression for the formulation of the joint model	http://www.bioconductor.org/packages/release/bioc/html/iClusterPlus.html	Pierre-Jean et al. (2019)
LRAcluster	Probabilistic The model with low-rank approximation	http://lifeome.net/software/lracluster/	LRAcluster (2008)

2.4 Representative Applications Based on Multi-Omics Big Data

The general approaches for multi-omics studies on biomedical applications follow the “multi-omics data integration -- > multi-omics data mining -- > biomedical pattern inference and validation” paradigm (Subramanian et al. 2020). Following this paradigm, many biomedical applications have been conducted based on multi-omics data, which in turn urged for more dependencies on multi-omics data by biomedical applications (Fig. 2.4) (Chakraborty et al. 2018; Khan and Azmir 2020).

Cancer poses a great threat to human health and is the focus of much biomedical research. The development of multi-omics has promoted cancer research, and cancer analysis has become one prominent example of multi-omics study (Khan and Azmir 2020).

Complementary Information

Cancer: Multi-omic analysis has a huge impact on the field of cancer mapping analysis, diagnosis and treatment. However, different types of mutations in cancers complicate analysis, and require specific treatment. In addition to the technical challenge of identifying somatic variation, most of the apparent genetic changes in

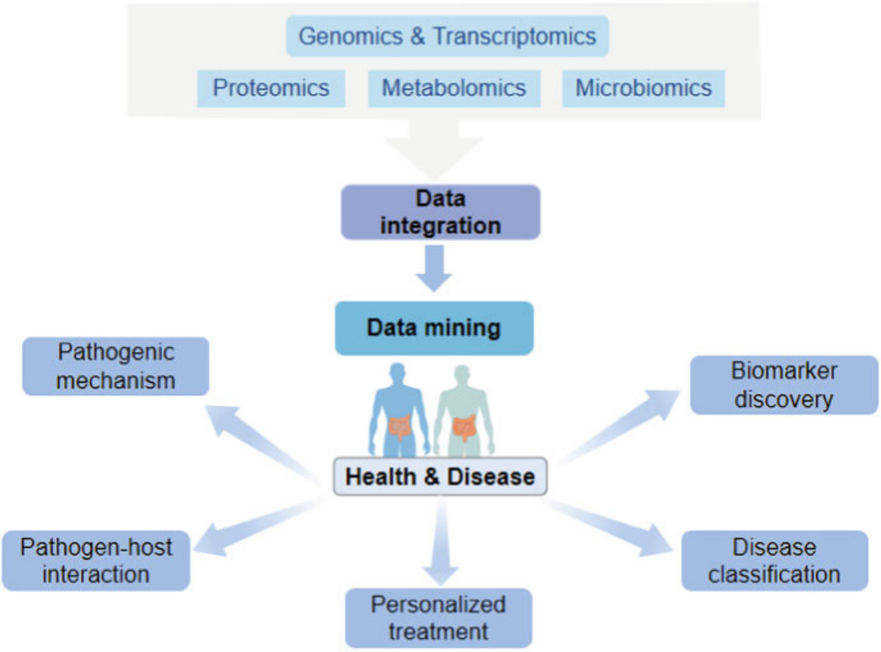

Fig. 2.4 Representative applications based on multi-omics approaches. Based on the integration of multi-omics data, a more comprehensive understanding of the molecular characteristics of diseases can be obtained, and promote the development of pathological research, marker discovery, etc.

cancer cases are benign and do not promote cancer cell growth. As a result, determining which mutations are driving the disease or which pathways are involved remains a significant challenge (Karczewski and Snyder 2018)

Identifying Driver Mutations: Whole-genome sequencing (WGS) of multiple tumors to identify mutated genes in common is a common method for identifying driver mutations. Because driver mutations are more likely to be present in genes expressed in specific cancers, adding functional data helps prioritize the likelihood of driver genes for these genes. In this way, combining additional multi-omics data with genetic data analysis provides a mechanism for filtering out a large number of genetic variations and eventually obtaining functionally related driving variants (Karczewski and Snyder 2018)

Molecular Signatures of Cancer: Multi-omics data can reveal biochemical pathways active in cancer and classify them into different subtypes, in addition to identifying driver mutations. As a result, it is a useful tool for determining which pathways are targeted in patients, even if strong candidate mutations (such as difficult-to-characterize non-coding mutations or indirect effects) are not found in these pathways (Karczewski and Snyder 2018)

2.4.1 Microbiome Mining for Cancer Research

Unknowns surround the microbiome and cancer. According to a widely quoted statistic, infectious pathogens cause 20% of all cancers worldwide. This may explain why microbes are often regarded as carcinogens that must be eradicated in cancer biology (Xavier et al. 2020).

In 2019, the IUCN stated that “there is no direct evidence that human symbiotic microbes are a key determinant of cancer pathogenesis”, it is critical to distinguish between the microbiome’s direct and indirect effects on cancer. When microbes come into direct contact with cancerous tissue and affect its performance, direct effects may occur. When the microbiome affects tumors at the distal end, it may have an indirect effect, as in the case of the intestinal microbiome, or it may influence cancer progression elsewhere by influencing host physiology or triggering systemic inflammation. In general, the human microbiome can interact with cancer through complex feedback loops. The effect of the microbiome on cancer may be direct or indirect. Direct interactions between the microbiota can occur in tissues where cancer arises, such as the skin microbiota described here directly interacting with melanoma; indirect interactions can occur between the microbiota and cancer in different tissues. For example, the gut microbiota alters circulating metabolites, which in turn affect the general physiology of the host and may have indirect effects on skin cancer progression or its response (Xavier et al. 2020).

The microbiome’s impact on cancer can be direct or indirect, which is an important distinction. As illustrated in the following hypothetical case: The skin microbiome interacts directly with melanoma; indirect interactions between the microbiome present in different tissues and cancer are also possible. The gut microbiome, for example, may influence host physiology by altering circulating metabolites, thereby indirectly influencing skin cancer progression or the host's response to treatment. Furthermore, diet may play a role, as it influences metabolite circulating levels and microbiome composition.

Viruses play a key role in our understanding of cancer as an inherited disease, and those that have been found to cause cancer in humans are further evidence that the link between microbes and cancer may help prevent it. Although there is a link between viruses and cancer, most microbiome research focuses on bacteria rather than viruses. Surprisingly, other microbes in the microbiome may also influence the progression of virus-caused cancers, which is significant because the majority of people infected with cancer-causing viruses do not develop cancer, and the factors that determine how viral infections develop are unknown.

Cervical squamous epithelial lesions and HPV infection significantly altered the composition of bacteria and fungi in a study of Puerto Rican women. Thus, changes in the cervicovaginal microbiome shifted from a Lactobacillus-dominated community to one dominated by strictly anaerobic bacteria (such as Sneathia sanguinigens and Gardnerella vaginallis), which colonizes the host-environment junction, may influence cervical cancer susceptibility (Xavier et al. 2020).

Dual Role of Microorganisms in Cancer

Helicobacter pylori, the most well-known cancer-causing bacterium, exemplifies a paradox: while it causes stomach cancer in humans, it is thought to be a normal component of the stomach microbiome, capable of spreading from person to person and co-evolving with humans. Furthermore, h. pylori colonization may be beneficial to human health because removing h. pylori increase the risk of many diseases, including severe gastroesophageal reflux disease and its sequelae, Barrett's esophagitis, and esophageal adenocarcinoma. Although it is a risk factor for diabetes and other diseases, it can also prevent asthma, multiple sclerosis, and inflammatory bowel disease (IBD). As a result, Helicobacter pylori may be both harmful and beneficial to human health. This duality also applies to other microbes. The type of inflammation caused by microbes may be the difference between good and bad. The human immune system is in balance, and an appropriate immune response is required for homeostasis.

Emerging Technology to Study the Role of Microbiota in Cancer

To investigate the role of the microbiome in cancer, it is necessary to distinguish between direct and indirect effects, as they necessitate different approaches and even different techniques. These studies may necessitate sophisticated in-situ imaging techniques to preserve the spatial structure of directly affected microbes, which can directly contact cancer cells and affect their performance.

A systematic approach to the function of the microbiome is required to analyze the indirect effects of the microbiome. For studying microbiome functions, several bioinformatics tools, such as BugBase93 and PICRUSt2, are already available. Other tools, such as QIIME 2, to multiple sets of studies provide a platform for microbial groups, integration of different types of microbial groups of data, such as taxonomy data, metagenomic and metabolomic data, to support research and what

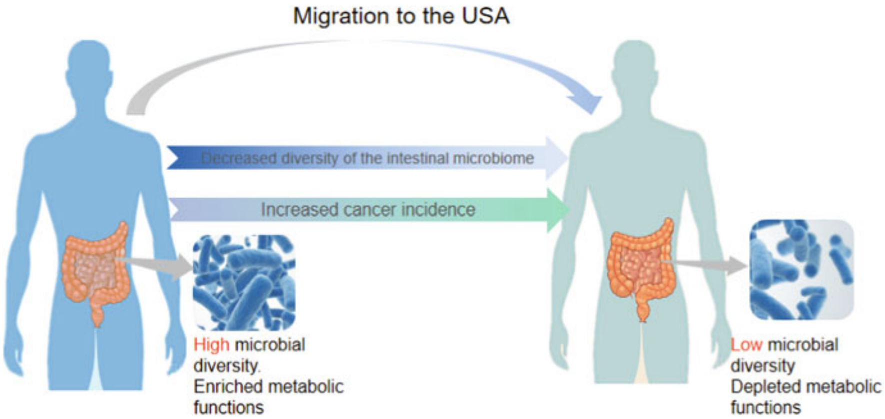
Migration to the USA
Decreased diversity of the intestinal microbiome
Increased cancer incidence
High microbial diversity
Enriched metabolic functions
Low microbial diversity
Depleted metabolic functions

Fig. 2.5 Migration to the United States reduces the diversity of intestinal microbiota. Accompanied with the host migrations to the USA, the lower microbial diversities, the depleted metabolic functions, as well as the increased cancer incidences also appear

microorganisms exist at the same time, their potential function and metabolic activity, can go beyond simply microbiome association studies, to study the mechanism of microbial groups.

Environmental Effects on Microbial Composition and its Role in the Pathogenesis of Cancer
A recent study of the microbiome of the immigrant population in the United States (Vangay et al. 2018) (Fig. 2.5) showed that the immigrant's gut microbiome changed after arrival in the United States: After migration, the native strains and functions of the immigrant's gut were immediately replaced by those typical of the American population. The longer immigrants stay in the United States, the more their microbiome changes. The change in the next generation was even greater, and importantly, it was associated with obesity. This suggests that migration can have an impact on microbiome diversity and further contribute to cancer-related health problems (such as obesity), which justifies the need for a microbiome library for remodel.

A recent study of the microbiome of the immigrant population in the United States (Fig. 2.5) revealed that the gut microbiome of the immigrant changed after arrival: The native strains and functions of the immigrant's gut were immediately replaced by those of the American population after migration. The more time immigrants spend in the United States, the more their microbiome changes. The change in the next generation was even greater, and it was linked to obesity. This suggests that migration can affect microbiome diversity and contribute to cancer-related health problems (such as obesity), justifying the need for a microbiome library for remodel.

This could explain the increased incidence of certain diseases, including obesity and some cancers, among certain groups of immigrants to the United States.

Studying this intriguing phenomenon could shed new light on the role of the microbiome in cancer.

2.4.2 *The Twin Astronauts*

NASA launched a genetic research project involving twin astronauts in 2017 to investigate the impact of the space environment on human genes. Ten research groups compared astronaut Scott Kelly's physiological data before, during, and after his mission to the space station to that of his twin brother, Mark Kelly, who is also an astronaut and is now retired and back on Earth. Based on the research, in April 2019, Science published an article titled "The NASA Twins Study: A Multidimensional Analysis of a Year-Long Human Spaceflight," which discussed the impact of a year in space from multiple perspectives.

At the genomic level, it was discovered that Scott and Mark each had hundreds of mutations, as well as over 200,000 RNA molecules that differed from those found in the normal population, after sequencing the DNA and RNA of white blood cells from the twins. Scott's DNA's chemical modification decreased during the flight and returned to normal immediately after landing. Mark's DNA chemistry changed halfway through the experiment, but it eventually returned to normal. The researchers believe that this demonstrates that genes are sensitive to environmental changes, whether in space or on Earth. Scott's telomeres grew during his time in space. When he returned to Earth, his telomeres had shrunk again. Telomeres typically shorten with age, which could imply "longer life," but researchers believe the change is likely due to Scott's increased exercise and decreased calorie intake while on the station. Telomerase, a reverse transcriptase that repairs and lengthens telomeres, was also studied in the twins. Both became more active in November 2015, possibly as a result that "major and intense family matter" that was taking place at the time. They also have different microbial communities, which is most likely due to the environment they live in and the food they eat. Scott's digestive tract was dominated by two major strains of bacteria once in space and once on Earth.

According to NASA's chief scientist John Charles human research projects (John Charles), the experiment, some results, such as the growth of the telomerase, require further studies to compare, verify whether be deceptive or transient changes, preliminary results show that the pressure of life in space for a year and there is no more in 6 months of life. More definitive results, including samples and supporting data, correlations between overall results, and genetic testing of astronauts on the current ISS mission and participants on future long-term missions, await further analysis.

2.4.3 Integrative Analysis of Genomics, Epigenomics, Transcriptomics

Genomics changes caused by DNA copy number aberrations or mutations frequently occur during tumorigenesis, promoting tumor development. Cancer genome epigenetic regulation via DNA methylation is also important in heterogeneous cancer behavior. Genome mapping studies, particularly in hepatocellular carcinoma (HCC), have revealed the enormous heterogeneity of genomic and epigenome disorders. The transcriptome disorder caused by these mutations is a major driving force in cancer progression. The limitations of using single omics data to analyze pathogenic factors have become increasingly apparent in recent years. When combined with multi-omics analysis, it allows for a more comprehensive understanding of tumors as well as the discovery of valuable tumor markers and related mechanisms. During tumor invasion, both DNA methylation and CNV occur, and both have an impact on transcription. It is unclear whether they have a synergistic effect.

The article "Integrative analysis of genomic and epigenomic regulation of the transcriptome in liver cancer" was published in Nature Communication in October 2017. The CNV, MET, and EXP samples from 64 liver cancer samples were chosen by the author. A difference analysis was performed using transcriptome data combined with genomic and epigenetic data to try to find abnormal genes or pathways that are closely related to the occurrence and progression of liver cancer.

The findings revealed that the abnormal frequencies of DNA copy number-related (CNVcor) and methylation-related expression (METcor) genes are significantly co-regulated. The integration of CNVcor and METcor genes in a multi-omics study revealed three prognostic subtypes of hepatocellular carcinoma, which can be confirmed by independent data. BAP1 mutations are common in the most aggressive subtypes expressing stem genes, indicating that it plays a role in aggressive tumor progression.

In conclusion, the study established a new research direction for the comprehensive examination of genome and epigenome regulation. As an in-depth understanding of the molecular biology principles underlying multi-omics data has already been obtained for biomedical applications, the "multi-omics data integration -- > multi-omics data mining -- > biomedical pattern inference and validation" paradigm has also evolved: more types of omics data are required for a holistic view of such patterns, as are more investigations of the associations between various omics data (Subramanian et al. 2020).

2.5 When Biocuration Meet Artificial Intelligence

2.5.1 The Current State of Biocuration

Biocuration, mostly done by biocurators, serve as the museum catalog of the Internet age. Biocuration has its data source from various biological databases and data described in numerous literature (Hirschman et al. 2016), and outputs high-quality annotation and curated data in standard format for scientists in various fields to further process (Bateman 2010). The biological administrator manages, collects, annotates, and verifies information disseminated through biological and model biological databases using computers (Bateman 2010).

This is not only about workload, but also about curation quality: as input data grows larger and the degree of heterogeneity increases, accurate manual curation will inevitably require more time, but time does not wait for biocurators. As a result, community biocuration has been proposed and demonstrated to be effective (Dauga 2015).

The development of machine learning approaches for omics data is a significant advancement in biocuration (König et al. 2018). Nowadays, it is difficult to find a biocurator who does all of her curations by hand. Machine learning tools may not only aid in more effective data pre-screening, but may also aid in the mining of previously overlooked valuable information from omics data (Ma and Zhang 2019). As a result, various automatic or semi-automatic methods for improving the efficiency and accuracy of biocuration have been proposed (Hu et al. 2004).

2.5.2 The Current State of Artificial Intelligence and its Application in Biocuration

Artificial Intelligence (AI) refers to the general or specific intelligence exhibited by computing machines or software (Hamet and Tremblay 2017). It focuses on central problems (or goals) including reasoning, knowledge, planning, learning, and perception (Russell and Norvig 2003).

AI has long been used in biocuration to increase efficiency and accuracy. Simple AI methods, such as K nearest neighbor (KNN), Support Vector Machine (SVM), Markov chain Monte Carlo (MCMC), and Neural Network (NN), were used on sequence analyses in the early years of biological data curation (Liao and Noble 2003), and they remain key methods for sequence annotations today, except human curation. More sophisticated AI for biocuration is based on more sophisticated computational models such as multi-layer MCMC and NN (Zobitz et al. 2011; Kim et al. 2019), new algorithms such as network analysis (DeGregory et al. 2018), and statistical methods such as PAML (Maldonado et al. 2016), for more diverse areas of biocuration including text mining (Hirschman et al. 2016), ontology analysis (Orchard and Hermjakob et al. 2015) and community curation (Dauga 2015).

However, AI has been in the stage of "weak AI" for decades, thus making current automatic or semiautomatic biocuration techniques also weak in AI. In recent years, "strong AI" has become more and more popular, which relies on multi-purpose and automatically-adjustable AI (Russell and Norvig 2003; Searle 1980). Similar to recent years' advancement of omics research based on multiple omics data, strong AI has its core model built based on a larger size of heterogeneous data, and this model is optimized with the accumulation of more data and becomes more accurate during this process (Gluth et al. 2013). One typical application of AI on biocuration is in biological text mining by means such as natural language processing. For example, the BioCreative initiative (Arighi et al. 2013) has been in smooth progress to tackle several biological text mining challenges using advanced information extraction methods. More recently, Google has made its text2vec text mining engine open-source (Johnson et al. 2019).

2.5.3 In Alliance Is the Trend

Automatic biocuration would become more powerful as a result of strong AI, and biocurators might be concerned about competition with them in big-data annotation and data mining. Actual debates have been going on for quite some time. However, they should be confident that manual curation has distinct advantages.

Firstly, current AI-based curations are still insufficient power for large-scale, effective, and predictive biocuration. Secondly, current computation hardware/software structures are rule-based, AI built based on the current computer system would follow rules. Thirdly, the ultimate goal of biological researchers is to help people. Making biocuration results understandable and readable by humans, including visualization, clinical interpretations, and public understanding, would thus require significant human intervention. Finally, and most importantly, because human intelligence and AI are based on different computation architectures (Searle 1980; Anderson 1964), a proper collaboration between the two would undoubtedly be beneficial. As a result, manual curation could be focused on more sophisticated, problem-oriented, and logic-intensive biocuration, while strong AI handles data collection and integration, batch processing, and result summary. This approach would be suitable for biological big-data research as well as improving the efficiency and accuracy of biocuration.

Current trends in biological research have also demonstrated that collaboration between biocurators and AI for better biological data curation is unavoidable. The collaboration of a biocurator and artificial intelligence would result in optimized biocuration that could best utilize biological big data for knowledge discovery.

Firstly, current grand projects may help us learn about biological intelligence, which we can then apply to biocuration and systems biology research. Furthermore, crowd-sourcing approaches for biocuration based on community power have been proposed (Hirschman et al. 2010). On the one hand, this would speed up biocuration, while on the other hand, it would provide more rules for strong AI to learn and even for AI systems to join for community biocuration. Finally, through the "computing everywhere" trend, the integration of hardware (hard disks, networks, etc.) in which databases and biocuration software or even strong AI systems live would be deeply integrated with the human (Mostéfaoui et al. 2008). Overall, while researchers in biological big data and bio-curators should continue to be cautious and patient in biological data curation, they should not be concerned about the future of their careers in biocuration.

Furthermore, we should keep in mind that the future of IT will not necessarily resemble human intelligence in terms of speed, accuracy, and data volume. Rather, the use of artificial intelligence to improve human intelligence is becoming more popular, particularly in bio-curation (Hirschman et al. 2010). However, decision-making, particularly those not based on a majority vote, would pose a significant challenge to IT development (New York Times report and human intelligence).

2.6 Conclusion

Multi-omics is an emerging field of omics research (Hasin et al. 2017). Some progress has been made in solving biomedical problems, but there are still many questions about the research of complex physiological processes such as human cells and diseases. The need of biomedical applications has prompted the development of multi-omics. Multi-omics platforms are considered to be the most complete system for obtaining and measuring biomedical data.

Cancer research, chronic kidney disease, infectious disease, and heart disease are all areas where multi-omics has been used. It provides more comprehensive biological information to answer the pathological process of the disease, allowing for in-depth study of the disease and the implementation of effective interventions.

A growing number of multi-omics studies have been published in recent years. The utility and benefit obtained through the multi-omics process cannot be obtained through any single omics method. The application of biomedicine will greatly promote the development of understanding the relationship between the molecular and clinical characteristics of diseases, and could naturally lead to the increasing development of tools and resources in the multi-omics platform.

References

Anderson AR. *Minds and machines*. Contemporary prospects in philosophy series, vol. viii. Englewood Cliffs, N.J: Prentice-Hall; 1964. p. 114.

Angermueller C, et al. Deep learning for computational biology. Mol. Syst. Biol. 2016;12(7): 878–878.

Arighi, C.N., et al., An overview of the BioCreative 2012 Workshop Track III: interactive text mining task. Database (Oxford), 2013. 2013: p. bas056.

Bateman A. Curators of the world unite: the International Society of Biocuration. Bioinformatics. 2010;26(8):991.

Buchan DWA, Jones DT. EigenTHREADER: analogous protein fold recognition by efficient contact map threading. Bioinformatics. 2017;33(17):2684–90.

Callaway E, '*It will change everything': DeepMind's AI makes gigantic leap in solving protein structures.* s, 2020.

Casanova R, et al. Comparing data-driven and hypothesis-driven MRI-based predictors of cognitive impairment in individuals from the atherosclerosis risk in communities (ARIC) study. Alzheimers Dement. 2022;18(4):561–71.

Chakraborty S, et al. Onco-multi-OMICS approach: a new frontier in cancer research. Biomed. Res. Int. 2018;2018:9836256.

Danaher P, et al. Pan-cancer adaptive immune resistance as defined by the tumor inflammation signature (TIS): results from the cancer genome atlas (TCGA). J. Immunother. Cancer. 2018;6 (1):63.

Dauga D. Biocuration: a new challenge for the tunicate community. Genesis. 2015;53(1):132–42.

DeGregory KW, et al. A review of machine learning in obesity. Obesity. Rev. 2018;19(5):668–685.

Ghandi M, et al. Next-generation characterization of the cancer cell line Encyclopedia. Nature. 2019;569(7757):503–8.

Gluth S, Rieskamp J, Büchel C. Deciding not to decide: computational and neural evidence for hidden behavior in sequential choice. PLoS. Comput. Biol. 2013;9(10):e1003309.

Guo JT, Ellrott Y, Fau-Xu K, and Xu Y. *A historical perspective of template-based protein structure prediction,*2008.

Hamet P, Tremblay J. Artificial intelligence in medicine. Metabolism. 2017;69:S36–40.

Happ MM, et al. Generating high density, low cost genotype data in soybean [Glycine max (L.) Merr.]. G3 (Bethesda, Md). 2019;9(7):2153–60.

Hasin Y, Seldin M, Lusis A. Multi-omics approaches to disease. Genome Biol. 2017;18(1):83.

Heng YJ, et al. The molecular basis of breast cancer pathological phenotypes. J. Pathol. 2017;241 (3):375–91.

Hirschman J, et al. A MOD(ern) perspective on literature curation. Mol. Gen. Genomics. 2010;283 (5):415–25.

Hirschman L, et al. Crowdsourcing and curation: perspectives from biology and natural language processing. Database. 2016;2016

Hu X, Zhang Z, Tao C. A robust method for semi-automatic extraction of road Centerlines using a piecewise parabolic model and Least Square template matching. Photogramm Eng Remote Sens. 2004;70:1393–8.

Johnson AJ, et al. Daily sampling reveals personalized diet-microbiome associations in humans. Cell Host Microbe. 2019;25(6):789–802.e5

Karczewski KJ, Snyder MP. Integrative omics for health and disease. Nat. Rev. Genet. 2018;19(5): 299–310.

Khan MS, Azmir J. Multi-omics for biomedical applications. J Appl Bioanal. 2020;6(3):97–106.

Kim B-H, Yu K, Lee PCW. Cancer classification of single-cell gene expression data by neural network. Bioinformatics. 2019;36(5):1360–6.

König C, et al. Using machine learning tools for protein database biocuration assistance. Sci. Rep. 2018;8(1):10148.

Krassowski M, et al. State of the field in multi-omics research: from computational needs to data mining and sharing. Front Genet. 2020;11:1598.

Li Y, et al. Deep learning in bioinformatics: introduction, application, and perspective in the big data era. Methods. 2019;166:4–21.

Li Y, et al. Patterns of somatic structural variation in human cancer genomes. Nature. 2020;578 (7793):112–21.

Liao L, Noble WS. Combining pairwise sequence similarity and support vector machines for detecting remote protein evolutionary and structural relationships. J Comput Biol. 2003;10(6): 857–68.

LRAcluster. 2008. http://lifeome.net/software/lracluster/.

Ma T, Zhang A. Integrate multi-omics data with biological interaction networks using multi-view factorization AutoEncoder (MAE). BMC Genomics. 2019;20(11):944.

Maldonado E, et al. LMAP: lightweight multigene analyses in PAML. BMC. Bioinformatics. 2016;17(1):354.

Mostéfaoui SK, Maamar Z, Giaglis GM. Advances in ubiquitous computing : future paradigms and directions, vol. xiii. Hershey PA: IGI Pub; 2008. p. 362.

Nahmias DO, Civillico EF, Kontson KL. Deep learning and feature based medication classifications from EEG in a large clinical data set. Sci Rep. 2020;10(1):14206.

Ning K, Chen T. Big data for biomedical research: current status and prospective. Chin Sci Bull (Chinese Version). 2015;60:534.

Nusinow DP, et al. Quantitative proteomics of the cancer cell line encyclopedia. Cell. 2020;180(2): 387–402.e16.

Orchard S and Hermjakob H, *Shared resources, shared costs—leveraging biocuration resources.* Database (Oxford), 2015. **2015**.

Pal S, et al. Big data in biology: the hope and present-day challenges in it. Gene Rep. 2020;21: 100869.

Pierre-Jean M, et al. Clustering and variable selection evaluation of 13 unsupervised methods for multi-omics data integration. Brief. Bioinform. 2019;21(6):2011–30.

Poplin R, et al. A universal SNP and small-indel variant caller using deep neural networks. Nat. Biotechnol. 2018;36(10):983–7.

Rudnick, P.A., et al., A description of the clinical proteomic tumor analysis consortium (CPTAC) common data analysis pipeline. J. Proteome. Res. 2016;15(3):p.

Russell SJ, Norvig P. Artificial intelligence: a modern approach (2nd ed.). Prentice Hall; 2003.

Sadeqi Azer E, et al. Tumor phylogeny topology inference via deep learning. iScience. 2020;23 (11):101655.

Searle J. Minds, brains and programs. Behav Brain Sci. 1980;3(3):417–57.

Serpell LC, Radford SE, Otzen D. AlphaFold: a special issue and a special time for protein science. J Mol Biol. 2021:167231.

Shen R, Olshen AB, Ladanyi M. Integrative clustering of multiple genomic data types using a joint latent variable model with application to breast and lung cancer subtype analysis. Bioinformatics. 2009;25(22):2906–12.

Strodel B. Energy landscapes of protein aggregation and conformation switching in intrinsically disordered proteins. J Mol Biol. 2021:167182.

Subramanian I, et al. Multi-omics data integration, interpretation, and its application. Bioinform. Biol. Insights. 2020;14:1177932219899051.

Vangay P, et al. US immigration westernizes the human gut microbiome. Cell. 2018;175(4): 962–72.e10

Wang, Z., M.A. Jensen, and J.C. Zenklusen, A practical guide to the cancer genome atlas (tcga), in statistical genomics: methods and protocols. In: E. Mathé, S. Davis, editors. Springer New York: New York, NY; 2016. p. 111–141.

Webb B, Sali A. Comparative protein structure modeling using MODELLER. Curr. Protoc. Bioinformatics. 2016;54:5.6.1–5.6.37.

Xavier JB, et al. The cancer microbiome: distinguishing direct and indirect effects requires a systemic view. Trends Cancer. 2020;6(3):192–204.

Yang X, et al. High-throughput transcriptome profiling in drug and biomarker discovery. Front Genet. 2020;11:19.

Zhang L, et al. From machine learning to deep learning: progress in machine intelligence for rational drug discovery. Drug Discov Today. 2017;22(11):1680–5.

Zobitz JM, et al. A primer for data assimilation with ecological models using Markov chain Monte Carlo (MCMC). Oecologia. 2011;167(3):599.

Chapter 3
-Omics Technologies and Big Data

Ansgar Poetsch and Yuxue Li

3.1 Multi-Omics Data Types and Underlying Technology

Biomedicine is an interdisciplinary field of academic research and innovation that applies biomedical information, medical imaging technology, genetic chips, nanotechnology, new materials and other technologies (Asif et al. 2018). With the development and gradual application of high-throughput DNA sequencing technology since the turn of the century, the amount of data in the field of life science has grown rapidly (Long et al. 2014). The emergence of second and third-generation sequencing technologies has enabled data avalanche growth, and "really" big data has been achieved. Nowadays biomedical big data comprises several types, including genomic data, metagenomic data, proteomic data, metabolomic data, single-cell data, and biomedical image data (Cirillo and Valencia 2019) et al.

Biomedical big data is divided into two categories: big data storage and big data analysis, with big data storage serving the in-depth analysis of big data (Luo et al. 2016). Big data analysis pertains to single-omics data analysis and multi-omics data integration analysis. What is the motivation for performing multi-omics data integration analysis? Because biological contexts are interacting networks, -omics levels are not independent. Thus, analysis limited to a single-omics level has provides a restricted representation of biology, whereas multi-omics integration analysis can

A. Poetsch (✉)
Queen Mary School, Medical College, Nanchang University, Nanchang, China

College of Marine Life Sciences, Ocean University of China, Qingdao, China

Y. Li
Plant Biochemistry, Ruhr University Bochum, Bochum, Germany

Key Laboratory of Molecular Biophysics of the Ministry of Education, Hubei Key Laboratory of Bioinformatics and Molecular-imaging, Center of Artificial Intelligence Biology, Department of Bioinformatics and Systems Biology, College of Life Science and Technology, Huazhong University of Science and Technology, Wuhan, China

K. Ning (ed.), *Methodologies of Multi-Omics Data Integration and Data Mining*,
Translational Bioinformatics 19, https://doi.org/10.1007/978-981-19-8210-1_3

better explain molecular information and causal relationships in biology including biomedicine.

3.1.1 Genomics & Transcriptomics Data Analysis

In 1986, American geneticist Thomas H. Roderick first proposed genomics. Genomics is an interdisciplinary biological discipline for collective characterization, quantitative research and comparative study of different genomes for all genes of an organism. Genomics includes genome sequencing and analysis to assemble and analyze the function and structure of entire genomes through the use of high-throughput DNA sequencing and bioinformatics. Genomics with the goal of collectively characterizing and quantifying all genes in an organism, its main tools and methods include bioinformatics, genetic analysis, gene expression measurement, and gene function identification. It is worth mentioning that Whole Genome Sequencing (WGS) currently refers to human whole-genome sequencing (Modi et al. 2021). The term "whole" refers to the complete genome sequence within the cell of the species, from the first DNA to the last DNA, complete detection and arrangement, allowing this technology to detect almost any type of mutation on the genome. Whole-genome sequencing has a high value for humans because it contains inherent associations between all genes and life traits, but it also means more data interpretation and greater technical challenges (Nikolayevskyy et al. 2019).

Transcriptomics is the study of gene transcription and regulation of transcription in cells at the overall level. Transcriptome refers to the sum of all RNAs that can be transcribed by a living cell, and is an important way to study cell phenotype and function. In contrast to the genome, the definition of the transcriptome includes temporal and spatial constraints. The gene expression of the same cell in different growth periods and growth environments is not exactly the same. For example, the same tissue expresses almost the same set of genes to distinguish it from other tissues, such as brain tissue or myocardial tissue, which differently express 30% of all genes, showing tissue specificity.

DNA sequencing has made several leaps in technology in recent decades with second and third generation technologies hallmarked by high sequencing speed and parallelization in different commercial realizations: 454 and Solexa, PacBio etc. In respect to data analysis, sequencing is simply the process of converting a physical signal (e.g. light pulse) from DNA sequencing into a computer-readable digital signal (Chen et al. 2021). However, with all high throughput -omics technologies comes the need for sophisticated data analyses to determine and extract valid signals from chemical noise and to limit and control errors in the conversion process. So far, with the rapid accumulation of high-throughput sequencing data, the analysis requirements for genomics and transcriptomics data integration, mining, and visualization are increasing (Karczewski and Snyder 2018). To meet the analytical and user requirements, it is urgent to optimize software and hardware systems for big data analysis, integrate the analysis process, and build an interactive visual analysis platform for big data analysis.

After obtaining DNA sequence information, it is necessary to perform gene prediction to obtain the contained functional genes. Gene prediction refers to the analysis of the assembled genome sequence to identify the functional regions such as genes contained in it according to the knowledge of the gene structure of known organisms or database sequences. Gene prediction is mainly based on sequence similarity search or ab initio prediction based on pattern sequence features. Commonly used prokaryotic gene prediction software include GeneMark, Glimmer, Prodigal, etc., and eukaryotic gene prediction software include GENSCAN, Augustus, GlimmerHMM, PASA, etc.

3.1.2 Metagenomics Data Analysis

Metagenomics was defined for the first time in 1998 as “the collection of all genomes in a microbial community” (Handelsman et al. 1998; Rondon et al. 1999). Metagenomics studies target the entire microbial community in a given habitat. DNA is often extracted directly from environmental samples to study the community structure, species taxonomy, phylogeny, gene function, and metabolic network of environmental microbes (Rondon et al. 1999).

Metagenomics borrows technology from genomics for sequencing, but comes with additional considerations and challenges in all experimental steps: (1) sample collection, processing, and sequencing; (2) sequencing data pre-processing; (3) taxonomic, functional group, and other genomics analysis of the microbiome; (4) statistical and biological function analysis (Thomas et al. 2012).

Sample collection: metagenomics samples face some challenges due to complex environments and individual differences, such as each person’s age, dietary habits, living environment, drug uptake (especially antibiotics), resulting in quite different gut microbiota structures. Moreover, choice of primers and amplification methods can introduce biases or limit metagenome coverage.

Sequencing and library preparation: library preparation is a standardized process, and there are many well-established tools. The choice of sequencing platforms must fit purpose of the experiment. If we want to access low abundance information in the sample, we may require a high-throughput, large-data sequencing result increasing time, costs, and data amount. If the goal is to analyze microbial components, lineages, and so on, the Illumina NextSeq and NovaSeq platforms reach TB level.

Data type and analysis: unlike genomics, which studies a specific species, metagenomics can obtain sequence information of all species in a specific environment. During sequencing data pre-processing, depending on the source of the samples and the target of the analysis, it is needed to filter out sequences that may cause interference (e.g. human genome), commonly used software are BWA, Bowtie, BBMap. Obtaining a feature table containing the relative abundance of bacteria is a critical step in metagenomics research, and most subsequent analyses are based on this. MetaPhlAn is widely used to analyze the composition and species abundance of microbial communities (bacteria, archaea, eukaryotes and viruses).

Statistical and biological function analysis: based on the species and functional composition of the microbiome. The overall difference and group specific difference analysis are often carried out from the analysis target. Sometimes an association analysis of environmental factors and microorganisms is also carried out.

3.1.3 Proteomics Data Analysis

Proteins are the material basis of life activity and the executor of life, so studying proteins is self-evidently important. The study of proteins and their interactions within specific systems is known as proteomics (McArdle and Menikou 2021; Patterson and Aebersold 2003). Proteomics essentially covers all technologies used for the characterization of proteins at a large-scale level, including protein expression levels, posttranslational modifications, protein-protein interactions, protein localizations, thus obtaining an overall and comprehensive representation of cell status for understanding disease biogenesis, cellular metabolism, and so on, a concept first proposed by Marc Wilkins in 1994 (Yadav 2007).

Proteomics is a product of the post-genomic era, far more complex than genomics (Maithal 2002). The presence of the genome is relatively stable, whereas the proteome between cells varies due to biochemical reactions of proteins and genes as well as environmental factors. Whilst the number of human protein coding genes is around 20,000, alternative splicing and posttranslational modifications result in a myriad of theoretical proteoforms with the total number existing in humans still awaiting experimental clarification (Smith et al. 2021). Different proteins can be expressed by the same organism in different parts of living organisms, at different stages of life, and in different environments. Moreover, the same protein may have different functions depending on posttranslational processing—a classical example is protein phosphorylation as signal transduction pathway on/off switch.

A mass spectrometer is an excellent tool for high-throughput proteomics sequencing (McArdle and Menikou 2021) and nowadays indispensable in obtaining any protein feature that can be identified by protein or protein fragment mass. If the sequencer is a ruler measuring the sequence and length of the gene base sequence, the mass spectrometer is a scale calling the quality of the protein fragmentation ions. Mass spectrometry techniques such as high-resolution multistage tandem mass spectrometry are mature in proteomics. Massive high-resolution one-dimensional mass spectrometry (MS) and two-dimensional mass spectrometry (MS/MS) data have been collected to complete some large-scale proteome qualitative and quantitative analysis (Bantscheff et al. 2012). Currently, proteomics methods move towards more comprehensive subjects (such as comprehensive first-level mass spectrometry data-independent acquisition (DIA) data studies) and higher throughput in clinical studies. Importantly, most high-throughput proteome analyses are done on the peptide “bottom up” and not protein level “top down”, which may lead to ambiguities in protein and PTM site identification and quantification (Dupree et al. 2020). It must be pointed out that so far state of the art proteomics (and metabolomics) yields incomplete data, a consequence of the sheer number of proteoforms and their abundance range covering several orders of magnitude ($> = 10^{10}$ in human plasma). Therefore, in practice, with a dedicated workflow focus of the analysis is on certain aspects of the proteome such as protein abundance or a certain protein modification. Inevitably missing data is not only important for proteome data interpretation, but a challenge that must be addressed during -omics data integration, including different -omics data types for mechanistic studies.

Proteome data is commonly analyzed from the following perspectives:

1. Proteins identification:

   Although fragment mass spectra can be used for de novo sequencing, proteins are usually identified by comparing the measured mass spectra against theoretical mass spectra calculated from all protein sequences suspected in the sample by search engines (Andromeda (Cox et al. 2011), SEQUEST (Griffiths et al. 2019), ...) thus previous genome sequencing and annotation is required. Here we list some commonly used protein databases: first comprehensive protein databases, such as NCBI, Uniprot (UniProt: the universal protein knowledgebase 2017), Ensembl (Yates et al. 2020); second species-specific protein databases, such as Arabidopsis (TAIR (Garcia-Hernandez et al. 2002)), rice (RAP-DB (Tanaka et al. 2008)), silkworm (silkdb (Lu et al. 2020)). The use of mass spectral libraries (even theoretical) instead can be advantageous and has gained popularity.

2. Protein quantification:

   The signal intensity of an analyte in mass spectrometry is based on the ionization efficiency of a compound, where analyte concentration is only one of contributing factors. Owing to ongoing quantification method development and dictated by experimental design and question, various methods for relative as well as absolute protein quantification are currently in use. Examples are label-free quantification, tandem mass tag labels, isotope labelled standards. Consequently, the representation of quantitative protein data and metadata will vary and with this the approaches and tools to glean the desired quantitative information for -omics data integration.

3. Protein feature information:

   One of nature's inventions is to modify protein functions by changing protein features such as protein length, modification, interaction with other biomolecules, localization. In principle proteomics is capable of obtaining most if not all feature information by employing tailored methods. In doing so, some of these methods go beyond processing of mass spectral data only. As several methods are specialized, not commonly used, and/or under development, data standards are less mature; rather frequently, published protein feature data does not adhere to any standard. Nonetheless, protein feature information can be invaluable for biological interpretation, justifying the extra effort in data analysis or conversion sometimes needed.

3.1.4 Metabolomics Data Analysis

Metabolomics is the comprehensive analysis of the metabolites present in a biological specimen. The metabolome can be highly dynamic, as some key cellular metabolites display an extremely fast turnover in living organisms. Metabolites display a rich structural and physicochemical diversity. Consequently, metabolomics is the analytically most challenging -omics technology and so far, no separation, identification, and quantification technology powerful enough to cover all metabolites in a single experiment exists. The complementary methods NMR and MS are the cornerstone of current metabolomics; NMR excels at structural elucidation, accuracy and reproducibility, whereas the strengths of MS lie mostly in sensitivity and throughput. Owing to the complexity of most biological samples, online or offline chromatographic (GC or LC) separation is usually done prior to analysis. Furthermore, all the metabolites present in humans remain to be fully elucidated, as quite frequently only partial molecular features (e.g. mass, chemical moieties) are known.

MS-based metabolite identification is far more challenging than protein identification. Still, the field of metbabolomics tolerates ambiguous identifications for instance purely based on retention time and intact mass. But even rich molecular information from MS and MS^n spectra may be insufficient for unequivocal metabolite identification. In fact, common is the confident identification of only 2% of all m/z features in an untargeted metabolomics MS analysis (Silva et al. 2015). Here, one must discern identification of known and unknown metabolites. The usual approach for identification of known metabolites is their matching to previously recorded mass spectra. So just like in proteomics, this comes with a degree of uncertainty that is determined by automated statistical data analysis. The unequivocal identification and structural elucidation (including stereochemistry when applicable) of an unknown metabolite, ideally including authentication with the synthesized compound, can be an arduous procedure often not followed in high-throughput metabolomics. Fortunately, scientific ingenuity in metabolomics data analysis and ever-increasing computational power have been constantly producing and improving identification of unknowns; exemplary approaches are prediction of compound structures based on chemical/enzymatic reaction pathways and prediction of MS fragmentation at astounding accuracy with AI.

NMR delivers connectivity and distance for atoms, and is therefore extremely powerful for the structural elucidation of unknown compounds. Yet employing all tools (e.g. COSY, TOCSY) to establish metabolite molecular structure takes hours or days, which prevents their use in high throughput metabolomics. Instead, methods (e.g. STOCSY, STORM) use spectral correlations between the NMR signal features in the sample and reference spectra for compounds in a spectral library (Dona et al. 2016).

Metabolite data is commonly analyzed from the following perspectives:

1. Metabolite identification:

   Aside from commercial spectral libraries and identification software, free databases for mass and NMR spectra are available (e.g. HMDB (Wishart et al.

2021)). For processing, vendor specific RAW data is usually converted into standardized format for MS (e.g. mzML (Martens et al. 2011)) and NMR (e.g. nmrML (Schober et al. 2018)). It is good practice to provide for each compound one of the four levels (from identified to unknown) of identification according to Metabolomics Standard Initiative (Sumner et al. 2007). Identification may require more than one MS or NMR method, thus identification may be based on several pieces of complementary information.

2. Metabolite quantification:

   In principle, quantification with NMR against one reference compound is straightforward and robust. However for complex bio-samples, both in NMR and MS results may be confounded by overlapping signals. MS spectra are processed with sophisticated algorithms (Tautenhahn et al. 2008) to correct for background signals etc. As already mentioned for proteomics, MS signal intensity is affected my many factors, hence standards, ideally an isotope labelled identical compound for every metabolite, are required for absolute quantification. Relevant identification and quantification information, as well as metadata can be stored in the mzTAB-M standard (Hoffmann et al. 2019). For metabolomics experiment result sharing, there are two popular public repositories, MetaboLights and NIH Metabolomics Workbench.

3.1.5 Single-Cell Data Analysis

The analysis of single-cell genomics data can disclose heterogeneities in cell populations, thus may yield more relevant and informative data (Miao et al. 2021). However, single-cell data analysis faces enormous challenges due to the unknown distribution of heterogeneity between single cells and correlation properties (such as gene structure, gene expression, and so on). At present, there are not many methods for analyzing single-cell genomic data, mainly improved genome assembly methods such as velvet and single-cell gene expression differentiation analysis methods, and there is no specialized technology for in-depth single-cell heterogeneity analysis. Existing methods for single-cell phenotype detection and single-cell sequencing have tentatively determined its feasibility, revealing its far-reaching implications in terms of single-cell methods. Current single-cell research methods are also limited to a small number of single cells in terms of hardware and software architectures, so data analysis is done with CPU clusters that are independent of one another. However, because this trend is accelerating, there are indications that the era of data analysis for hundreds or even thousands of single cells has arrived.

Further, although single-cell sequencing can identify changes in gene expression regulatory networks in various types of cells, it cannot determine the source (DNA level) of gene expression network abnormalities. Genomics analysis can find disease-related susceptibility genes, but it is difficult to answer which cell types and which gene pathways are mainly affected by mutations in susceptibility genes.

Therefore, multi-omics analysis of combined single-cell data, including genome analysis, has higher value for the interpretation of many biological problems.

3.1.6 Biomedical Image Data Analysis

Biomedical images are human measurements at various scales (micro-, macroscopic, etc.). They use a variety of imaging modes (CT scanners, ultrasonic instruments, and so on) to assess the physical properties of the human body. These images are interpreted by clinical experts and have a significant impact on physician's decision-making. Biomedical images are typically three-dimensional (3D), with an additional time dimension (4D) and multiple channels (4-5D).

The characteristics of biomedical images: (1) highly dependent on imaging equipment and imaging environment and there are many different types of images, which are difficult to fuse; (2) the image pixels are large, but the signal-to-noise ratio and the image resolution are low; (3) there are differences and variability among biological individuals. With the advancement of optical imaging instruments and high-precision cell operation technology, biomedical image-related data is rapidly accumulating, and the associated image processing technologies are changing with each passing day. With the reinvestment of the United States and the European Union in biomedical research fields such as brain science, many 2D and 3D medical image processing methods adapted to high throughput and high accuracy of TB level have been proposed, and the potential of related applications has gradually been recognized. The 2012 Nature Methods Bioimage Processing album provides a systematic overview of existing high-throughput bioimage processing methods such as Universal imageJ (Schneider et al. 2012) and PhenoRipper for biomedical images (Rajaram et al. 2012).

Deep learning, which has recently emerged, has largely replaced many other machine learning methods because it avoids the creation of manually engineered features, thereby removing a critical source of error from the process. It can be applied to image segmentation, image registration, lesion detection and auxiliary diagnosis, imaging omics biomarker extraction in biological images. For example, convolutional neural network (CNN) can be used to extract image features or directly complete tasks such as classification and detection. Fully convolutional neural network (FCNN) can obtain pictures of the same resolution as the original picture, which is often used for tasks such as segmentation. Faster Region-proposal based neural network (FRCNN) can be used to detect a variety of objects in images.

Since 2013, the journal BMC Bioinformatics has been collecting and arranging biological image data, analyzing related research papers, and related methodological articles have emerged one after another. At present, biomedical image data is mainly stored in three formats, DICOM, Analyze and NIfTI, and data processing develops rapidly. However, there is still a lack of recognized and standardized biomedical image storage and processing platforms.

3.2 Biological Big Data Research

A trait of all presented -omics technologies (single cell, metagenomic data, etc) is the production of big data, which spurred computational developments to handle all aspects of big data. Generally, humanity has entered the era of big data in all aspects of life and research fields. Big data research, following the general laws of technological innovation, development, and maturity, transforms traditional a priori knowledge-driven research methods into data-driven research.

We will briefly illustrate the history of big data in biomedical sciences and key developments. The development history of multi-omics big data is actually a process of development and innovation of different -omics data collection technologies. The discovery in 1952 that the genetic material is DNA, not protein, started the process of in-depth research on DNA sequences. In 1977, Frederick Sanger developed a DNA sequencing technique and completed the sequencing of the first complete genome, known as the phiX174 virus, which opened up new possibilities for genomics. In 2005, the next-generation sequencing technology 454 was invented, and Illuminas came out the following year. In 2008, the third-generation sequencing technology nanopore sequencing began to be used. All of this information is stored in large generic databases such as NCBI or EBI. Simultaneously, with the development and application of high-throughput sequencing technology, as well as the integration of biotechnology and information technology, the biomedical data types and data scales in large general databases such as NCBI are constantly increasing (Fig. 3.1), which provide a good opportunity for research in the biomedical field.

3.2.1 Research Trend of Biological Big Data

Several directions can be identified for the utilization of big data (Ning and Chen 2015), demonstrating new possibilities and the great power of biomedical big data (Fig. 3.2) to drive research:

In the following, we will elaborate on important major research directions enabling the utilization of big data in biomedicine:

1. Grand ecology

   The development of biomedical big data allows us to better understand the world. The main applications in grand ecology include public health and public safety. The volume of data recorded by hospitals, such as human health status and immunizations, is huge, but without big data, this data is meaningless. Big data analytics can standardize and integrate raw patient data to enrich public health records, and a rich variety of public health records can provide better care. In COVID-19, the successful application of biomedical big data in public safety has played an important role in COVID-19 prevention and control. The location-based spatiotemporal big data results, such as COVID-19 thematic maps, “health codes”, intelligent robots have been applied in many fields related to epidemic

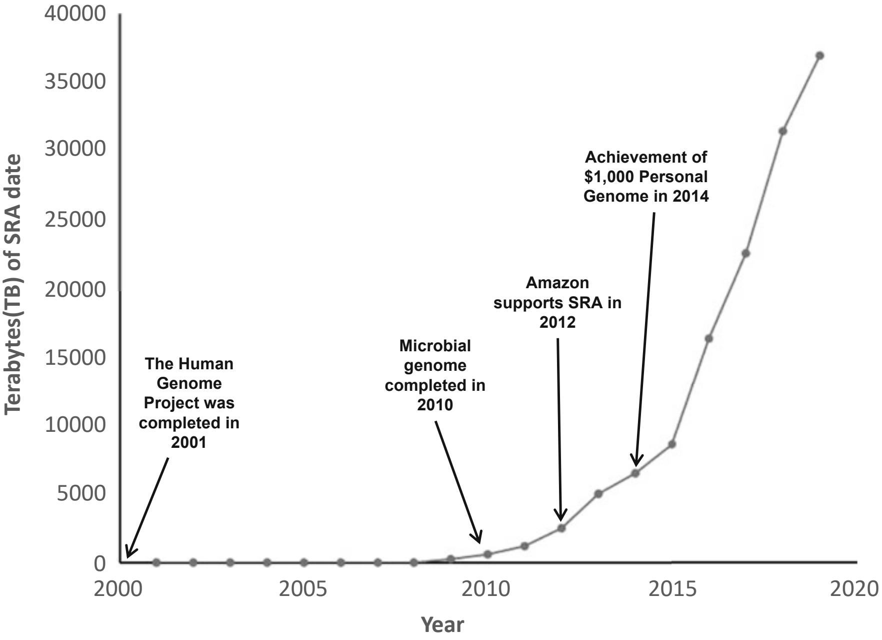

Fig. 3.1 The data growth in the GenBank SRA database in recent years. With the rapid increase of the data sizes, several milestone works have also emerged, as annotated in this figure

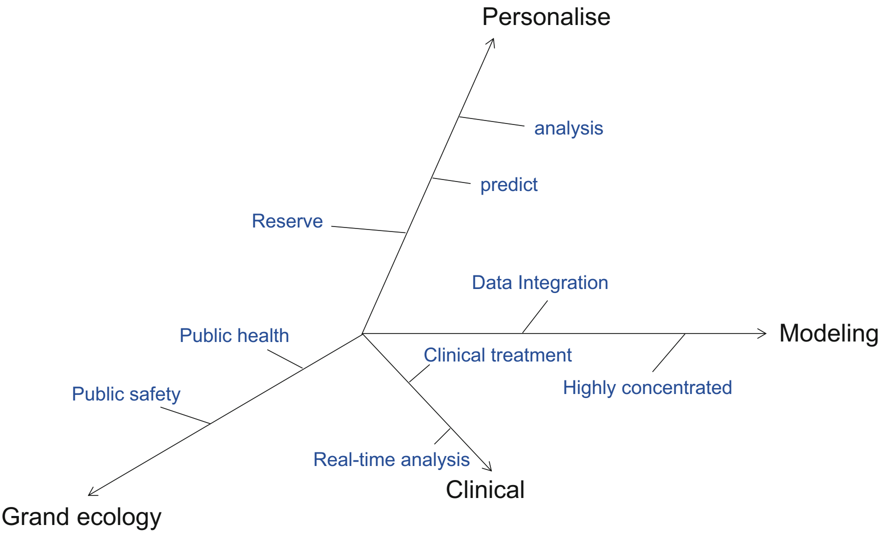

Fig. 3.2 The direction of big data-driven biomedical research. Biomedical big data drives research in personalization, grand ecology, clinical, modeling, and there is personalised research in each field

prevention and control. However, in the process, it was found that the current biomedical big data cannot meet the needs of public health, the spatiotemporal trajectory data is incomplete and the accuracy is not high, and the contradiction between data sharing and privacy is an urgent problem to be solved at present, and it is also the direction of further development in the future.

2. Clinical

   On the other hand, real-time analysis and clinical processing of biomedical data, involving rapid and accurate sampling, data mining, and knowledge discovery, as well as clinical processing or other real-time feedback, is another hot topic in biomedical data research. As many as three million people die each year due to lower respiratory tract infections. Previously, diagnosis was mainly based on microbial culture, but it was time-consuming and had poor sensitivity. In July 2019, the first rapid and economical metagenomic sequencing method using nanopore technology was proposed. This method can accurately and quickly identify bacterial pathogens directly from patient respiratory samples, and can accurately detect resistance genes within 6 h. In cancer diagnosis, there are studies collecting tumor tissue, paracancerous tissue and blood of patients, and analyzing the whole genome and whole transcriptome data of samples in the TCGA database to characterize the cancer-related microbiome. In general, fast and accurate sampling, data mining, knowledge discovery, and clinical processing or other real-time feedback have gradually developed into a system framework that will further promote the clinical application of biomedical big data.

3. Modeling

   Big data integration is a universal problem of biomedical big data, involving several issues such as data format, data contradiction, and data indexing. Intelligent data modeling and analysis will aid in the resolution of the aforementioned issues, can greatly deepen understanding, and is a major focus of biomedical data research (Fig. 3.3). Biomedical data entails the deep integration of different types and scales of data: integration of sample phenotypes, genotypes, and metadata, integration of different samples, and finally the construction of an all-around data model. Modeling needs to be trained with a large amount of actual data to achieve continuous optimization. The final model can accurately and quickly realize the integration and analysis of biomedical big data, and mine effective information from it.

4. Personalisation

   Finally, personalised analysis, prediction, and safe preservation of biomedical data (next-generation electronic medical records) have a wide range of applications. On the one hand, biomedical big data collection and analysis processing will collect genotypic and epigenetic data from massive samples; on the other hand, the relevant data will be sorted out and analyzed, and then a personalized prediction will be provided. At the same time, personalised data will necessitate the security preservation of massive amounts of data.

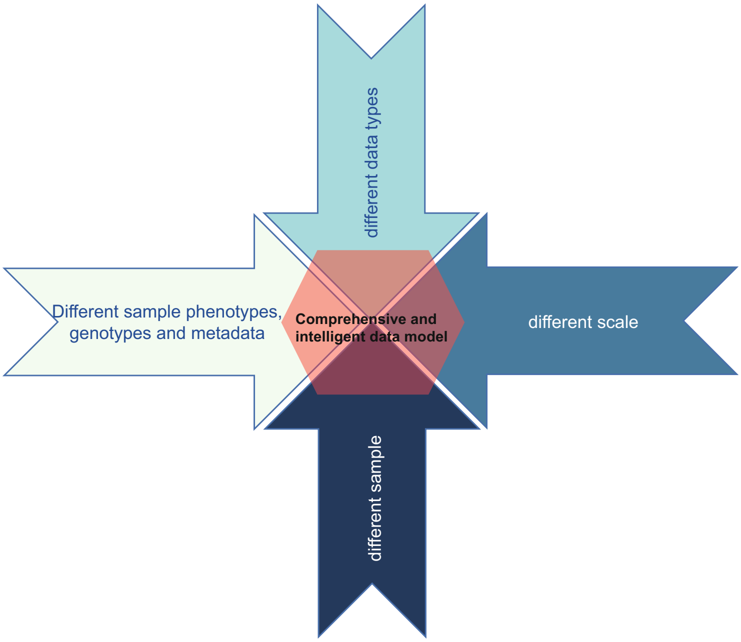
different data types
Different sample phenotypes, genotypes and metadata
Comprehensive and intelligent data model
different scale
different sample

Fig. 3.3 The biomedical big data integration and modeling. Generating data models for different types, scales, samples, phenotypes of data can greatly deepen understanding and is a major focus of biomedical data research

3.2.2 Challenges in -Omics Research

Although all individual -omics datasets may not have the four "V" characteristics associated with "big data" integration, volume, variety, velocity and veracity, they present similar challenges, especially in large sample data research. Each -omics platform has its own characteristics and also faces unique challenges. It is important to understand these when developing methods and approaches to integrate -omics data, as the complexity and completeness of each data type is different. Linking phenotype and genotype is a major challenge, nucleic acid amplification from small amounts of biological material, reliable quantification and molecular annotation based on sequence identity enables high-throughput sequencing. However, the interpretation of many biological problems cannot be based only on genomic and transcriptomic data, but has to explain these phenomena in a specific biological context, that is, the impact of specific variants on phenotypic variation. Combining proteomic and metabolomic data with genomics and transcriptomic data can potentially link genetic and epigenetic variation to phenotypic variation by providing molecular information. In addition, it is also a challenge for the quantification of proteome and metabolome without amplification methods.

There are still many common problems in the -omics data platform. (1) data handling, -omics data generally needs to be filtered, cleaned, converted, standardized for better follow-up analysis, but so far, there is no standardized process for -omics data, and different analysis methods and processes will have a significant impact on the analysis results. (2) data annotating, the quality of -omics data annotation directly affects the accuracy of subsequent analysis. For model organisms, the relevant reference data resources are comprehensive and the available tools are abundant, but for non-model organisms, achieving high-quality annotation is a big challenge. (3) data storaging and sharing, -omics datasets lack standard naming rules and unified data formats, making it difficult to achieve unified access in public databases. It is necessary to formulate standard rules to classify and organize data, and store it in a large public database in a more standardized manner, so as to achieve more effective data sharing and promote research.

3.2.3 Multi-Omics Data Integration Tools and Databases

Extracting meaningful correlations and real interactions from massive -omics data is an extremely difficult process. Especially in multi-omics data analysis, the heterogeneity of different -omics data and the nonlinear interactions and multi-factor combined effects in biological systems make it more complicated to identify real biological signals from random noise. Biological systems, analytical platforms, and data types can all contribute to noise. At present, a large number of tools for integrating -omics data have been developed, including web-based tools that do not require computing experience, and more functional tools for those with computing experience, catering to the needs of different researchers. The following tables (Tables 3.1 and 3.2) list some commonly used multi-omics data integration tools and databases for readers to choose from.

3.2.4 Auxiliary Data and Tools for Multi-Omics Data Integration

3.2.4.1 Relevant Metadata

A large amount of biological big data has accumulated as a result of the advancement of high-throughput sequencing technology (Churko et al. 2013). The increase in data volume allows us to explore many biological issues deeper, and the emergence of many bioinformatics tools and platforms makes it possible to combine samples from different studies for analysis (Magi et al. 2010). However, data integration is very challenging due to the wide range of data sources and the differences in the analysis methods. Metadata can assist in the integration of data, and data analysis benefits from metadata, as it may reveal otherwise hidden patterns in data structure (e.g. sex/age effects). The important role of metadata in data analysis has prompted the development of metadata-related databases and analytical platforms containing

Table 3.1 Tools for multi-omics data integration

Tools	Function	Link address	Reference
MapMan	Visualize and display large datasets on metabolic pathway maps	https://mapman.gabipd.org/	Bolger et al. (2021)
WGCNA	Gather R functions for weighted correlation network analysis	https://labs.genetics.ucla.edu/horvath/CoexpressionNetwork/Rpackages/WGCNA/	Langfelder et al. (2008)
iCluster	Detect novel biomarkers based on transcriptomic and proteomic datasets	https://www.mskcc.org/departments/epidemiology-biostatistics/biostatistics/icluster	Tian et al. (2021)
3Omics	Integrate multiple inter-transcriptomic or intra-transcriptomic, proteomic and metabolomic data	http://3omics.cmdm.tw/	Kuo et al. (2013)
Omics Integrator	Integrating proteomic data, gene expression data, and/or epigenetic data using protein-protein interaction networks	http://fraenkel.mit.edu/omicsintegrator, https://github.com/fraenkel-lab/OmicsIntegrator	Kedaigle et al. (2018)
MixOmics	Enables data exploration, integration, dimensionality reduction and visualization of biological datasets	http://mixomics.org/	Rohart et al. (2017)

Table 3.2 Databases and resources for multi-omics data integration

Databases	Function	Link address	Reference
NCBI	Genomics and transcriptomics	https://www.ncbi.nlm.nih.gov/	Schoch et al. (2020)
MOPED	The database has integrated 250 publicly available protein and mRNA abundance profiles from four different model organisms	http://moped.proteinspire.org	Kolker et al. (2011)
OMICtools	Repository to aid in the integration of -omics datasets	https://omictools.com/	Borsatto et al. (2021)
MetaboLights	Metabolomics datasets	http://www.ebi.ac.uk/metabolights/	Haug et al. (2020)
PeptideAtlas repository	Proteomics	http://www.peptideatlas.org/PASS/PASS00512	Desiere et al. (2006)
Omics Database Generator	Use genome files and the output of various programs to create a graph database for querying genomic data across domains	https://github.com/jguhlin/odg	Guhlin et al. (2017)

metadata. The TCGA database (Tomczak et al. 2015), which contains cancer metadata information, and Qiita (Gonzalez et al. 2018) metadata analysis platform are widely used.

As exemplary metadata analysis tool Qiita will be portrayed: it is free and open source online for microbiome analysis, designed to make it easier for non-bioinformatics scientists to analyze their data or combine data in the database for meta-analysis using standardized analysis processes (such QIIME 2 can start analysis from raw DNA sequences, standardizing the process directly to obtain publication-grade statistical and graphical results.). Qiita has the following features: (1) users do not need any command-line tool knowledge system to perform microbiome data analysis; (2) perform a variety of metagenomic data analyses, such as 16S, 18S, ITS, and WGS, among others; (3) to obtain similar research, upload complete sample information, clarify the research software and parameters, and upload sequence data or intermediate files; (4) submit the biological sequence data used for analysis to EBI-ENA for storage; (5) connect to the database, combine your data with other research data for meta-analysis, realize data sharing within the research team. Qiita stores a wealth of metadata information for various types of samples. For example, a total of 1573 samples are stored for COVID-19, and the sample metadata information includes age, gender, the time of sample acquisition, whether antibiotics were used, and so on (Gonzalez et al. 2018).

A prime example for metadata as well as integration of -omics with clinical data is the TCGA (The Cancer Genome Atlas) project (Huang et al. 2020). In 2006, the National Cancer Institute (NCI) and the National Human Genome Research Institute (NHGRI) collaborated to launch the project, a watershed moment in cancer genome research. More than 20,000 primary cancers and matched normal samples from 33 cancer types were molecularly described (Wang et al. 2016). In the TCGA, we can find available clinical information for liver cancer (may include demographic information, treatment information, survival data, etc.); pathology report (partial samples); mRNA data (measured by mRNA chip or RNA-Sequencing); copy number (the ratio of each segment of chromosome obtained by SNP chip sequencing to the tumor compared with normal tissue); mutation (nucleotide changes obtained from tumor tissue sequencing data relative to the reference genome sequence, including changes such as insertions, deletions, and so on); protein (the expression level of approximately 200 common cancer genes (the degree of DNA methylation was obtained by methylation chip sequencing).

3.2.4.2 Quality Assurance Example

Shotgun sequencing, which randomly interrupts mRNA at one time and selects sequences with a suitable length for large-scale sequencing, is used in second-generation sequencing technology. The obtained random short sequences are then merged, and short sequences with the same part are spliced together. The short sequence has a high number of repetitive parts in this process, and the presence of merging codons frequently causes errors. As a result, different outcomes will be measured for a given base position, and the quality value can be calculated by running these outcomes through a series of calculations. The goal of quality control is to eliminate extremely low-quality sequences (Kchouk et al. 2017).

FastQC (de Sena Brandine et al. 2019) is a Java-based software that provides a straightforward method for quality control of raw sequence data from high-throughput sequencing pipelines. It is capable of performing a fast multi-threaded quality evaluation of sequencing data (Quality Control). It is used to quickly determine whether there are any issues with the data, and one should pay attention to these issues before proceeding with further analysis. It can be used to quickly determine whether there are any issues with the data, and one should be aware of these issues before moving forward with further analysis. FastQC's primary functions are as follows: (1) import data from BAM, SAM, or FastQ files; (2) provide a quick overview and indicate which areas may have issues; (3) summarize charts and tables to quickly evaluate your data; (4) export the results to a permanent HTML-based report. (5) offline operation enables automatic report generation without the need for interactive applications to be run (de Sena Brandine et al. 2019). For the output results of FastQC, we need to focus on, (1) Basic Statistics; (2) per base sequence quality; (3) per sequence quality scores; (4) per base sequence content; (5) sequence duplication levels.

3.3 Case Studies on Multi-Omics Data Integration: Resources and Applications

Multi-omics research can provide more comprehensive biological information than single-omics research. Some progress has been made in the solution of biomedical problems, which will hopefully be applied to the study of complex physiological processes such as human cells and diseases. Multi-omics research has been widely used to connect cancer and genetic information.

3.3.1 Multi-Omics Data Resources for Human Brain Diseases

In recent years, multi-omics data research on human brain diseases has grown in scope. American researchers published "Review of multi-omics data resources and integrative analysis for human brain disorders" in Briefings in Functional Genomics in May 2021, summarizing the brain multi-omics data resources of the healthy control group and neuropsychiatric diseases, such as schizophrenia, autism, bipolar disorder, Alzheimer's disease, Parkinson's disease, progressive supranuclear palsy, and others (Dong et al. 2021).

ENCODE is a project that began in 2003 to determine the regulatory function of about 1% of the human genome. Although ENCODE is primarily concerned with cell lines, recent updates have included some -omics data from the human brain, primary neurons, or neuronal cell lines (The ENCODE (ENCyclopedia Of DNA Elements) Project 2004). Roadmap Epigenomics is primarily concerned with the collection of RNA-seq, ChIP-seq (histone), DNase-seq, and methylation data from human blood and 22 tissue types (Kundaje et al. 2015). OmicsDI is a search platform for multiple -omics data sets. It combines data sets from multiple databases in proteomics, genomics, metabolomics, and transcriptomics. Searches for the keyword "brain" produced 116,261 results as of December 2020, 10 of which are multi-omics data sets (Perez-Riverol et al. 2017). PsychENCODE, the data features the largest brain collection (2793 unique donors), including control and disease group data including schizophrenia, bipolar disorder, and autism. The frontal cortex is the main brain area studied by the consortium (Akbarian et al. 2015). The BRAIN Initiative and related brain research projects generate -omics data, for which NeMO, a data repository, is dedicated to storing and sharing. NeMO data includes transcriptional activity, methylation, histone modification profile, and chromatin accessibility for humans, mice, and marmosets. A current search of human data on the BICCN website reveals that 418 scRNA-seq ($n = 412$) and scATAC-seq ($n = 6$) samples are publicly available. More human brain single-cell -omics data (for example, single-cell PLAC-seq, ATAC-seq, and RNA-seq data used to define cell type-specific 3D epigenomes) are available on NeMO via restricted access (Song et al. 2020).

3.3.2 Multi-Omics Data Resources for Cancer Cell Lines

Cancer cell lines are the most widely used model for studying cancer biology, identifying cancer targets, and assessing drug efficacy. Previous cell line research was limited to a few commonly used cell lines or up to 60 cell lines using NCI60 panels. As a result of the scarcity of large-scale, robust, and well-defined cancer cell line models, the high sensitivity of cancers with activating EGFR mutations was initially overlooked. As the Cancer Genome Anatomy (TCGA) project began to define the genetic basis of human cancer, it became clear that similar efforts would be required to determine the characteristics of cancer cell lines.

In 2008, the "Cancer Cell Line Encyclopedia Project" was launched. The goals are to (1) perform detailed gene and drug characterization on a large number of human cancer models; (2) develop comprehensive computational analysis to link unique drug vulnerabilities with characteristic inheritance, gene expression, and cell lineage patterns; and (3) integrate cell lines. Cancer patients can be stratified using genomics. There are 1457 different cell lines, 84,434 genes, 136,488 different databases, 1,159,663 mutation entries, 118,661,636 distribution scores, and 411,948,577 methylation site scores in the CCLE database. To put it another way, the CCLE database contains gene expression profile data (Affy chip and RNA-seq), copy number data, mutation data, and methylation data from various cell lines.

We can see six different data set modules in the CCLE database's function module: (1) Achilles shRNA knockdown; (2) Copy number; (3) DNA methylation;

(4) Protein Array; (5) mRNA expression (Affy); (6) mRNA expression (RNA-seq). The corresponding -omics data is downloaded based on individual analysis requirements. CCLE is a vital data resource in pan-cancer research.

3.3.3 Multi-Omics Research for Retinoblastoma

Multi-omics studies have investigated cancer-causing tumors related to genetic background. A comprehensive genomics study showed that there is a strong correlation between the genetic status of tumors and multi-omics data (Khan and Azmir 2020). The multi-omics in cancer project carried out a "multi-omics method to identify disease progression biomarkers of retinoblastoma" study in November 2016. Retinoblastoma is a type of pediatric eye cancer that typically affects children under the age of five. It's a complicated disease caused primarily by biallelic inactivating mutations in the RB1 gene (Limonte et al. 2020). Tumors, aqueous humor, vitreous, and tear fluid samples were obtained from nine patients whose eyes were removed, and the retina, aqueous humor, and vitreous were obtained from the eyes of the two children who died. Their deaths were not caused by any ophthalmological diseases.

The authors used microarrays for mRNA and miRNA gene expression, and then performed pathway analysis to determine gene enrichment, thereby realizing the functional characterization of tumors. Differential expression analysis showed that 108 ($p \leq 0.05$, fold change $\geq$10) genes are unique genes in patients with high risk of metastasis. Pathway analysis revealed the key pathways involved in the progression of retinoblastoma, including the cell cycle and Rap1 signaling pathway. In this study, transcriptomics, metabolomics, and proteomics are integrated, and 18 miRNAs of retinoblastoma cancer cell lines that may be related to retinal cancer are reported.

3.3.4 Multi-Omics Research for Cardiovascular Disease

Cardiovascular disease is the leading cause of morbidity and death worldwide (Pagidipati and Gaziano 2013). Discrete transcriptional regulatory pathways and global protease inhibitors interfere with the expression profile of cardiac proteins, causing them to change in pathological myocardial hypertrophy. Many studies are now using transcriptomics (such as RNA-seq) and proteomics (such as mass spectrometry) to identify disease characteristics and pathogenic mechanisms in cardiac hypertrophy and other diseases. However, because of the poor correlation between the differential expression of transcripts and the differential expression of proteins, it is still debatable whether transcriptomics and proteomics experiments reflect the same biological laws. Furthermore, does the irrelevance of transcription for protein reflect the dominance of different translational regulation in protein abundance, or is it simply the result of unexplained technological variation?

Previous genomics research has linked coronary heart disease to a group of 150 genomes (Deloukas et al. 2013). In January 2018, the journal *Nature* Communication published an article titled "Integrated omics dissection of proteome dynamics during cardiac remodeling," which integrated transcriptomics, proteomics, and protein turnover for multi-omics research and evaluated their effects in vitro. The data had a synergistic effect for the myocardial hypertrophy model mice.

This study looked at the reproducible hypertrophy features in each -omics data type of six mouse genetic strains and discovered that combining transcript abundance, protein abundance, and protein turnover data results in disease candidate genes. Furthermore, the inclusion of protein turnover measurements enables the discovery of different post-transcriptional regulatory pathways and implies the presence of different disease proteins not found in steady-state transcription and protein abundance data.

3.3.5 Multi-Omics Research for Infectious Disease

Infectious diseases in wild animals that are newly emerging pose a serious threat to biodiversity, and infectious diseases are the cause of a decline in the number of large numbers of wild animals. Chytrid disease, caused by Dendrobium chytrid, has severely impacted many amphibian populations and species around the world (Rebollar et al. 2016). A potential research strategy for its research is to enhance the amphibian skin with antifungal bacteria through probiotic organisms. In vivo experiments using bioaugmentation strategies have yielded mixed results, necessitating a more informed strategy for selecting successful probiotic candidates. Multi-omics integrated analysis, such as metagenomics, transcriptomics, and metabolomics, can better guide probiotic selection and optimize the selection method.

This study employs bioinformatics and statistical tools to integrate multiple sets of data, and it is possible to identify species involved in pathogen suppression using an in silico model that connects bacterial community structure and bacterial defense functions. For in-field investigations and experiments, the authors used 16S rRNA gene amplification and sequencing, index species analysis, Kolmogorov-Smirnov measurement, and symbiosis network methods to identify bacteria related to pathogen resistance. In addition to 16S amplicon sequencing, methods for the in-depth understanding of symbiosis functions, such as shotgun metagenomics, metatranscriptomics, or metabolomics, are recommended to increase the likelihood of finding beneficial bacteria candidates. Probiotics can be isolated and tested in ongoing and clinical trials.

In summary, multi-omics data integration and organization techniques have been developed, and a broad-spectrum of applications have been conducted based on these techniques and the organized multi-omics data. These advancements in data integration and organization have also called for data mining techniques that could best utilize multi-omics data for novel knowledge discovery.

References

Akbarian S, et al. The PsychENCODE project. Nat Neurosci. 2015;18(12):1707–12.

Asif MRA, et al. Role and impact of biomedical engineering discipline for developing country perspective. Int J Innov Res Comput Sci Technol. 2018;6:87–90.

Bantscheff M, et al. Quantitative mass spectrometry in proteomics: critical review update from 2007 to the present. Anal Bioanal Chem. 2012;404(4):939–65.

Bolger M, Schwacke R, Usadel B. MapMan visualization of RNASeq data using Mercator4 functional annotations. Methods Mol Biol. 2021;2354:195–212.

Borsatto KC, et al. Omics tools applied to the study of Chagas disease vectors: cytogenomics and genomics. Am J Trop Med Hyg. 2021;104(6):1973–7.

de Sena Brandine G, Smith AD. Falco: high-speed FastQC emulation for quality control of sequencing data. F1000Res. 2019a;8:1874.

Chen K, et al. Electrical DNA sequence mapping using oligodeoxynucleotide labels and nanopores. ACS Nano. 2021;15(2):2679–85.

Churko JM, et al. Overview of high throughput sequencing technologies to elucidate molecular pathways in cardiovascular diseases. Circ Res. 2013;112(12):1613–23.

Cirillo D, Valencia A. Big data analytics for personalized medicine. Curr Opin Biotechnol. 2019;58:161–7.

Cox J, et al. Andromeda: a peptide search engine integrated into the MaxQuant environment. J Proteome Res. 2011;10(4):1794–805.

Deloukas P, et al. Large-scale association analysis identifies new risk loci for coronary artery disease. Nat Genet. 2013;45(1):25–33.

Desiere F, et al. The PeptideAtlas project. Nucleic Acids Res. 2006;34(Database issue):D655–8.

Dona AC, et al. A guide to the identification of metabolites in NMR-based metabonomics/metabolomics experiments. Comput Struct Biotechnol J. 2016;14:135–53.

Dong X, Liu C, Dozmorov M. Review of multi-omics data resources and integrative analysis for human brain disorders. Brief Funct Genomics. 2021;20(4):223–34.

Dupree EJ, et al. A critical review of bottom-up proteomics: the good, the bad, and the future of this field. Proteomes. 2020;8(3):14.

Garcia-Hernandez M, et al. TAIR: a resource for integrated Arabidopsis data. Funct Integr Genomics. 2002;2(6):239–53.

Gonzalez A, et al. Qiita: rapid, web-enabled microbiome meta-analysis. Nat Methods. 2018;15(10): 796–8.

Griffiths RL, et al. Direct mass spectrometry analysis of protein complexes and intact proteins up to >70 kDa from tissue. Anal Chem. 2019;91(11):6962–6.

Guhlin J, et al. ODG: Omics database generator – a tool for generating, querying, and analyzing multi-omics comparative databases to facilitate biological understanding. BMC Bioinformatics. 2017;18(1):367.

Langfelder P, Horvath S. WGCNA: an R package for weighted correlation network analysis. BMC Bioinformatics. 2008;9(1):559.

Limonte CP, et al. A targeted multiomics approach to identify biomarkers associated with rapid eGFR decline in type 1 diabetes. Am J Nephrol. 2020;51(10):839–48.

Handelsman J, et al. Molecular biological access to the chemistry of unknown soil microbes: a new frontier for natural products. Chem Biol. 1998;5(10):R245–9.

Hoffmann N, et al. mzTab-M: a data standard for sharing quantitative results in mass spectrometry metabolomics. Anal Chem. 2019;91(5):3302–10.

Huang F, et al. CILP2 overexpression correlates with tumor progression and poor prognosis in patients with colorectal cancer in the cancer genome atlas (TCGA) study. World J Surg Oncol. 2020;18(1):274.

Haug K, et al. MetaboLights: a resource evolving in response to the needs of its scientific community. Nucleic Acids Res. 2020;48(D1):D440–4.

Modi A, et al. The illumina sequencing protocol and the NovaSeq 6000 system. Methods Mol Biol. 2021;2242:15–42.

Karczewski KJ, Snyder MP. Integrative omics for health and disease. Nat Rev Genet. 2018;19(5): 299–310.

Kchouk M, Gibrat JF, Elloumi M. Generations of sequencing technologies: from first to next generation. Biol Med. 2017;09:03.

Kedaigle AJ, Fraenkel E. Discovering altered regulation and signaling through network-based integration of transcriptomic, epigenomic, and proteomic tumor data. Methods Mol Biol. 2018;1711:13–26.

Khan MS, Azmir J. Multi-omics for biomedical applications. J Appl Bioanal. 2020;6(3):97–106.

Kolker E, et al. MOPED: Model Organism Protein Expression Database. Nucleic Acids Res. 2011;40(D1):D1093–9.

Kundaje A, et al. Integrative analysis of 111 reference human epigenomes. Nature. 2015;518 (7539):317–30.

Kuo T-C, Tian T-F, Tseng YJ. 3Omics: a web-based systems biology tool for analysis, integration and visualization of human transcriptomic, proteomic and metabolomic data. BMC Syst Biol. 2013;7(1):64.

Long Q, et al. The development and application of high throughput cultivation technology in bioprocess development. J Biotechnol. 2014;192:323–38.

Lu F, et al. SilkDB 3.0: visualizing and exploring multiple levels of data for silkworm. Nucleic Acids Res. 2020;48(D1):D749–55.

Luo J, et al. Big data application in biomedical research and health care: a literature review. Biomed Informatics Insights. 2016;8:1–10.

Magi A, et al. Bioinformatics for next generation sequencing data. Genes. 2010;1(2):294–307.

Maithal K. Proteomics—a new player in the post-genomic era. Indian J Biochem Biophys. 2002;39 (5):291–302.

Martens L, et al. mzML—a community standard for mass spectrometry data. Mol Cell Proteomics : MCP. 2011;10(1):R110.000133-R110.000133.

McArdle AJ, Menikou S. What is proteomics? Arch Dis Child Educ Pract Ed. 2021;106(3):178–81.

Miao Z, et al. Multi-omics integration in the age of million single-cell data. Nat Rev Nephrol. 2021;17(11):710–24.

Nikolayevskyy V, et al. Role and value of whole genome sequencing in studying tuberculosis transmission. Clin Microbiol Infect. 2019;25(11):1377–82.

Ning K, Chen T. Big data for biomedical research: current status and prospective. Chin Sci Bull. 2015;60(0023-074X):534.

Pagidipati NJ, Gaziano TA. Estimating deaths from cardiovascular disease: a review of global methodologies of mortality measurement. Circulation. 2013;127(6):749–56.

Patterson SD, Aebersold RH. Proteomics: the first decade and beyond. Nat Genet. 2003;33(3): 311–23.

Perez-Riverol Y, et al. Discovering and linking public omics data sets using the omics discovery index. Nat Biotechnol. 2017;35(5):406–9.

Rajaram S, et al. PhenoRipper: software for rapidly profiling microscopy images. Nat Methods. 2012;9(7):635–7.

Rebollar EA, et al. Using "omics" and integrated multi-omics approaches to guide probiotic selection to mitigate chytridiomycosis and other emerging infectious diseases. Front Microbiol. 2016;7:68.

Rohart F, et al. mixOmics: an R package for 'omics feature selection and multiple data integration. PLoS Comput Biol. 2017;13(11):e1005752.

Rondon MR, et al. Toward functional genomics in bacteria: analysis of gene expression in Escherichia coli from a bacterial artificial chromosome library of Bacillus cereus. Proc Natl Acad Sci U S A. 1999;96(11):6451–5.

Schneider CA, Rasband WS, Eliceiri KW. NIH image to ImageJ: 25 years of image analysis. Nat Methods. 2012;9(7):671–5.

Schober D, et al. nmrML: a community supported open data standard for the description, storage, and exchange of NMR data. Anal Chem. 2018;90(1):649–56.

Schoch CL, et al. NCBI Taxonomy: a comprehensive update on curation, resources and tools. Database (Oxford). 2020;2020.

Silva RRD, Dorrestein PC, Quinn RA. Illuminating the dark matter in metabolomics. Proc Natl Acad Sci. 2015;112(41):12549–50.

Smith LM, et al. The human Proteoform project: defining the human proteome. Sci Adv. 2021;7 (46):eabk0734-eabk0734.

Song M, et al. Cell-type-specific 3D epigenomes in the developing human cortex. Nature. 2020;587 (7835):644–9.

Sumner LW, et al. Proposed minimum reporting standards for chemical analysis Chemical Analysis Working Group (CAWG) metabolomics standards initiative (MSI). Metabolomics. 2007;3(3): 211–21.

Tanaka T, et al. The Rice annotation project database (RAP-DB): 2008 update. Nucleic Acids Res. 2008;36(Database issue):D1028–33.

Tautenhahn R, Böttcher C, Neumann S. Highly sensitive feature detection for high resolution LC/MS. BMC Bioinformatics. 2008;9(1):504.

Tian S, Wang C. An ensemble of the iCluster method to analyze longitudinal lncRNA expression data for psoriasis patients. Hum Genomics. 2021;15(1):23.

The ENCODE (ENCyclopedia Of DNA Elements) Project. Science. 2004;306(5696):636–40.

Thomas T, Gilbert J, Meyer F. Metagenomics—a guide from sampling to data analysis. Microb Inform Exp. 2012;2(1):3.

Tomczak K, Czerwińska P, Wiznerowicz M. ReviewThe cancer genome atlas (TCGA): an immeasurable source of knowledge. Contemp Oncol/Współczesna Onkologia. 2015:68–77.

UniProt: the universal protein knowledgebase. Nucleic Acids Res. 2017;45(D1):D158–d169.

Wang Z, Jensen MA, Zenklusen JC. A practical guide to the cancer genome atlas (TCGA). Methods Mol Biol. 2016;1418:111–41.

Wishart DS, et al. HMDB 5.0: the human metabolome database for 2022. Nucleic Acids Res. 2021;50(D1):D622–31.

Yadav SP. The wholeness in suffix -omics, −omes, and the word om. J Biomol Techniques: JBT. 2007;18(5):277.

Yates AD, et al. Ensembl 2020. Nucleic Acids Res. 2020;48(D1):D682–d688.

Chapter 4
Multi-Omics Data Mining Techniques: Algorithms and Software

Min Tang, Yi Liu, and Xun Gong

Abbreviations

ADMM	Alternating direction method of multipliers
AI	Artificial intelligence
API	Application programming interface
ARMI	Assisted robust marker identification
AUC	Area under the curve
BRCA	Breast cancer dataset
BXD	Murine liver dataset
CD	Coordinate descent
CNV	Copy number variation
COAD	Colon adenocarcinoma
DL	Deep learning
EDA	Exploratory data analysis
EM	Expectation–maximization
FAIR	Findable, accessible, interoperable, and reproducible
FDR	False discovery rate
GBM	Glioblastoma
GE	Gene expression
GPU	Graphics processing unit
GWAS	Whole genome association study
KEGG	Kyoto Encyclopedia of Genes and Genomes
LAD	Least absolute deviation
LASSO	Least absolute shrinkage and selection operator
LDA	Linear discriminant analysis

M. Tang (✉) · Y. Liu · X. Gong
School of Life Sciences, Jiangsu University, Zhenjiang, Jiangsu, China

Department of Rheumatology & Immunology, the Affiliated Hospital of Jiangsu University, Zhenjiang, China
e-mail: mt3138@ujs.edu.cn

K. Ning (ed.), *Methodologies of Multi-Omics Data Integration and Data Mining*,
Translational Bioinformatics 19, https://doi.org/10.1007/978-981-19-8210-1_4

LIHC	Liver hepatocellular carcinoma
LPP	Locality preserving projections
LRMs	Linear regulatory modules
LUSC	Lung squamous cell carcinoma
MALA	Microarray logic analyzer
MCCA	Multiple canonical correlation analysis
MCIA	Multiple co-inertia analysis
MCMC	Markov chain Monte Carlo
MCP	Minimax concave penalty
ML	Machine learning
MOFA	Multi-omics factor analysis
NGS	Next generation sequencing
OR	Odds ratio
PCA	Principal component analysis
PMC	PubMed Central
QC	Quality control
R	Statistical programming language R
rMKL	Robust multiple kernel learning
SARC	Sarcoma Alliance for Research through Collaboration
SCAD	Smoothly clipped absolute deviation
SKCM	Skin cutaneous melanoma
SNF	Similarity network fusion
SNP	Single nucleotide polymorphism
TCGA	The Cancer Genome Atlas

With the aid of cost-effective next-generation sequencing technologies, the datasets with multiple dimensions, called multi-omics or integrated omics, have been dramatically accumulated. Because of the limitation of application individual omics, multi-omics efforts have been playing the lead in bioinformatics and biomedical research—from simple computation to data mining. As multi-omics, a merger of biology, informatics, data science and computational sciences, has incredible high complexity, the multi-omics data mining techniques are indigestible to researchers new to this field. The present review is to provide an overview of the current state of the field. On the one hand, we do our best to summarize the algorithms and software designed for the horizontal or vertical integration of the omics data from the various high-throughput sequencing platforms. For each method, we give a complete survey on software and their algorithms that are frequently used coupled with a brief discussion about the principles for applying these computational strategies and considerations, especially in cancer research. On the other hand, we also give a summary of the user-friendly tools suited for multi-omics data interpretation, analysis, and visualization. To our knowledge, this is the most complete and updated summary of publicly available resources about multi-omics data mining. We hope the readers can get inspiration here for their own multi-omics data analysis.

4.1 Introduction

Huge volumes of biology data have been generated from various sequencing machines at multiple levels, including the genome, transcriptome, proteome, epigenome, and metabolome, since the advent of high-throughput technologies. Furthermore, the omics spectrum can be expanded to include the lipidome, glycolproteome, and phosphoproteome. These so-called "multi-omics" have unprecedentedly aided in the capture of so much information related to life sciences that have temporally aided the progress of personalized medicine (Ghosh et al. 2018; Karczewski and Snyder 2018), agriculture (Ichihashi et al. 2020), microbiology (Quinn et al. 2016), plant science (Liu et al. 2016). As further research has become increasingly dependent on data mining, an increasing number of algorithms and software have been developed for multi-omics data.

Although individual omics analysis has been widely used in biology-related studies, an integrative analysis on multi-omics data not only provides manyfold more meaningful results than individual omics, but also maximizes comprehensive biological insight via jointed data mining (Krassowski et al. 2020). Individual omics are generally combined sequentially or simultaneously by integrated approaches to reveal the interplay of molecules (Biswas and Chakrabarti 2020; Zhang et al. 2012). In addition, to help in assessing the information flow from one-dimension data to the other and then enrich the availability results, the integrative approaches can potentially enable researchers to overcome the formidable obstacles that individual omics internally have (Misra et al. 2019; Gomez-Cabrero et al. 2014). However, during this process, they also inevitably bring about new challenges of data fusion, clustering, visualization, and functional characterization (Pinu et al. 2019). Although many tools use a combination of approaches, data integration algorithms can be broadly classified as fusion-based, network-based, Bayesian-based, correlation-based, similarity-based, and other multivariate methods. Approaches developed for multi-omics data mining in cancer research, in particular, can reveal tumor subtypes, biological mechanisms, and identify driver genomic alternations. As a result, they have been widely used in diagnostics, tumor classification, and prognosis (Ickstadt et al. 2018). Especially, the approaches developed for multi-omics data mining in cancer research can reveal can tumor subtypes, biological mechanism and identify driver genomic alternations. Thus, they have been widely employed for diagnostics, tumor classifications and prognostications (Biswas and Chakrabarti 2020; Menyhárt and Győrffy 2021; Mantini et al. 2021; Kong et al. 2020; Rappoport and Shamir 2018).

In this review, we provide an overview of popular software for multi-data integration, interpretation, and visualization in genomics, proteomics, transcriptomics, metabolomics, and epigenomics, among other fields. We concentrate on approaches that use parallel data set the integration to integrate at least two omics data sets derived from at least partially overlapping samples and are readily available in algorithms and software.

4.2 Software for Multi-Omics Data Integration

Hundreds of integrative methods for multi-omics have been developed over the last two decades. The vast majority of those present were supervised, unsupervised, or semi-supervised. It is unfortunate that some are no longer maintained or have lost their code. We do our best to summarize the entire information of each method for the benefit of the readers. Several studies compared those methods to real-world data and/or simulations (Meng et al. 2016; Tini et al. 2017; Bersanelli et al. 2016; Pierre-Jean et al. 2019; Song et al. 2020).

4.2.1 Matrix Factorization Methods

4.2.1.1 Joint/Integrative Non-negative Matrix Factorization (jNMF, iNMF)

The Non-negative Matrix Factorization (NMF) method has been widely used for unsupervised data integration, which projects differences between datasets onto dimension-reduced space (Bakal et al. 2019; Gligorijević et al. 2019). Zhang et al. proposed splitting a non-negative matrix into two matrices: a specific coefficient matrix and a common basis matrix for multi-omics data integration (Zhang et al. 2012; Zhang et al. 2011). It is worth noting that the NMF rationale is to parse data onto a common basis space rather than simply correlation. This allows one to take stock of the elements with significant z-scores before detecting coherent patterns among datasets. Furthermore, this method has several extensions: The iNMF framework considers heterogeneous effects during data integration (Yang and Michailidis 2015), whereas jNMF can detect homogeneity in datasets (Zhang et al. 2012). The homogeneous and heterogeneous patterns are then combined using a combination of NMF and jNMF objective functions. A homogeneity parameter aids in accounting for dataset heterogeneity. Because the objective function of iNMF is nonconvex, the process should be repeated many times to obtain the optimal minimal objective function.

However, because of the wide range of distributions and variability, the application of NMF necessitates not only non-negative input matrices, but also an appropriate normalization step for the inputs. Furthermore, NMF necessitates a large amount of memory and is time-consuming.

4.2.1.2 iCluster

Shen et al. (Shen et al. 2009) created a regularized joint latent variable for integrative clustering that is similar to the common factor but does not have non-negative constraints. The unsupervised method reduces the dimensionality of the datasets in a single framework while also incorporating flexible modeling of the associations between different data types and the variance-covariance structure, resulting in a single cluster assignment for all samples. A likelihood-based inference is computed using the expectation-maximization algorithm. It cannot, however, handle both categorical and continuous variables.

4.2.1.3 iCluster+

Mo et al. (Mo et al. 2013) proposed iCluster+, a framework for joint modeling of discrete and continuous variables resulting from genomic, transcriptomic, and epigenomic profiles. It is an advanced version of iCluster that employs generalized linear regression to create a joint model of categorical and numerical variables (continuous and count). According to their hypothesis, a set of orthogonal latent variables representing distinct molecular drivers can be used to predict diverse molecular phenotypes, which can then reveal clinical and biologically significant subgroups. To induce sparsity, a penalized likelihood approach with lasso penalty terms was used. The lasso regression can determine the subset of features that contribute to biological variation between subtypes. However, due to the computationally intensive approach and the use of penalized regression, iCluster is constrained by esoteric statistical inference.

4.2.1.4 Multiple Factor Analysis (MFA)

MFA is popular software that integrates several ‘omics’ data (numerical and/or categorical) through a projection and pushes them into a low-dimensional variable space. As a result, the integration of numerical and categorical variables is possible, which aids in the incorporation of a supplementary group of data into the downstream analysis. Each dataset is also subjected to principal component analysis (PCA) to identify the individual pattern. The variance-covariance matrix is used in global analysis to identify the common structure in each dataset. A variable matrix for visualizing individual and common structures is also generated. The implementable code is included in FactomineR’s R package (http://factominer.free.fr) as one of the multivariate methods (de Tayrac et al. 2009).

4.2.1.5 Joint and Individual Variation Explained (JIVE)

JIVE is a variation of the NMF category and an extension of PCA with clear advantages over popular two-block methods such as Canonical Correlation Analysis (CCA) and Partial Least Squares. It separates the joint and individual effects of the datasets as a general decomposition of variation for the integrated analysis multi-omics by decomposing the datasets into three terms, residual noise, a low-rank approximation capturing joint variation between omics, and a low-rank approximation for structured variation individual to each omics. As a result, JIVE can quantify the amount of joint (shared) variation between omics, reduce dimensionality, and enable visual exploration of the individual (omic-specific) and joint structure, leading to new directions for visual exploration of joint and individual structure. Although outliers reduce robustness, JIVE estimates common features more accurately (Lock et al. 2013). The R package of R.JIVE developed in CRAN is a wrapper of JIVE with important extensions, which improves the JIVE accessibility to the bioinformatics community, and the speed and flexibility of result visualization (O'Connell and Lock 2016).

4.2.1.6 Joint Bayes Factor

A beta-Bernoulli process is used in the Joint Bayes Factor, which, like JIVE, assumes a common factor for data-specific and a shared factor across all datasets (Ray et al. 2014). The original inputs are decomposed into residual noise, common factors shared across omics data, and omic-type specific factors. Unlike JIVE, this non-parametric method employs penalties to introduce sparsity and assumes a beta-Bernoulli process for the two types of factors (Thibaux and Jordan 2007). The factor loading matrix, like many other biological studies, should be sparse. To add the sparsity, a student-t sparseness-promoting prior or a spike-slab prior is used (Tipping 2001; Chen et al. 2011; Carvalho et al. 2008). Sparsity was also imposed on the factor scores for modeling the heterogeneous genomic data. The assumption of a close relationship between multilevel datasets and a linear relationship between the observed space and the latent space is a significant limitation of the Joint Bayes Factor.

4.2.2 Bayesian Approach

4.2.2.1 Bayesian Consensus Clustering (BCC)

Consensus clustering (CC) is a widely used methodology for combining multiple clustering algorithms, and it can also be used to integrate datasets from multiple sources (Jovanovski and Kocarev 2019). BCC is a versatile clustering method that can model the heterogeneity and dependence of multi-omics data using a finite Dirichlet mixture model. Separate clustering is formed for each omic data that is loosely connected to the overall clustering of multi-omics data to undergo post-hoc integration of separated clusters. BCC performs both omic-specific clustering and CC at the same time. Because BCC implementation assumes normally distributed data, CC is derived from the distribution with a higher probability to clusters that are present in specific regions. A heuristic approach is also provided for selecting the optimal number of clusters for a given set of omic data.

4.2.2.2 Multiple Dataset Integration (MDI)

Kirk et al.(Kirk et al. 2012) proposed multiple dataset integration (MDI) as a Bayesian method for unsupervised integrative modeling of multi-omics data. Each omic can be represented using the Dirichlet-Multinomial Allocation (DMA) mixture model, with the parameters describing their agreement defining their dependencies. Additionally, the models can be connected through the use of multilevel variables assigned to components (e.g., genomic features). Linkages at the component variable level aid in capturing the dependencies between multiple datasets. As a result, MDI can examine the shared data captured by parameters describing the agreement between multiple omics datasets. In different datasets, the same gene allocation affects each other. For instance, a group of genes assigned to the same MDI component can be clustered across all datasets.

Cho et al., like MDI, developed a probabilistic framework called Prob GBM to build a patient similarity network, where nodes represent patients and edges represent phenotypic similarities among patients (Cho and Przytycka 2013). This method illustrates the phenotypic similarities using features from multi-omics data rather than directly calculating phenotypic similarity between patients and providing potential explanatory features. Explanatory characteristics explain the phenotypic similarities of patients derived from gene expression data (e.g., mutations and miRNA expression, CNVs). As a result, each patient is defined as a combination of the genetic characteristics of each subtype, which is represented by a feature distribution. Finally, the most likely subtype is assigned to each patient. Prob GBM can be used to model relationships between gene expression similarity and a variety of genetic causes in general.

4.2.2.3 COpy Number and EXpression in Cancer (CONEXIC)

CONEXIC is a Bayesian network-based method developed by Akavia et al. from Module Networks (Segal et al. 2003). It is capable of integrating gene expression data and matched CNV (amplifications and deletions) exclusively (Akavia et al. 2010). This method employs a score-guided search algorithm to identify the combination of modulators that best explain the behavior of each module (gene expression) across all samples and seeks the highest score within deleted or amplified regions. The outputs include a ranked list of modulators that correlate with DEGs across samples and are present in deleted or amplified regions. Rather than identifying mutation drivers, CONEXIC aids in elucidating the probable roles of candidate drivers.

4.2.2.4 Multi-Omics Factor Analysis (MOFA)

MOFA, an unsupervised model-based method, can integrate multi-omics data from the same or partially overlapped samples (Argelaguet et al. 2018). This method can handle missing values and aid in determining the main sources of technical and biological variation in multiple datasets as a set of hidden factors. It provides a model formulation that supports the combination of different noise models to integrate different data types such as numerical and categorical data by using a probabilistic Bayesian framework. The use of linear models to represent relationships between datasets, on the other hand, can yield strong nonlinear relationships within and between omics.

4.2.2.5 Patient-Specific Data Fusion (PSDF)

Yuan et al. (Yuan et al. 2011) presented a non-parametric Bayesian model for the integration of gene expression data and CNVs based on a Dirichlet process hierarchy. This method is made up of three steps: a. extract concordance from discordant signals; b. select informative features; c. provide an estimate of the number of disease subtypes. The concordance of each sample's gene expression and CNVs was assessed, and a binary state was assigned based on their concordance. The samples that agreed were fused, while the others did not, resulting in patient-specific fusion models. The Markov chain Monte Carlo (MCMC) sampling method was used to predict the probability of each fused sample. PSDF feature selection reduces noise in datasets by selecting only those features that aid in clustering. The feature selection is also a binary indicator that is identified separately for each dataset. As a result, patient-specific data fusion results in patient-specific consistent fusion, which in turn determines the number of clusters. However, like CONEXIC, this method restricts the inputs to only two types of data (CNVs and gene expression), limiting its flexibility and application range.

4.2.3 Network-Based Methods

4.2.3.1 Similarity Network Fusion (SNF)

SNF builds individual networks per omic type and iteratively optimizes the networks to promote their similarity until they concentrate onto a single network using a nonlinear network fusion method (Wang et al. 2014). The fusion step, based on message-passing theory, makes the network more similar to the others within each iteration (KNN and graph diffusion). SNF, because it is based on sample networks, can scale a large number of genes to derive useful information even from a small cohort of samples while remaining robust to data heterogeneity and noise. The main reason is that noise (weak connections) is removed, leaving only the strong connections to converge. SNF, to its credit, does not impose data format constraints, allowing it to accommodate more data types. The only restriction is that the samples must be consistent across all datasets. In terms of sample classification, Morgane et al. claimed that SNF outperformed the other 12 unsupervised methods (Pierre-Jean et al. 2019).

4.2.3.2 Low-Rank Approximation Based Multi-Omics Data Clustering (LRAcluster)

Wu et al. proposed a method called LRAcluster that uses the low-rank approximation method to find the principal low-dimension subspace for classification of multi-omics data (Wu et al. 2015). Each omics data set is conditional on a size-matched parameter matrix in this method, and this low-rank parameter matrix can be represented in a low-dimensional space. The dimension parameter and the number of clusters aid in the rapid reduction of dimension and the better clustering of disease subtypes (Cantini et al. 2021). As a result, LRAcluster is a very useful tool for faster and more efficient unsupervised clustering of samples (Vaske et al. 2010).

4.2.3.3 Pathway Representation and Analysis by Direct Reference on Graphical Models (PARADIGM)

PARADIGM can be used to infer patient-specific genetic variations in a probabilistic graphical model framework by incorporating curated pathway interactions among genes (Vaske et al. 2010). Multiple genome-scale measurements (e.g., CNVs, gene expression) from a sample can be combined in the PARADIGM model to infer gene activities and products and summarize the inputs/outputs of a single National Cancer Institute (NCI) pathway. Each NCI pathway was converted to a distinct probabilistic model, represented by a factor graph with both hidden and observed states, during this process. A pathway was abstracted as an acyclic graph, with edges representing either positive or negative influence on downstream nodes, and nodes relying on the combined input signals. A gene was abstracted as a factor graph consisting of a set of interconnected variables representing a gene's expression level and activity, which were combined with other high-through measurements. The integrated pathway activity (IPA) scores can describe the specific measurement with an altered degree of alteration in a specific pathway as a result of PARADIGM. However, the PARADIGM pathways are measured independently without taking into account the interactions between pathways, which may reduce their accuracy and robustness.

4.2.3.4 NetICS

NetICS, using a graph diffusion-based model framework, can accommodate a wide range of data types, including but not limited to gene expressions, somatic mutations, CNVs, miRNA expressions, methylation patterns, and protein expressions (Dimitrakopoulos et al. 2018). The main advantage of this method is the ability to predict the effect of epigenetic changes, genetic aberrations, and miRNAs on downstream genes and proteins in the interaction network by identifying mediators that orchestrate downstream expression changes and are located between aberrant and DEGs. It employs a per-sample network-diffusion model on a directed functional interaction network to generate a population-level gene ranking by aggregating individual rankings, as well as a global ranking for all samples. The method ranks genes based on their proximity to upstream genetic anomalies and downstream differentially expressed genes. Proteins from each sample are then combined using a powerful rank aggregation technique.

4.2.3.5 Perturbation Clustering for Data INtegration and Disease Subtyping (PINS) and PINSPLUS

Nguyen et al. proposed a radically different unsupervised method named PINS to integrate multi-omic data including but not limited to gene expression, CNVs, DNA methylation, and noncoding microRNA (Strehl and Ghosh 2003) based on the resilience of patient connectivity and cluster ensembles (Nguyen et al. 2017). PINS computes a similarity matrix for one block of an omics dataset, followed by hierarchical clustering of the patients that are cut for each possible number of clusters. Then, for each partitioning in the clusters, this method employs a pairwise connectivity matrix. Then, by adding Gaussian noise to the original data, perturbed matrices are generated to evaluate the stability of clustering. PINS computes connectivity matrices and an average of the matrices using the perturbed datasets. Finally, if there are no significant differences between the original connectivity matrix and the average of perturbed connectivity matrices, it means that the perturbations do not affect the clustering results. A hierarchical structure search on the average connectivity matrix is used to address the heterogeneous subgroup of patients within a cluster, and the best number of subgroups has the smallest difference between the original and perturbed connectivity matrix. Furthermore, PINSPlus is reported to be much faster and more powerful when running on large omics datasets (Nguyen et al. 2018).

4.2.4 Multiple Kernel Learning Methods and Multi-Step Analysis-Based Methods

4.2.4.1 Feature Selection Multiple Kernel Learning (FSMKL)

The FSMKL method is a supervised classification method that uses the multiple kernel learning algorithm (Seoane et al. 2013). This new scheme measures the similarity of multi-omics data and computes a statistical score for feature selection per omic and pathway. FSMKL creates classifiers with a decision function based on a variety of different types of input data and pathway-based kernels. Each omic is treated as a base kernel, and composite kernels are created by linearly combining them. To further incorporate biological information into the algorithm, specific groups of genes, including membership in a KEGG pathway, are used to construct kernels independently. Statistical methods complete feature selection, allowing the most relevant kernels and associated features to be discovered. Following the feature selection steps, the most appropriate decision function over kernels is completed, which contributes to an integrative function over base kernels. Including clinical factors in the classifier's high-throughput profiles can improve prediction accuracy.

4.2.4.2 Regularized Multiple Kernel Learning Locality Preserving Projections (rMKL-LPP) & Web-rMKL

rMKL-LPP (Speicher and Pfeifer 2015) is an unsupervised version of Speicher and Pfeifer's machine-learning-based biomedical data fusion methods, which is an extension of the current multiple kernel learning method (MKL-DR) (Lin et al. 2011). Depending on the multiple kernel learning with a graph embedding framework algorithm called Locality Pre-serving Projections, the method can reduce dimensionality for the clustering of samples. One key feature of the method is the accommodation for numerical and sequence matrices and stability for small datasets, even several kernel matrices per omic. Moreover, as rMKL-LPP provides a variety of kernels per omic and different choices of dimension reduction methods, it claims to offer comparable results with much more flexibility (Zeng and Lumley 2018). Like mixOmics (Rohart et al. 2017) and BioNMF (Mejía-Roa et al. 2008) which provide web services, a web-server named web-rMKL (Röder et al. 2019) that employs rMKL-LPP had been established for the users' convenience.

4.2.4.3 CNAmet

Louhimo et al. proposed a cutting-edge multi-step integration method called CNAmet (Louhimo and Hautaniemi 2011) to exclusively integrate CNVs, DNA methylation, and gene expression data. The method assumes that gene upregulation occurs as a result of increased copy number and hypomethylation, whereas gene downregulation occurs as a result of decreased copy number and hypermethylation. CNAmet consists of three major steps: calculating signal-to-noise statistics to measure copy number and methylation aberrations relative to expression values (weight calculation); combining the weights to infer deterministic scores (score calculation), which aids in identifying the type of gene modification; and performing a permutation test on the combined scores and correcting the P-values (significance evaluation). During these steps, the genes regulated by methylation and CNVs work together to provide better characterization and a better understanding of biological processes. However, one significant limitation is that the sample set in all omics data must be the same.

4.2.4.4 Integrative Bayesian Analysis of Genomics Data (iBAG)

iBAG is a versatile supervised multi-step technique with an embedded hierarchical model for incorporating biologically meaningful data from an arbitrary number of omics platforms (Wang et al. 2012). This approach is composed of two stages: (1) The first step is to apply a regression model mechanistically to partition gene expression data into discrete segments, which include the CNV principal component, the methylation principal component, and unknown components. (2) The next stage is to create a clinical model. Clinical data, such as survival statistics and binary outcomes, are modeled as the result of joint regression based on the components identified in the first step regression. The Normal-Gamma (NG) prior is used to account for sparsity and facilitate effect size estimation.

4.3 Software for Multi-Omics Data Interpretation and Visualization

4.3.1 UCSC Xena

To meet the need for easy-to-use genomics visualization software, UCSC Xena was developed for both private datasets and large public repositories (Goldman et al. 2020). It has multiple advanced features over the UCSC Cancer Brower: (1) easy to use by installing a Xena Hub on users' computers; (2) view public and private data together on the same platform; (3) security advantages as no need to upload private data to a public server; (4) integration of multi-omics data by combining multiple hubs. Besides, statistical tools are embedded which allow the significance of associations and dynamic quantification (Sanborn et al. 2010; Goldman et al. 2014; Goldman et al. 2012).

4.3.2 LinkedOmics

LinkedOmics is the first multi-omics database that consists of mass spectrometry (MS)-based global proteomics data produced by the Clinical Proteomic Tumor Analysis Consortium (CPTAC). Three analysis modules (LinkFinder, LinkCompare and LinkInterpreter) were developed to comprehensively analyze and explore The Cancer Genome Atlas (TCGA) data, including 32 cancer types and a total of 11,158 patients. The LinkFinder module examines the associations between and within datasets and clinical interested attributes. The LinkCompare module performs a comparison of the associations in multi-omics and pan-cancer analyses identified by LinkFinder. The LinkInterpreter module decodes the identified associations into biological sense through network and pathway analysis.

4.3.3 NetGestalt

To address the challenge of increasing data complexity and network size, a web application named NetGestalt which can integrate multi-omics data over biological networks was developed by Zhiao et al. (Shi et al. 2013). To reduce the visualization complexity of large biological networks, this method places the nodes in a single horizontal dimension on the ground of the hierarchical modular architecture. Depending on the computational algorithm of the network seriation and modularization (NetSAM) package, NetGestalt uncovers the hierarchical organization of biological networks. Compared with other network visualization software, it provides multi-scale representation and navigation of the data, pathways, statistical analysis, and cross-omic comparisons. To facilitate data integration, it enables simultaneous visualization of different types of data within the same framework.

4.3.4 3Omics

The 3Omics is a web-based systems biology software that performs professional integration and comparative analysis of human transcriptomics, proteomics, and metabolomics (Kuo et al. 2013). If only two of the three omics data are available in running, this method automatically captures the missing protein, transcript or metabolite information by text-mining the PubMed literature. Five commonly used analysis methods are intensively combined: correlation analysis, phenotype mapping, co-expression profiling, GO enrichment analysis, and pathway enrichment analysis on each omic via a single platform. The main output of 3Omics are the inter-omic correlation networks, by which users can visualize the relationships within the omics data with respect to experimental conditions or time for all proteins, transcripts and metabolites.

4.3.5 Paintomics 3

Paintomics 3 is also a web-based software specially designed for integrated visualization and exploration of multi-omics data (Hernández-de-Diego et al. 2018). In addition to traditional transcriptomics, metabolomics, and proteomics, it supports region-based approaches such as ChIP-seq or ATAC-seq data. This method has a comprehensive KEGG pathway analysis workflow, including automatic pathway enrichment feature name/identifier conversion, multi-layered feature matching, trend charts, network analysis, interactive heatmaps, and so on. Moreover, it is rich in auxiliary and customization functions that enable id conversion data and job storage, re-coloring, filtering, rescaling. In different cases, this method creates the most informative representation of the multi-omics dataset.

4.3.6 MethHC & MethHC 2

The database of DNA Methylation and gene expression in Human Cancer (MethHC) is a portal for mRNA/microRNA expression profiles and a large collection of DNA methylation data from TCGA, comprising 18 cancer types, 6000 samples, 12,567 RNA sequencing and 6548 microarrays data (Huang et al. 2014). The specialty of MethHC is to implement identification of differentially methylated genes, clustering, and correlation analysis with gene expression. To interpret the results, three databases (UCSC genome browser, miRStart, and KEGG pathways) are integrated in this method. The updated version, MethHC 2, consists of 33 human cancers, 50,118 microarray and RNA sequencing data from TCGA and GEO (Huang et al. 2020). Besides, the key features including clinical-pathological data, mutations and CNVs, a multiplicity of information (enhancer regions, gene regions, and CGI regions), even circulating tumor DNA methylation profiles were supplemented into the multi-omic data.

4.4 Challenges of Multi-Omics Data Manipulation

Multi-omics data mining, colloquially referred to as “data munging,” frequently begins with a difficult and time-consuming data wrangling phase. To maintain uniformity across different sequencing platforms, a large amount of data should undergo data filtering, systematic normalization, batch effect removal, and quality checks prior to transformation and mapping. Typically, the former is used to harmonize disparate datasets via data normalization, scaling, and imputation. The latter can be accomplished through an ID harmonization process or a labor-intensive meta-data annotation process. Additionally, the majority of multi-omics software requires inputs in a particular format (sample X feature matrix). Thus, it is critical to use preprocessing steps carefully, as they may have a significant latent influence on the downstream analysis. A thorough orientation for newcomers regarding sample registration and robust metadata recording prior to data generation and analysis is extremely beneficial in mitigating errors and avoiding artificial waste.

Data heterogeneity is an obvious additional challenge when dealing with multi-omics data due to the variety of technologies and platforms used. For example, due to batch effects and other factors, even datasets generated from the same cohort of samples within a unified framework cannot be combined easily (Su et al. 2014). Additionally, because pre-processing steps for individual datasets have not been democratized, the software can be applied only to representative studies (Shen et al. 2009; Robinson et al. 2017; Pineda et al. 2015). Without a doubt, integrating non-omics data (e.g., clinical and epidemiological data) that contribute to the explanation of disease-related trait variation can likely improve the outcomes of omics-only algorithms (Lichtenstein et al. 2000). However, additional challenges arise when integrating omics and non-omics data, including the ability to integrate the two types of large-scale data and their relationship (i.e., ascertainment bias), the nature and heterogeneity of non-omics data, the fairness of the models (independent, conditional, and joint modeling), and the presence of interactions (López de Maturana et al. 2019).

In this contribution, we have enumerated widely used algorithms and software for data integration, analysis, visualization. The performance of the unsupervised software had been compared with both simulated and real data several times in (Tini et al. 2017; Bersanelli et al. 2016; Ritchie et al. 2015; Subramanian et al. 2020; Pucher et al. 2018; Chauvel et al. 2019). The reviews focusing on integrative clustering software are also reported sometimes (Chalise et al. 2014),. From a lack of an accurate and unifying 'Gold Standard' framework, robust conclusions do not come into being. In other words, no exiting software has the domination in the multi-omics data mining filed. The supervised software is indeed model-based and has the characteristics of regression analysis which identities the subset of only identity the subset of relevant omics features (e.g., candidate biomarkers) (Richardson et al. 2016; Ma et al. 2020).

4.5 Conclusions and Future Perspectives

The advantages of multi-omics data studies have been widely acknowledged in various domains of cancer, microbial, plant, biomedical and animal scientific research in the past decade. Currently, the integrative approaches coupled with computational challenges offer the opportunity to derive the most relevant biological insights such as disease subtyping, prognosis, and diagnosis. In this document, we collected typical algorithms and software in multi-omics data mining techniques, focusing on the data integration, analysis, and visualization. For data imputation, readers can refer to the review paper which summarizes the currently available imputation methods for handling missing values at the stage of data quality control with an emphasis on multi-omics imputation (Song et al. 2020). For computational environment sharing, new researchers can find guidance in (Krassowski et al. 2020). For the state-of-the-art single-cell multi-modal omics (scMulti-omics) studies (Ma et al. 2020), several integration software are also available, such as MAGAN, UnionCom (Cao et al. 2020), LIGER (Wilson et al. 2019), MOFA+ (Argelaguet et al. 2020). As easy-to-use guidance mainly for new researchers, such a summary can hardly accommodate all software matching this chapter, even for a single integration part. Thus, we had pruned the whole catalogue according to the following principles: (1) the semi-supervised software was excluded (e.g., GeneticInterPred (You et al. 2010)); (2) the software without installation source was no more considered (e.g., iPAC (Aure et al. 2013), MCD (Chari et al. 2010)); (3) others (e.g., CoxPath (Mankoo et al. 2011), MKGI (Kim et al. 2016)). Besides, we prefer the software with high citations.

To date, the most primal of motivations to code the software above is to solve the biological questions of interest. To get the best results, consideration of the sample type, environmental parameters, multi-omics data and integrative methods is vital before decision making. A crucial application is cancer subtyping or molecular classification, which serves as a reference to further advance the precision treatment (Robinson et al. 2017; Kim et al. 2016; Xiao et al. 2021; Kristensen et al. 2014). However, most of the cancer-derived multi-omics software is partly limited to the algorithms employed which results in its utilization only to specific scenarios such as fixed multi-omics types or poor statistical significance. Thus, it is quite worthy to compare the software in multiple large-scale multi-omics studies to confirm their strong point. Furthermore, the application to other complex diseases such as Alzheimer's and Parkinson's dementia should be meaningful.

Collectively, based on multi-omics big-data, hundreds of data mining algorithms and software have been developed, serving for diverse purposes. The application of these algorithms and software on multi-omics big-data have already revealed rich information about the regulation patterns in diverse biomedical objectives. And the details of these applications will be described in following chapters.

References

Akavia UD, Litvin O, Kim J, Sanchez-Garcia F, Kotliar D, Causton HC, Pochanard P, Mozes E, Garraway LA, Pe'er D. An integrated approach to uncover drivers of cancer. Cell. 2010;143 (6):1005–17.

Argelaguet R, et al. Multi-omics factor analysis-a framework for unsupervised integration of multi-omics data sets. Mol Syst Biol. 2018;14(6):e8124.

Argelaguet R, et al. MOFA+: a statistical framework for comprehensive integration of multi-modal single-cell data. Genome Biol. 2020;21(1):111.

Aure MR, et al. Identifying in-trans process associated genes in breast cancer by integrated analysis of copy number and expression data. PLoS One. 2013;8(1):e53014.

Bakal G, Kilicoglu H, Kavuluru R. Non-negative matrix factorization for drug repositioning: experiments with the repoDB dataset. AMIA Annu Symp Proc AMIA Symp. 2019;2020: 238–47.

Bersanelli M, et al. Methods for the integration of multi-omics data: mathematical aspects. BMC Bioinformatics. 2016;17(2):S15.

Biswas N, Chakrabarti S. Artificial intelligence (AI)-based systems biology approaches in multi-omics data analysis of cancer. Front Oncol. 2020;10:2224.

Cantini L, et al. Benchmarking joint multi-omics dimensionality reduction approaches for the study of cancer. Nat Commun. 2021;12(1):124.

Cao K, et al. Unsupervised topological alignment for single-cell multi-omics integration. Bioinformatics. 2020;36(Supplement_1):i48–56.

Carvalho CM, et al. High-dimensional sparse factor modeling: applications in gene expression genomics. J Am Stat Assoc. 2008;103(484):1438–56.

Chalise P, et al. Integrative clustering methods for high-dimensional molecular data. Transl Cancer Res. 2014;3(3):202–16.

Chari R, et al. An integrative multi-dimensional genetic and epigenetic strategy to identify aberrant genes and pathways in cancer. BMC Syst Biol. 2010;4(1):67.

Chauvel C, et al. Evaluation of integrative clustering methods for the analysis of multi-omics data. Brief Bioinform. 2019;21(2):541–52.

Chen M, et al. Predicting viral infection from high-dimensional biomarker trajectories. J Am Stat Assoc. 2011;106(496):1259–79.

Cho D-Y, Przytycka TM. Dissecting cancer heterogeneity with a probabilistic genotype-phenotype model. Nucleic Acids Res. 2013;41(17):8011–20.

Dimitrakopoulos C, et al. Network-based integration of multi-omics data for prioritizing cancer genes. Bioinformatics. 2018;34(14):2441–8.

Ghosh D, et al. Leveraging multilayered "omics" data for atopic dermatitis: a road map to precision medicine. Front Immunol. 2018;9:2727.

Gligorijević V, Panagakis Y, Zafeiriou S. Non-negative matrix factorizations for multiplex network analysis. IEEE Trans Pattern Anal Mach Intell. 2019;41(4):928–40.

Goldman M, et al. The UCSC cancer genomics browser: update 2013. Nucleic Acids Res. 2012;41 (D1):D949–54.

Goldman M, et al. The UCSC cancer genomics browser: update 2015. Nucleic Acids Res. 2014;43 (D1):D812–7.

Goldman MJ, et al. Visualizing and interpreting cancer genomics data via the Xena platform. Nat Biotechnol. 2020;38(6):675–8.

Gomez-Cabrero D, et al. Data integration in the era of omics: current and future challenges. BMC Syst Biol. 2014;8(2):I1.

Hernández-de-Diego R, et al. PaintOmics 3: a web resource for the pathway analysis and visualization of multi-omics data. Nucleic Acids Res. 2018;46(W1):W503–9.

Huang H-Y, et al. MethHC 2.0: information repository of DNA methylation and gene expression in human cancer. Nucleic Acids Res. 2020;49(D1):D1268–75.

Huang W-Y, et al. MethHC: a database of DNA methylation and gene expression in human cancer. Nucleic Acids Res. 2014;43(D1):D856–61.

Ichihashi Y, et al. Multi-omics analysis on an agroecosystem reveals the significant role of organic nitrogen to increase agricultural crop yield. Proc Natl Acad Sci U S A. 2020;117(25):14552–60.

Ickstadt K, Schäfer M, Zucknick M. Toward integrative Bayesian analysis in molecular biology. Annu Rev Stat Its Appl. 2018;5(1):141–67.

Jovanovski P, Kocarev L. Bayesian consensus clustering in multiplex networks. Chaos. 2019;29 (10):103142.

Karczewski KJ, Snyder MP. Integrative omics for health and disease. Nat Rev Genet. 2018;19(5): 299–310.

Kim D, et al. Using knowledge-driven genomic interactions for multi-omics data analysis: metadimensional models for predicting clinical outcomes in ovarian carcinoma. J Am Med Inform Assoc. 2016;24(3):577–87.

Kirk P, et al. Bayesian correlated clustering to integrate multiple datasets. Bioinformatics. 2012;28 (24):3290–7.

Kong L, et al. Multi-omics analysis based on integrated genomics, epigenomics and transcriptomics in pancreatic cancer. Epigenomics. 2020;12(6):507–24.

Krassowski M, et al. State of the field in multi-omics research: from computational needs to data mining and sharing. Front Genet. 2020;11:1598.

Kristensen VN, et al. Principles and methods of integrative genomic analyses in cancer. Nat Rev Cancer. 2014;14(5):299–313.

Kuo T-C, Tian T-F, Tseng YJ. 3Omics: a web-based systems biology tool for analysis, integration and visualization of human transcriptomic, proteomic and metabolomic data. BMC Syst Biol. 2013;7(1):64.

Lichtenstein P, et al. Environmental and heritable factors in the causation of cancer—analyses of cohorts of twins from Sweden, Denmark, and Finland. N Engl J Med. 2000;343(2):78–85.

Lin Y, Liu T, Fuh C. Multiple kernel learning for dimensionality reduction. IEEE Trans Pattern Anal Mach Intell. 2011;33(6):1147–60.

Liu H, et al. MODEM: multi-omics data envelopment and mining in maize. Database. 2016;2016

Lock EF, et al. Joint and individual variation explained (JIVE) for integrated analysis of multiple data types. Ann Appl Stat. 2013;7(1):523–42.

López de Maturana E, et al. Challenges in the integration of omics and non-omics data. Genes. 2019;10(3):238.

Louhimo R, Hautaniemi S. CNAmet: an R package for integrating copy number, methylation and expression data. Bioinformatics. 2011;27(6):887–8.

Ma A, et al. Integrative methods and practical challenges for single-cell multi-omics. Trends Biotechnol. 2020;38(9):1007–22.

Mankoo PK, et al. Time to recurrence and survival in serous ovarian tumors predicted from integrated genomic profiles. PLoS One. 2011;6(11):e24709.

Mantini G, et al. Computational analysis of Phosphoproteomics data in multi-omics cancer studies. Proteomics. 2021;21(3-4):e1900312.

Mejía-Roa E, et al. bioNMF: a web-based tool for nonnegative matrix factorization in biology. Nucleic Acids Res. 2008;36(suppl_2):W523–8.

Meng C, et al. Dimension reduction techniques for the integrative analysis of multi-omics data. Brief Bioinform. 2016;17(4):628–41.

Menyhárt O, Győrffy B. Multi-omics approaches in cancer research with applications in tumor subtyping, prognosis, and diagnosis. Comput Struct Biotechnol J. 2021;19:949–60.

Misra BB, et al. Integrated omics: tools, advances and future approaches. J Mol Endocrinol. 2019;62(1):R21–45.

Mo Q, et al. Pattern discovery and cancer gene identification in integrated cancer genomic data. Proc Natl Acad Sci. 2013;110(11):4245.

Nguyen H, et al. PINSPlus: a tool for tumor subtype discovery in integrated genomic data. Bioinformatics. 2018;35(16):2843–6.

Nguyen T, et al. A novel approach for data integration and disease subtyping. Genome Res. 2017;27 (12):2025–39.

O'Connell MJ, Lock EF. R.JIVE for exploration of multi-source molecular data. Bioinformatics. 2016;32(18):2877–9.

Pierre-Jean M, et al. Clustering and variable selection evaluation of 13 unsupervised methods for multi-omics data integration. Brief Bioinform. 2019;21(6):2011–30.

Pineda S, et al. Framework for the integration of genomics, epigenomics and transcriptomics in complex diseases. Hum Hered. 2015;79(3-4):124–36.

Pinu FR, et al. Systems biology and multi-omics integration: viewpoints from the metabolomics research community. Meta. 2019;9(4):76.

Pucher BM, Zeleznik OA, Thallinger GG. Comparison and evaluation of integrative methods for the analysis of multilevel omics data: a study based on simulated and experimental cancer data. Brief Bioinform. 2018;20(2):671–81.

Quinn RA, et al. From sample to multi-omics conclusions in under 48 hours. mSystems. 2016;1(2): e00038–16.

Rappoport N, Shamir R. Multi-omic and multi-view clustering algorithms: review and cancer benchmark. Nucleic Acids Res. 2018;46(20):10546–62.

Ray P, et al. Bayesian joint analysis of heterogeneous genomics data. Bioinformatics. 2014;30(10): 1370–6.

Richardson S, Tseng GC, Sun W. Statistical methods in integrative genomics. Annu Rev Stat Appl. 2016;3(1):181–209.

Ritchie MD, et al. Methods of integrating data to uncover genotype–phenotype interactions. Nat Rev Genet. 2015;16(2):85–97.

Robinson DR, et al. Integrative clinical genomics of metastatic cancer. Nature. 2017;548(7667): 297–303.

Röder B, et al. web-rMKL: a web server for dimensionality reduction and sample clustering of multi-view data based on unsupervised multiple kernel learning. Nucleic Acids Res. 2019;47 (W1):W605–9.

Rohart F, et al. mixOmics: an R package for 'omics feature selection and multiple data integration. PLoS Comput Biol. 2017;13(11):e1005752.

Sanborn JZ, et al. The UCSC cancer genomics browser: update 2011. Nucleic Acids Res. 2010;39 (suppl_1):D951–9.

Segal E, et al. Module networks: identifying regulatory modules and their condition-specific regulators from gene expression data. Nat Genet. 2003;34(2):166–76.

Seoane JA, et al. A pathway-based data integration framework for prediction of disease progression. Bioinformatics. 2013;30(6):838–45.

Shen R, Olshen AB, Ladanyi M. Integrative clustering of multiple genomic data types using a joint latent variable model with application to breast and lung cancer subtype analysis. Bioinformatics. 2009;25(22):2906–12.

Shi Z, Wang J, Zhang B. NetGestalt: integrating multidimensional omics data over biological networks. Nat Methods. 2013;10(7):597–8.

Song M, et al. A review of integrative imputation for multi-omics datasets. Front Genet. 2020;11: 1215.

Speicher NK, Pfeifer N. Integrating different data types by regularized unsupervised multiple kernel learning with application to cancer subtype discovery. Bioinformatics. 2015;31(12):i268–75.

Strehl A, Ghosh J. Cluster ensembles—a knowledge reuse framework for combining multiple partitions. J Mach Learn Res. 2003;3(null):583–617.

Su Z, et al. A comprehensive assessment of RNA-seq accuracy, reproducibility and information content by the sequencing quality control consortium. Nat Biotechnol. 2014;32(9):903–14.

Subramanian I, et al. Multi-omics data integration, interpretation, and its application. Bioinform Biol Insights. 2020;14:1177932219899051.

de Tayrac M, et al. Simultaneous analysis of distinct omics data sets with integration of biological knowledge: multiple factor analysis approach. BMC Genomics. 2009;10(1):32.

Thibaux R, Jordan M. Hierarchical beta processes and the Indian buffet process. J Mach Learn Res—Proceedings Track. 2007;2:564–71.

Tini G, et al. Multi-omics integration—a comparison of unsupervised clustering methodologies. Brief Bioinform. 2017;20(4):1269–79.

Tipping ME. Sparse bayesian learning and the relevance vector machine. J Mach Learn Res. 2001;1:211–44.

Vaske CJ, et al. Inference of patient-specific pathway activities from multi-dimensional cancer genomics data using PARADIGM. Bioinformatics. 2010;26(12):i237–45.

Wang B, et al. Similarity network fusion for aggregating data types on a genomic scale. Nat Methods. 2014;11(3):333–7.

Wang W, et al. iBAG: integrative Bayesian analysis of high-dimensional multiplatform genomics data. Bioinformatics. 2012;29(2):149–59.

Wilson CM, et al. Multiple-kernel learning for genomic data mining and prediction. BMC Bioinformatics. 2019;20(1):426.

Wu D, et al. Fast dimension reduction and integrative clustering of multi-omics data using low-rank approximation: application to cancer molecular classification. BMC Genomics. 2015;16(1): 1022.

Xiao W, et al. Toward best practice in cancer mutation detection with whole-genome and whole-exome sequencing. Nat Biotechnol. 2021;39(9):1141–50.

Yang Z, Michailidis G. A non-negative matrix factorization method for detecting modules in heterogeneous omics multi-modal data. Bioinformatics. 2015;32(1):1–8.

You Z-H, et al. A semi-supervised learning approach to predict synthetic genetic interactions by combining functional and topological properties of functional gene network. BMC Bioinformatics. 2010;11(1):343.

Yuan Y, Savage RS, Markowetz F. Patient-specific data fusion defines prognostic cancer subtypes. PLoS Comput Biol. 2011;7(10):e1002227.

Zeng ISL, Lumley T. Review of statistical learning methods in integrated omics studies (an integrated information science). Bioinform Biol Insights. 2018;12:1177932218759292.

Zhang S, et al. A novel computational framework for simultaneous integration of multiple types of genomic data to identify microRNA-gene regulatory modules. Bioinformatics. 2011;27(13): i401–9.

Zhang S, et al. Discovery of multi-dimensional modules by integrative analysis of cancer genomic data. Nucleic Acids Res. 2012;40(19):9379–91.

Part II
Applications of Multi-omics Analyses

Chapter 5
Multi-Omics Data Analysis for Cancer Research: Colorectal Cancer, Liver Cancer and Lung Cancer

Hantao Zhang, Xun Gong, and Min Tang

5.1 Introduction

Cancer is a disease mainly caused by the accumulation of mutations in two gene classes, which are proto-oncogenes and tumor suppressor genes (Weinberg 1996). With its incidence growing rapidly, cancer is regarded as an important obstacle to human life extension (Torre et al. 2016b). In terms of cancer deaths worldwide for both men and women, lung cancer, colorectal cancer, and liver cancer are top three cancer types (Sung et al. 2021).

Since the twentieth century, lung cancer started to become the most common cause of cancer death as well as the second most commonly occurring cancer in both men and women internationally (Alberg and Samet 2003). It also ranks the most frequently diagnosed cancer and the leading cause of cancer mortality in men (Sung et al. 2021). In the United States, every year the number of patients who die from lung cancer is higher than the combined death toll from colon, breast, and prostate cancer (Spiro and Silvestri 2005). Tobacco smoking is regarded as the leading cause of lung cancer (Salgia and Skarin 1998). Compared with non-smokers, smokers have a 20- to 30-fold increase in lung cancer risk (Minna et al. 2002). Hence, the

Publisher's Disclaimer: The work described has not been published before; that it is not under consideration for publication anywhere else; that its publication has been approved by all co-authors, if any, as well as by the responsible authorities—tacitly or explicitly—at the institute where the work has been carried out. The publisher will not be held legally responsible should there be any claims for compensation.

H. Zhang · M. Tang (✉)
School of Life Sciences, Jiangsu University, Zhenjiang, Jiangsu, China
e-mail: mt3138@ujs.edu.cn

X. Gong
Department of Rheumatology & Immunology, the Affiliated Hospital of Jiangsu University, Zhenjiang, China

K. Ning (ed.), *Methodologies of Multi-Omics Data Integration and Data Mining*,
Translational Bioinformatics 19, https://doi.org/10.1007/978-981-19-8210-1_5

industrialized countries, where the smoking prevalence first took place, have the highest lung cancer incidence rates (Alberg et al. 2005). Over the past several decades, because of tobacco control policies and smoking cessation, the smoking prevalence keeps decreasing in those countries (De Groot et al. 2018), thus the burden of lung cancer shifts to developing countries (Torre et al. 2016a). People also gain more knowledge in lung cancer biology. The majority of lung cancers have been divided into four histological types, which are small-cell lung cancer (SCLC) and three non-small-cell lung cancer (NSCLC) types including squamous cell carcinoma, adenocarcinoma, and large cell carcinoma (Wistuba and Gazdar 2006). However, the mortality rates of lung cancer still remain high (Barta et al. 2019), which might be explained by nonspecific symptoms of this disease at early stages (Van Meerbeeck et al. 2011). When seek medical treatment, most patients present with advanced disease which is nearly incurable (Patz et al. 2000).

Nowadays, colorectal cancer (CRC) is the second most common cause of cancer-related death worldwide and the third most common malignant disease (Center et al. 2009). Generally, colorectal cancer has been thought as a disease of the elderly, with rare people being diagnosed before 50, but it also strikes younger people (O'connell et al. 2004). In addition, colorectal cancer is the only type that strikes both men and women with approximately equal frequency (Potter 1999), since it is the second most common cancer in females and the third most common cancer in males (Siegel et al. 2014). What’s more, the incidence rates of colorectal cancer vary greatly around the world (Stintzing 2014). It is well-known that most cases of colorectal cancer are detected in western countries (Mármol et al. 2017), because people in longstanding developed countries often exhibit same factors playing important roles in the development of colorectal cancer, which might include obesity, unhealthy diet, smoking, alcohol consumption, and physical inactivity (Fearon 1995; Weinberg and Schoen 2014). However, in recent years, high incidence rates of CRC have been observed in newly developed countries where the risk of suffering from colorectal cancer was once quite low (Mármol et al. 2017).

Liver cancer is regarded as an aggressive and heterogeneous tumor which ranks the third most common cause of cancer-related death as well as the second leading cause of cancer-related death in man around the world (Yamashita and Wang 2013; Gao et al. 2019). In addition, liver cancer can be divided into primary liver cancer and secondary liver cancer in nature (Mckillop and Schrum 2005). As for primary liver cancer (PLC), based on different histological features, it can be categorized into six subtypes, which are hepatocellular carcinoma (HCC), intrahepatic cholangiocarcinoma (iCCA), mixed hepatocellular-cholangiocarcinoma (HCC-CCA), fibrolamellar HCC, and the pediatric neoplasm hepatoblastoma (Mcglynn et al. 2001; Srivatanakul et al. 2004). Among these histological types, HCC is the commonest primary liver cancer worldwide which accounts for nearly 90% of all cases of primary liver malignancies (Ariff et al. 2009). The second most frequent type of primary liver cancer is iCCA, the incidence rates increase steadily (Sia et al. 2017). What’s more, the incidence rates of liver cancer in different countries vary significantly (Bosch et al. 1999). Blaming for hepatitis B virus (HBV) infection, the East and Southeast Asia as well as the Middle and Western

Africa have the highest liver cancer rates (Bosch et al. 2004). Thanks to HBV vaccine, liver cancer incidence rates is decreasing in several highest-risk areas (Chen and Zhang 2011). However, in some low-risk western countries, the rates continue to increase. Risk factors such as obesity, cigarette smoking, hepatitis C virus (HCV) infection and chronic alcohol abuse are believed to be related to liver cancer in these areas (Bishayee 2014). Gender is another risk factor for liver cancer development, males are more susceptible than females, as the incidence rates of liver cancer among men is over twice that among women (Liu et al. 2015).

Through molecular and genetic studies of cancer, multiple biomarkers of colorectal cancer, liver cancer, and lung cancer have been identified (Zochbauer-Muller and Minna 2000; Bishayee 2014; Dienstmann et al. 2017). However, it is quite difficult to find diagnostic, prognostic, and therapeutic targets from these outcomes, and the morality is still high for patients all over the world (Chakraborty et al. 2018). With the advancement of high-throughput omics technologies, researchers are now able to study genomics, transcriptomics, proteomics, and phosphoproteomic data at the same time (Ahmed 2020). Although through analyzing single omics data set, one can observe the alternation and association of biological entities at that level, the interaction between multiple molecular layers cannot be fully assessed (Biswas and Chakrabarti 2020). Hence, in lung cancer, liver cancer, and colon cancer research, many multi-omics analyses have been conducted in order to gain a holistic view of the molecular dynamics underlying cancer progression and to make a progress in early detection and prognosis (Sun and Hu 2016). Also, because of the heterogeneous nature of cancer, different patient may have different clinical responses to the same treatment (Du and Elemento 2015). For this problem, multi-omics studies at an individual level have been conducted to develop precision cancer medicine (Ghosh et al. 2018; Mantini et al. 2021).

In this review, we introduced different types of omics data used in the research of colorectal cancer, liver cancer, and lung cancer. In addition, we summarized currently used technologies for high-throughput multi-omics data analysis. We also reviewed integrative analyses using genomic, epigenomic, transcriptomic, proteomic, and metabolomics data that helped reveal the molecular pathology of colorectal cancer, liver cancer, and lung cancer. Finally, we discussed challenges and envisioned the future of precision cancer medicine.

5.2 Various Multi-Omics Data Types and Selected Repositories

With the advent of sequencing technologies, biomolecules in a given biological samples can be identified and quantified at multiple omics levels (Das et al. 2020). Next-generation sequencing (NGS) is now frequently used for whole-genome or whole-exome sequencing (Behjati and Tarpey 2013). ChIP-seq (chromatin immunoprecipitation) and DNase1-seq (DNase I hypersensitive sites-sequencing) are used for detection of DNA-protein interactions. RNA-seq can be used to identify and quantify RNA molecules (Kim and Dekker 2018; Lu et al. 2019). As for proteomic and metabolomic study, mass-spectrometry based techniques are widely used (Domon and Aebersold 2006). Omics data generated by these techniques, including but not limited to genomic, epigenomic, transcriptomic, proteomic, and metabolomic, is together called as multi-omics data (Liu et al. 2019). There are several publicly accessible databases listed in references (Huang et al. 2017; Subramanian et al. 2020), which accommodate multiple omics data sets and serve as rich resources for understanding the etiology of human cancer.

5.2.1 DriverDB v3

The DriverDB database (http://ngs.ym.edu.tw/driverdb/) contains numerous exome-seq data that was extracted from The Cancer Genome Atlas (TCGA), The International Cancer Genome Consortium (ICGC), Prostate Cancer Genetics Project (PCGP), The Therapeutically Applicable Research to Generate Effective Treatments (TARGET), and published papers (Cheng et al. 2014). More exome-seq data as well as additional RNA-seq data from TCGA, ICGC, and published papers were added to updated DriverDB v2 (Chung et al. 2016). DriverDB v3, the latest version, incorporated not only new exome-seq and RNA-seq datasets but also copy number variation (CNV), methylation, and smRNA-seq datasets. By applying various bioinformatic tools it contains, users can identify abnormalities at multi-omics levels and discover driver genes and mutations (Liu et al. 2020a).

5.2.2 TCGA Portal

The Cancer Genome Atlas (TCGA) was launched by The National Institute of Health (NIH) in 2006 aiming to reveal genomic and epigenomic alternations associated with 32 types of human cancers (Wang et al. 2016). For each type of human cancer, various kinds of data including gene expression, exon expression, miRNA expression, protein expression, single nucleotide polymorphism (SNP), copy number variation (CNV), loss of heterozygosity (LOH), and DNA methylation has been generated and processed (Tomczak et al. 2015). The aforementioned data are stored in a free-access database, namely the TCGA Data Portal (https://tcga-data.nci.nih.gov/tcga/). Without a doubt, the wealth of TCGA data has led to the discovery of diagnostic biomarkers and development of new cancer therapies (Colaprico et al. 2016).

5.2.3 ICGC

The International Cancer Genome Consortium (ICGC; https://icgc.org/) mainly contains mutational genomic data in nearly 50 cancer types. The International Cancer Genome Consortium Data Portal (https://dcc.icgc.org) is a user-friendly platform which helps users visualize, analyze, and interpret cancer-related genetic, molecular, and clinical data it contains. This may lead to deeper understanding of tumor biology as well as development of better diagnostic methods and drugs (Zhang et al. 2019).

5.2.4 CCLE

In order to promote the translation of genetic and pharmacological data generated by cancer cell line studies into understanding of cancer progression and development of novel therapies, Cancer Cell Line Encyclopedia (https://portals.broadinstitute.org/ccle) was built by the collaboration between the broad Institute and the Novartis Institute (Barretina et al. 2012). The original release of CCLE contains a large-scale genomic data set from 947 human cancer cell lines and pharmacological profiling of 24 anticancer drugs across 479 of those cell lines. Later, whole genome sequencing, RNA-seq, miRNA profiling, and histone profiling were added to it (Nusinow et al. 2020).

5.2.5 LinkedOmics

The LinkedOmics database (http://www.linkedomics.org) contains mass spectrometry (MS)-based global proteomics data which was downloaded from the Clinical Proteomic Tumor Analysis Consortium (CPTAC). Multi-omics data including genomic, epigenomic, and transcriptomic data as well as clinical data for 32 TCGA cancer types which were downloaded from The Cancer Genome Atlas (TCGA) project were also added to this database. Aiming to allow users to analyze these data in detail, LinkedOmics provided three analysis modules, namely LinkFinder, LinkCompare, and LinkInterpreter. For each cancer cohort, the LinkFinder module allows user to find associations between an attribute of interest and all other attributes. These associations can be compared with query attributes through the LinkCompare module and interpreted through the LinkInterpreter module. The results are presented in the form of plot or heatmap, which may effectively help users gain biological understanding (Vasaikar et al. 2018).

5.2.6 RHPCG

Consisting of a group of kinases, hippo signaling pathway is a highly conserved pathway which plays important roles in controlling cell proliferation, apoptosis, and migration. Dysregulation of Hippo signaling pathway is involved in the initiation and progression of cancer, such as breast cancer, lung cancer and so on. The Regulation of the Hippo Pathway in Cancer Genome database (http://www.medsysbio.org/RHPCG) can serve as an open resource for visualizing alternations of Hippo pathway genes as well as understanding the roles of Hippo pathway in cancer, because RHPCG was designed to allow users easily search, view, and download alternations of core Hippo-protein-encoding genes in 33 cancer types at levels of genomics, epigenomics, and transcriptomics (Wang et al. 2019).

5.2.7 MOBCdb

The Multi-Omics Breast Cancer Database (http://bigd.big.ac.cn/MOBCdb/) was constructed in order to facilitate identification of breast cancer subtypes and discovery of novel biomarkers. MOBCdb contains SNV, gene expression, microRNA expression, DNA methylation, clinical, and drug response data that were downloaded from the TCGA data portal, GENECODE, miRBase, PharmGKB, and NCBI. With more than 10,000 files stored in the database, MOBCdb provides several methods to help users effectively gain information. In addition, by using the genome-wide browser in MOBCdb, users can visualize different omics data easily. The survival module was designed to help users find new biomarkers (Xie et al. 2018).

5.2.8 Target

The Therapeutically Applicable Research to Generate Effective Treatments database (https://ocg.cancer.gov/programs/target) was built by the cooperation of extramural and NCI investigators. TARGET originated with two pilot projects, now it contains the clinical information, gene expression, miRNA expression, copy number, and sequencing data of 24 molecular types of cancer. The effort of TARGET researchers has undoubtedly accelerated discoveries of genomic alterations in cancer and facilitated rapid translation of those findings into the clinic (Wu et al. 2021a).

There is much information that can be obtained from the data sets stored in the aforementioned databases. For instance, genomic studies can reveal the associations between tumorigenesis and genetic mutations (Ghosh et al. 2018). Also, Epigenomic data can lead to knowledge regarding how chemical modifications of DNA and protein drive tumorigenesis (Rhee 2018). Similarly, transcriptomic profiling can be used to detect the association between cancer and dysregulated genes (Canzler et al. 2020). Proteomic data can help researchers better understand its function in human cancer (Matthiesen and Jensen 2008). Because each omics data type only provides a partial view of the complexity of cancer, biological mechanisms can be fully captured only through integrating different omics data types (Hao et al. 2019).

5.3 Selected Integrative Tools for Multi-Omics Analysis

Cancer is a consequence of malfunction and alteration in multiple molecular layers (Hausman 2019). With decreasing time and cost to generate multiple omics datasets from biological samples, an increased need for large-scale omics analysis tools emerged to explore relationships between different biological readouts (Altenbuchinger et al. 2020). Usually, steps to conduct an integrative analysis of these readouts include data normalization, variable selection, cluster analysis, and dimensional reduction (Meng et al. 2016; Chauvel et al. 2020; Nicora et al. 2020). In this section, we review eight computational integrative tools that are capable of multi-omics data analysis. The first five tools were designed to reveal the biological mechanisms connecting identified key drivers and pathways to diseases. The remaining three tools can be used to discover new therapeutic interventions or support clinical decision making.

Integrative Omics Data Analysis (iODA) is a software for omics data analysis, which is written in Java and able to run on Windows or Linux operating systems. iODA can integrate and refine data generated by RNA-seq, miRNA-seq, and ChIP-seq, which leads to the revelation of complex pathogenesis of human cancer. There are six statistical methods included, namely Least Sum of Ordered Subset Squared, Cancer Outlier Profile Analysis, Maximum Ordered Subset T-statistics, Outlier Robust T-statistics, Outlier Sum, and t-test, which can be selected by users to process their input data. Then, differentially expressed genes and miRNAs as well as transcription factor binding sites are extracted for the following pathway enrichment analysis and consistency analysis. The dysfunctional molecules are mapped on the KEGG pathway, and the consistent molecular signatures are identified as key pathogenic factors in cancer. The source code as well as executable file of iODA can be downloaded at http://www.sysbio.org.cn/iODA for free (Yu et al. 2020).

The interactive tool for statistical analysis of omics and clinical data (IOAT in short) is a R and Python-based Windows application for analyzing and visualizing multi-omics and clinical data. IOAT is a user-friendly tool designed for non-programmers. It can perform feature screening, risk assessment, clustering, and survival analysis after reading a comma-separated value text file imported by users and preprocessing the multi-omics and clinical data contained in the file. All results are displayed in a report, which enables users to view the outcomes of each step and thus gain a better understanding of their data. Additionally, IOAT considers data breaches. After downloading an executable file from https://github.com/WlSun shine/IOAT-software, users can use this desktop tool without the need for network connectivity, ensuring the security of their personal data (Wu et al. 2021b).

MEXPRESS is a simple and user-friendly web tool for visualizing and interpreting multiple omics data that does not require clinical researchers to be programmers. Users can view gene expression, DNA methylation, and clinical data extracted from TCGA by entering a gene name and selecting a cancer type. MEXPRESS can also be used to conduct statistical analyses on these datasets and determine their correlation, which is extremely useful for biomarker discovery (Koch et al. 2015). While the core functions of MEXPRESS remain unchanged in the new version released in 2019, new data types, statistical methods, and options are included. All code is available for free download at https://github.com/akoch8/mexpress (Koch et al. 2019).

PROMO is a powerful and integrative Windows software written in Matlab that is designed to analyze large genomic and clinical datasets contained in multi-omics databases effectively. It includes several features such as data preprocessing, exploration and visualization, clustering, enrichment analysis, biomarker discovery, and classification of cancer subtypes. After importing a multi-omics dataset into PROMO, users can discover correlations between features at various multi-omics levels as well as the genes involved in biological differences, resulting in a better understanding of biological mechanisms and the discovery of new biomarkers. PROMO is freely accessible to the public at http://acgt.cs.tau.ac.il/promo/ (Netanely et al. 2019).

Chromatin structures, such as topologically associating domains (TAD) and TAD boundaries, are critical for gene expression regulation. Changes in the structure of chromatin may contribute to the progression of human cancer (Valencia and Kadoch 2019). PredTAD is a machine learning tool that uses the Gradient Boosting Machine (GBM) algorithm to predict 3D chromatin structures. It makes use of genomic and epigenomic data to predict and detect TAD boundary variants in normal and cancer cell genomes. Correlations between TAD boundary alternations and the expression of nearby genes can be identified using RNA-seq data analysis. Because genes located near altered boundaries may be involved in a cascade of oncogenic signaling pathways, PredTAD is an effective tool for transforming genomic and ChIP data into an understanding of the roles of chromatin structures in cancer progression. The source code for PredTAD is available at https://github.com/jchyr-sbmi/PredTAD/ (Chyr et al. 2021).

IOBR is a computational tool for interpreting multi-omics data; its application in immuno-oncology biological research has the potential to shed new light on tumor-immune interactions and accelerate the development of immunotherapies. It is composed of four functional modules: signature and tumor microenvironment (TME) estimation, phenotype estimation, mutation estimation, and module construction. IOBR is capable of identifying signature genes and phenotype-relevant signatures, analyzing signature-associated mutations, and building models using previously identified signatures. These models can be used to forecast therapy response, prognosis for cancer, and tumor resistance. The IOBR R package can be downloaded from https://github.com/IOBR/IOBR (Zeng et al. 2021).

DrugComboExplorer, a computational systems biology tool, predicts drug combinations for specific cancer types by integrating DNA-seq, RNA-seq, methylation, and gene copy number data. It processes multi-omics data from cancer patients, identifies driver signaling networks, and quantifies the efficacy of combinatorial drugs on these networks using multiple algorithms. Combinations of optimal drugs that target driver signaling networks may be a way to copy resistance progression. The source code for DrugComboExplorer is available at https://github.com/Roosevelt-PKU/drugcombinationprediction (Huang et al. 2019).

OncoPDSS is a system that interprets multi-omics variants detected in cancer samples as supporting evidence for clinical pharmacotherapy decision-making. It contains the OncoPDSS knowledgebase (OncoPDSSkb), which was created to store data on drug-drug interactions, clinical trials for cancer, and drug indications. OncoPDSS imports user-uploaded variants. It uses a classification strategy to determine whether pharmacotherapies are potentially effective or not based on OncoPDSSkb mutation records, cancer records, and drug records that serve as oncology pharmacotherapy evidence. As a result, this tool will significantly aid clinicians and physicians in making clinical decisions, while also providing cancer researchers with novel treatment strategies. OncoPDSS is accessible via the following link: https://oncopdss.capitalbiobigdata.com (Xu et al. 2020a).

Recent cancer projects as well as multi-omics databases provide the research community with a wealth of omics data and clinical information on cancer patients (Cieslik and Chinnaiyan 2020). Integrative analysis of these data is challenging and requires bioinformatics, statistical, and programming skills (Chakraborty et al. 2018; Park et al. 2020). Numerous tools have been built to solve this problem. However, some limitations still exist. For instance, iODA only supports the analysis of mRNA, miRNA, and ChIP-seq data (Yu et al. 2020). Efforts should be devoted to develop new tools that can be applied for all omics data types. In addition, several tools utilize the R language, which is not friendly for researchers with limited biostatistical or programming knowledge (Eicher et al. 2020; Graw et al. 2021). Web-based interfaces should be created to allow fundamental researchers to leverage the merits of multi-omics tools.

5.4 Overview of Cancer Multi-Omics Research

5.4.1 Lung Cancer

Lung cancer is a highly complex and heterogeneous disease (De Sousa and Carvalho 2018). In recent decades, cancer researches focusing on the discovery of prognostic indicators and therapeutic targets have already been made (Jones and Baldwin 2018). Li proposed a novel method for mining cancer-related gene modules based on multi-omics data. First, genome-wide regulatory networks were constructed using key regulatory factors identified by feature selection method. Second, dysregulated gene sets were identified by comparing regulatory networks in variant and non-variant samples, which were then used to generate cancer-related gene modules. This new mining method has been proved to be applicable to lung cancer research (Li et al. 2019). By analyzing genomic, transcriptomics, and proteomic data, Kong et al. identified abnormal expressed membrane proteins in highly metastatic lung cancer cells. The high expression level of *CDH2*, *EGFT*, *ITGA3*, *ITGB1*, *ITGA5* and low expression level of *CALR* were found to be associated with cancer metastasis (Kong et al. 2020).

Small cell lung cancer (SCLC) and non-small cell lung cancer (NSCLC) are two major types of lung cancer (Wu et al. 2020). Patients diagnosed as NSCLC accounts for nearly 85% of all lung cancer patients, which makes NSCLC the most common histological type of lung cancer (Wang et al. 2018). Chen et al. performed gene expression, prognosis, DNA methylation, and gene mutation analysis of *NUF2* gene. It was shown that that the more *NUF2* expressed, the poorer prognosis patients had. Thus, *NUF2* might be considered as a prognostic biomarker of NSCLC and can be used for cancer treatment (Chen et al. 2014). Luan et al. integrated DNA methylation, RNA, miRNA and DNA copy number data to construct a survival risk model. Based on this, the chromosome regions 17q24.3 and 11p15.5 were identified as the copy number variation regions that were associated with NSCLC patient survival (Luan et al. 2020).

NSCLC can be further divided into three main subtypes, lung adenocarcinoma (LUAD), lung squamous cell carcinoma (LUSC), and large cell carcinoma (LCC) (Herbst et al. 2018).

Numerous potential biomarkers have been identified as a result of advancements in the molecule biology of LUAD. Paula et al. used proteomic data, ChIP-seq and RNA-seq assays to demonstrate that *MGA* gene, which is mutated and copy number deleted in LUAD, acts as a tumor suppressor by repressing genes activated by the MYC pathway. This discovery may open new therapeutic avenues (Llabata et al. 2020). Zhang et al. estimated different tumor microenvironment infiltration patterns and the correlation between these patterns and the genetic or epigenetic alterations by analyzing expression, RNA-seq, WES, and DNA methylation profiles. A prognosis model was constructed using the detected genetic and epigenetic alternations, which may aid in the development of a more accurate prognostic predictor for human LUAD (Zhang et al. 2020b). Ken built a SVM to subclass patients based on their survival based on clinical data from LUAD. By combining RNA expression and miRNA expression data of these subtypes, six genes were efficiently identified to be associated with LUAD patient survival: *ERO1B*, *DPY19L1*, *NCAM1*, *RET*, *MARCH1*, and *SLC7A8* (Asada et al. 2020). Lee et al. applied mRNA, miRNA, DNA methylation and CNV data to develop a deep learning autoencoding approach for survival risk stratification. They successfully identify significant prognostic difference between two groups of LUAD patients using this model (Lee et al. 2020).

LUSC has a worse prognosis than LUAD (Zhang et al. 2020a). Numerous studies have already been conducted to ascertain the molecular characteristics of this subtype. According to Zhang, an integrative analysis of methylation and gene expression data revealed that 113 methylation features and 23 gene expression features are strongly associated with lung cancer. SFTA3 and LPP may serve as molecular markers for subtyping NSCLC (Zhang et al. 2020a). Additionally, Xu et al. investigated the gene expression changes associated with DNA copy number or DNA methylation in LUSC patients by integrating genomic, transcriptomic, and epigenetic data. Seven genes expressed at a high level, which could be due to CNV or methylation and result in a poor prognosis (Xu et al. 2020c). Additionally, Hu et al. examined multi-omics differences between LUSC patients with high and low levels of programmed death 1 expression (PD1). It was discovered that 178 genes involved in immunity were significantly upregulated in the high expression group, which may contribute to a better understanding of the relationship between PD1 and immunotherapy effect (Hu et al. 2020).

Pulmonary sarcomatoid carcinomas (PSC) is a rare tumor in the family of NSCLC (Antoine et al. 2016). Yang et al. conducted multi-omics analysis of PSC samples and found out that PSC may be converted from the epithelial components and can be divided into five subtypes based on different histological morphologies (Yang et al. 2020b). Also, it was delineated that a large portion of patients had mutations in the p53, RTK/RAS, and PI3K pathways, suggesting that targeted therapy could be an option for patients with PSC (Yang et al. 2020b). Totally, their study shed light on the biological nature and brought entry points for the treatment of this rare malignancy (Yang et al. 2020b).

5.4.2 Colorectal Cancer

Colorectal cancer (CRC) is a heterogeneous disease (Berg et al. 2017; Almusawi et al. 2021). Various studies performed in recent years have provided insights into the molecular characteristics of CRC. Xu et al. explored genes related to CRC prognosis and incidence (Xu et al. 2020b). Genes annotated with single nucleotide mutation sites, copy number variation sites, and methylation sites along with differentially expressed genes were identified as candidate genes (Xu et al. 2020b). Moreover, a weighted gene co-expression network analysis was performed to search for hub genes (Xu et al. 2020b). Finally, *LRRC26* and *REP15* were identified as CRC-specific driving genes (Xu et al. 2020b). Yuan et al. attempted to link genetic variants, genes, and risk of CRC (Yuan et al. 2021). They conducted expression quantitative trait loci (eQTL) analysis, meta-analysis, and methylation quantitative trait loci (mQTL) analysis of 131 lead SNPs to explore potential target genes (Yuan et al. 2021). In addition, a colocalization analysis of genes identified in the previous step was performed, which revealed 66 putative susceptibility genes in CRC (Yuan et al. 2021). Ayiomamitis et al. investigated the roles of cyclooxygenase 2 (COX-2), an enzyme that promotes prostaglandin E2 (PGE2) production, and human telomerase reverse transcriptase (hTERT), a component of telomerase, in the onset of CRC (Ayiomamitis et al. 2019). By analyzing the expression levels of COX-2, PGE2, and hTERT along with telomerase activity, they demonstrated that COX-2 plays a key role in the initial stages of CRC development (Ayiomamitis et al. 2019). Also, high COX-2 expression was found to be associated with low hTERT expression and a better survival among CRC patients (Ayiomamitis et al. 2019). To gain a better understanding of the clinical relevance between obesity and CRC, Holowatyj et al. performed transcriptomic analysis on visceral adipose and tumor tissues and metabolomics analysis on blood samples of CRC patients (Holowatyj et al. 2020). Combining results generated by each omics measurement, they elucidated that glycolytic metabolism, GPVI signaling, and fibrosis participated in the adipose-tumor crosstalk and could promote CRC development (Holowatyj et al. 2020). Ghaffari et al. investigated the underlying mechanisms that drive metastatic progression (Ghaffari et al. 2021). They performed RNA-seq, ChIP-seq, and ATAC-seq on a CRC cell line (Ghaffari et al. 2021). Then, a statistical model was used to comprehensively analyze these multi-omics profiles along with TF-DNA binding information (Ghaffari et al. 2021). It was elucidated that JunD, a TF, plays a crucial role in CRC migration and invasion (Ghaffari et al. 2021).

It is widely accepted that most colorectal cancers arise as a result of transformation from adenoma to adenocarcinoma (Lam et al. 2016), which is triggered by the stepwise accumulation of genetic and epigenetic mutations (Aarons et al. 2014). Using the deep learning framework, Lv et al. constructed a prognostic model for patients with colon adenocarcinoma (COAD) using the TCGA and GEO databases (Lv et al. 2020). After applying this model to the TCGA dataset, it was discovered that two subgroups with significantly different survival rates existed. Further analysis of these two subgroups revealed 1217 differentially expressed genes and ten differentially expressed miRNAs, which may aid in deciphering the mechanisms underlying COAD development (Lv et al. 2020). Yin et al. proposed an approach to detect potential prognosis risk biomarkers (PRBs) (Yin et al. 2020). First, based on gene expression, exon expression, DNA methylation, and somatic mutation profiles along with clinical information of COAD patients, the multi-omics-based prognostic analysis (MPA) model was used to select features closely related to the prognosis of COAD patients (Yin et al. 2020). Second, they applied the protein-protein interaction (PPI) network to annotate the functions of these features (Yin et al. 2020). Finally, 13 features were identified as PRBs through the further validation, which may serve as drug targets in COAD treatment (Yin et al. 2020).

CRC is also known as bowel and colon cancer, which makes colon cancer (CC) a subset of it (Jahanafrooz et al. 2020). Tong et al. successfully constructed a prognostic prediction model of CC patients by integrating clinical features, gene expression, miRNA expression, and DNA methylation data extracted from TCGA (Tong et al. 2020). Compared with models based on clinical and gene expression data, this integrative prognostic model was more effective, suggesting that the more types of omics data integrated, the better the cancer prognostic model would perform (Tong et al. 2020). Yang et al. also established a prognostic model for CC (Yang et al. 2020a). They first conducted an identification of differentially methylated genes, differentially expressed genes and miRNAs between tumor samples and normal samples (Yang et al. 2020a). Then, using omics features correlated with prognosis, the prognostic model was built, which might be helpful for CC research (Yang et al. 2020a). Yi et al. explored the underlying mechanisms of Wnt/β-catenin signaling regulating EMT program (Yi et al. 2020). It was validated that the RUNX2 expression activated by Wnt signaling pathway would lead to an increase in the expression of EMT-associated genes (Yi et al. 2020). Because EMT has been proved to be highly correlated with metastasis formation and tumorigenesis (Pastushenko and Blanpain 2019), RUNX2 might serve as a prognostic biomarker for CC. Arora et al. detected the dysregulated expression pattern of seven classical non-homologous end joining (c-NHEJ) pathway genes in CC (Arora et al. 2020). Compared to normal tissues, *XRCC5*, *XRCC6*, *PRKDC*, and *PAXX* were observed to be overexpressed in tumor tissues, whereas the expression level of *LIG4* and *NHEJ1* were downregulated (Arora et al. 2020). In addition, *PAXX* was identified as a prognostic biomarker (Arora et al. 2020). Thus, their study may help reveal the clinical significance of c-NHEJ pathway genes in CC. Using a novel upstream analysis strategy, Kel et al. deciphered the molecular mechanisms of the resistance to methotrexate (MTX) in CC (Kel et al. 2016). This strategy mainly contains two steps, i.e., the identification of transcription factors (TFs) and master regulators that activate these TFs (Kel et al. 2016). After applying this approach to transcriptomics, proteomics, and ChIP-seq data, PKC-alpha, TGF-alpha, TGF-beta, and alpha9-integrin were identified as anti-resistance targets (Kel et al. 2016). Their findings would provide new insight into oncology drug resistance research.

Left-sided colon cancer (LCC), which originates from the hindgut, and right-sided colon cancer (RCC), which originates from the midgut, are two subtypes of CC (Song et al. 2020). In addition to the different tumor locations, there are many differences between them (Shen et al. 2015). To gain a better understanding of these differences, Huang et al. analyzed transcriptomics, clinical, and somatic mutation data of patients with CC (Huang et al. 2021). A total of 360 differentially expressed genes were observed (Huang et al. 2021). Among them, it was indicated that *BRAF* and *KRAS* mutations were frequently presented in RCC, whereas *APC* mutation was frequently presented in LCC (Huang et al. 2021). In addition, the 4-mRNA and 6-mRNA were identified as prognostic signatures for LCC and RCC, respectively (Huang et al. 2021). Similarly, Hu et al. conducted a study on the differences in molecular features between LCC and RCC (Hu et al. 2018). It was revealed that *PARC* was hypermethylated in RCC, whereas *CDX2* was hypermethylated in LCC (Hu et al. 2018). Also, the expression levels of miR31, miR155, and miR625 were observed to be upregulated in RCC, whereas the expression levels of miR-296 and miR592 were downregulated in LCC (Hu et al. 2018). In addition, compared with LCC, the mutation rate of *KRAS* and *BRAF* was higher in RCC, which was believed to be associated with a worse prognosis (Hu et al. 2018). Yi et al. performed a systematic analysis on the regulatory mechanisms between gene mutations and tumor immune microenvironment (TIME) in LCC and RCC cells (Yi et al. 2021). It was revealed that the mutations of top mutated genes were strongly correlated with TIME, DNA methylation levels of some immune checkpoints, and immune-related genes and miRNAs in RCC. However, these associations were less significant in LCC (Yi et al. 2021).

5.4.3 Liver Cancer

Liver cancer, one of the extraordinarily heterogeneous diseases, is caused by the interplay of various internal and environmental factors (Li and Wang 2016; Marengo et al. 2016). The development of omics strategies has helped us gain a holistic view of tumor biology. Shen et al. distinguished two molecular subtypes by analyzing genomic, epigenomic, and transcriptomic data from patients with liver cancer (Shen et al. 2021b). In addition, two prognostic molecular targets, ANXA2 and CHAF1B, were highly expressed in tumor tissues and identified to be strongly related to the prognosis of liver cancer patients (Shen et al. 2021b). Their research findings could provide new insight into the exploration of key biomarkers and mechanisms of liver cancer (Shen et al. 2021b).

Primary liver cancer is a serious public health issue, with HCC as the most common pathological subtype (Lin et al. 2016). Significant effort has been made to reveal the biological nature of HCC. Based on multi-omics datasets of HCC samples downloaded from TCGA and GEO databases, Liu et al. conducted an investigation on the methyltransferase-like 3 (METTL3) as well as methyltransferase-like 14 (METTL14), which were both core molecules of a multicomponent methyltransferase complex (MTC) that catalyzed the formation of N6-methyladenosine (m6A) (Liu et al. 2020b). It was clarified that METTL3 and METTL14 influence distinct signaling pathways and biological processes, thus may play opposite regulatory roles in HCC (Liu et al. 2020b). Using several databases, Jin et al. investigated the impact of the expression levels of *CDK1*, *CCNB1*, and *CCNB2* in the survival of HCC patients (Zou et al. 2020). The upregulation of *CDK1*, *CCNB1*, and *CCNB2*, which might be caused by low levels of methylation or genomic alternations, was found to be highly correlated with poor prognosis in HCC patients (Zou et al. 2020). Using multi-omics analysis of metabolomics and absolute quantification proteomics, Dan et al. conducted an investigation on the effects of canagliflozin (CANA) on the proliferation of HCC cell lines (Nakano et al. 2020). It was shown that CANA, the sodium glucose co-transporter 2 (SGLT2) inhibitor, mainly altered oxidative phosphorylation metabolism, fatty acid metabolism, and DNA synthesis, which may suppress cell proliferation of Hep3B and Huh7 cells (Nakano et al. 2020). Shen et al. performed a multi-omics analysis to explore the metabolic impact of estrogen and its receptors in HCC cells (Shen et al. 2021a). It was suggested that estrogen acts on its receptors to suppress HepG2 cell growth via altering glucose and lipid metabolism, which might be part of the reason why women have a lower risk of HCC development as compared to men worldwide (Shen et al. 2021a). Woo et al. integrated CNV, DNA methylation, and mRNA expression data of a cohort of HCC patients to identify DNA copy-number-correlated (CNVcor) and methylation-correlated (METcor) genes (Woo et al. 2017). The frequencies of CNVcor gene aberration were indicated to be significantly correlated with frequencies of METcor gene aberration, demonstrating that the concomitant regulation of transcriptomes by alternations in DNA copy numbers and methylation should be took into consideration in liver cancer research (Woo et al. 2017).

In developing countries prevalent for hepatitis B virus (HBV) infection, HBV still remains the most common etiologic agent of HCC (Chang 2014). Much work also has been done to uncover the direct and indirect mechanisms that are involved in HCC oncogenesis by HBV (Xie 2017). Through the integration of proteomics and metabolomics assays, Xie et al. conducted an exploration on the mechanisms of HBV-induced HCC (Xie et al. 2017). They demonstrated that HBV core protein might contribute to the progression of HCC by modifying the metabolism of glycolysis and amino acid (Xie et al. 2017). Consequently, HBV core protein could represent a promising target for antiviral therapy (Xie et al. 2017). Aiming to identify novel biomarkers in HCC, Miao et al. performed multi-omics analyses integrating genomic, transcriptomics, and clinicopathological data of patients with HBV-related multifocal HCC (Miao et al. 2014). Six genes with abnormal expression levels were identified (Miao et al. 2014). Among them, *TTK* might be an overall prognostic indicator for HCC, because the expression level of *TTK* was shown to be highly correlated with metastatic potential, postsurgical recurrence, and survival of HCC patients (Miao et al. 2014). Gao et al. conducted a comprehensive proteogenomic characterization of tumor and adjacent liver samples from 159 HCC patients with HBV infection (Gao et al. 2019). Two metabolic enzymes, PYCR2 and ADH1A, were identified to participate in HCC metabolic reprogramming (Gao et al. 2019). Because the upregulation of PYCR2 or downregulation of ADH1A may result in HCC progression, they were also validated as potential prognostic biomarkers (Gao et al. 2019).

Since accurate stratification is essential for clinical decision making (Preisser et al. 2020), different stratification methods applied to cohorts of HCC patients have been developed. Kumardeep et al. proposed a deep learning-based model derived from RNA-seq, miRNA-seq, CpG methylation and clinical data of HCC samples to identify two subgroups with significantly different survival (Chaudhary et al. 2018). It was illuminated that the more aggressive subgroup is associated with *TP53* inactivation mutations and Wnt pathway activation (Chaudhary et al. 2018). Therefore, this risk stratification model may be useful at HCC prognosis prediction as well as therapeutic intervention (Chaudhary et al. 2018). Xiao et al. formed an integration method and used this method for an analysis of mRNA expression data, DNA-methylation data, somatic mutation data, and clinical information of HCC samples (Ouyang et al. 2020). 34 differentially expressed genes (DEGs) were identified, some of them were verified as diagnostic biomarkers for HCC (Ouyang et al. 2020). According to the gene expression data of the aforementioned DEGs, tumor samples were divided into three subtypes that displayed different biological processes (Ouyang et al. 2020). Hence, what they found out might help improve precision medicine regarding HCC (Ouyang et al. 2020).

The advanced molecular biological techniques as well as improving understanding of complex mechanisms of liver cancer has driven the development of precision medicine (Yoo et al. 2018). Yildiz analyzed datasets generated by high-throughput drug screening and genomic and transcriptomic studies on HCC cell lines (Yildiz 2018). He divided HCC cells into two subtypes that responded differently to the same drug treatments (Yildiz 2018). 6 molecular targets were revealed to be associated with drug sensitivity, which could aid the development of effective molecular therapies (Yildiz 2018). Also, the EGFR/PI3K/AKT/mTOR signaling pathway was believed to play a central role in the regulation of sensitivity and resistance to drug treatments in HCC (Yildiz 2018). Christos et al. utilized a computational approach to explore the novel drug targets in mTOR-driven HCC (Dimitrakopoulos et al. 2021). 74 mediators under the impact of upstream genetic aberrations and changes in miRNA expression were identified, among which YAP1, GRB2, HDAC4, SIRT1, and LIS1 were validated to be dysregulated in human HCC (Dimitrakopoulos et al. 2021). Thus, inhibitors of these mediators may be potentially useful in HCC treatment (Dimitrakopoulos et al. 2021).

5.5 Conclusion

Multiomics clearly has advantages when it comes to translating the biological characteristics of cancer into understandable and clinically interpretable data. The advancement of multiomics research in the context of a specific cancer reveals numerous “invisible” but critical correlations. Multiple biomarkers have a higher specificity than previous single-gene markers, laying the groundwork for future research in this field. The identification of specific markers enables the diagnosis of cancer and subsequent treatment, as well as better stratifying patients and developing more effective and personalized treatment methods.

As mentioned previously, multi-omics methods have been successfully applied to colorectal cancer, liver cancer, and lung cancer, yielding a wealth of biological data. As methods and resources for multi-omics analysis mature, multi-omics research will play an increasingly important role in understanding the pathogenesis of cancer and developing effective treatment measures.

However, there is a growing gap between the ability to integrate, process, and interpret data and the ability to generate large amounts of omics data. The majority of data standardization efforts and development of a central public database of omics data have been abandoned. Simultaneously, the majority of tools for multi-omics integration are insufficiently robust, prone to errors, and are only suitable for advanced users with programming expertise. There is still a long way to go before multi-omics analysis is widely applied and its value is maximized in cancer research.

Acknowledgments We thank all the individuals who have helped us in this study. We acknowledge the valuable work of the many investigators whose published articles we were unable to cite owing to space limitations.

Conflict of Interest

The authors declare that they have no conflict of interest.

Fundings This work was financially supported by the grants from the National Natural Science Foundation of China (31861143051, 31872425, 32002235, 31602008) and Senior Talent Foundation of Jiangsu University (19JDG039).

References

Aarons CB, Shanmugan S, Bleier JI. Management of malignant colon polyps: current status and controversies. World J Gastroenterol. 2014;20:16178–83.

Ahmed Z. Practicing precision medicine with intelligently integrative clinical and multi-omics data analysis. Hum Genomics. 2020;14:35.

Alberg AJ, Brock MV, Samet JM. Epidemiology of lung cancer: looking to the future. J Clin Oncol. 2005;23:3175–85.

Alberg AJ, Samet JM. Epidemiology of lung cancer. Chest. 2003;123:21S–49S.

Almusawi S, Ahmed M, Nateri AS. Understanding cell-cell communication and signaling in the colorectal cancer microenvironment. Clin Transl Med. 2021;11:e308.

Altenbuchinger M, Weihs A, Quackenbush J, Grabe HJ, Zacharias HU. Gaussian and mixed graphical models as (multi-)omics data analysis tools. Biochim Biophys Acta Gene Regul Mech. 2020;1863:194418.

Antoine M, Vieira T, Fallet V, Hamard C, Duruisseaux M, Cadranel J, Wislez M. Pulmonary sarcomatoid carcinoma. Ann Pathol. 2016;36:44–54.

Ariff B, Lloyd CR, Khan S, Shariff M, Thillainayagam AV, Bansi DS, Khan SA, Taylor-Robinson SD, Lim AKP. Imaging of liver cancer. World J Gastroenterol. 2009;15:1289–300.

Arora M, Kumari S, Singh J, Chopra A, Chauhan SS. PAXX, not NHEJ1 is an independent prognosticator in colon cancer. Front Mol Biosci. 2020;7:584053.

Asada K, Kobayashi K, Joutard S, Tubaki M, Takahashi S, Takasawa K, Komatsu M, Kaneko S, Sese J, Hamamoto R. Uncovering prognosis-related genes and pathways by multi-omics analysis in lung cancer. Biomol Ther. 2020;10

Ayiomamitis GD, Notas G, Vasilakaki T, Tsavari A, Vederaki S, Theodosopoulos T, Kouroumalis E, Zaravinos A. Understanding the interplay between COX-2 and hTERT in colorectal cancer using a multi-omics analysis. Cancers (Basel). 2019;11:1536.

Barretina J, Caponigro G, Stransky N, Venkatesan K, Margolin AA, Kim S, Wilson CJ, Lehar J, Kryukov GV, Sonkin D, Reddy A, Liu M, Murray L, Berger MF, Monahan JE, Morais P, Meltzer J, Korejwa A, Jane-Valbuena J, Mapa FA, Thibault J, Bric-Furlong E, Raman P, Shipway A, Engels IH, Cheng J, Yu GK, Yu J, Aspesi P Jr, De Silva M, Jagtap K, Jones MD, Wang L, Hatton C, Palescandolo E, Gupta S, Mahan S, Sougnez C, Onofrio RC, Liefeld T, Macconaill L, Winckler W, Reich M, Li N, Mesirov JP, Gabriel SB, Getz G, Ardlie K, Chan V, Myer VE, Weber BL, Porter J, Warmuth M, Finan P, Harris JL, Meyerson M, Golub TR, Morrissey MP, Sellers WR, Schlegel R, Garraway LA. The cancer cell line Encyclopedia enables predictive modelling of anticancer drug sensitivity. Nature. 2012;483:603–7.

Barta JA, Powell CA, Wisnivesky JP. Global epidemiology of lung cancer. Ann Glob Health. 2019;85:8.

Behjati S, Tarpey PS. What is next generation sequencing? Arch Dis Child Educ Pract Ed. 2013;98: 236–8.

Berg KCG, Eide PW, Eilertsen IA, Johannessen B, Bruun J, Danielsen SA, Bjornslett M, Meza-Zepeda LA, Eknaes M, Lind GE, Myklebost O, Skotheim RI, Sveen A, Lothe RA. Multi-omics of 34 colorectal cancer cell lines - a resource for biomedical studies. Mol Cancer. 2017;16:116.

Bishayee A "The Role of Inflammation in Liver Cancer," in *Inflammation and Cancer,* eds. B.B. Aggarwal, B. Sung & S.C. Gupta.), 2014 401–435.

Biswas N, Chakrabarti S. Artificial intelligence (AI)-based systems biology approaches in multi-omics data analysis of cancer. Front Oncol. 2020;10:588221.

Bosch FX, Ribes J, Borras J. Epidemiology of primary liver cancer. Semin Liver Dis. 1999;19:271–85.

Bosch FX, Ribes J, Diaz M, Cleries R. Primary liver cancer: worldwide incidence and trends. Gastroenterology. 2004;127:S5–S16.

Canzler S, Schor J, Busch W, Schubert K, Rolle-Kampczyk UE, Seitz H, Kamp H, Von Bergen M, Buesen R, Hackermuller J. Prospects and challenges of multi-omics data integration in toxicology. Arch Toxicol. 2020;94:371–88.

Center MM, Jemal A, Smith RA, Ward E. Worldwide variations in colorectal. Cancer. 2009;59: 366–78.

Chakraborty S, Hosen MI, Ahmed M, Shekhar HU. Onco-multi-OMICS approach: a new frontier in cancer research. Biomed Res Int. 2018;2018:9836256.

Chang MH. Prevention of hepatitis B virus infection and liver cancer. Recent Results Cancer Res. 2014;193:75–95.

Chaudhary K, Poirion OB, Lu L, Garmire LX. Deep learning-based multi-omics integration robustly predicts survival in liver cancer. Clin Cancer Res. 2018;24:1248–59.

Chauvel C, Novoloaca A, Veyre P, Reynier F, Becker J. Evaluation of integrative clustering methods for the analysis of multi-omics data. Brief Bioinform. 2020;21:541–52.

Chen JG, Zhang SW. Liver cancer epidemic in China: past, present and future. Semin Cancer Biol. 2011;21:59–69.

Chen Z, Fillmore CM, Hammerman PS, Kim CF, Wong KK. Non-small-cell lung cancers: a heterogeneous set of diseases. Nat Rev Cancer. 2014;14:535–46.

Cheng WC, Chung IF, Chen CY, Sun HJ, Fen JJ, Tang WC, Chang TY, Wong TT, Wang HW. DriverDB: an exome sequencing database for cancer driver gene identification. Nucleic Acids Res. 2014;42:D1048-1054.

Chung IF, Chen CY, Su SC, Li CY, Wu KJ, Wang HW, Cheng WC. DriverDBv2: a database for human cancer driver gene research. Nucleic Acids Res. 2016;44:D975-979.

Chyr J, Zhang Z, Chen X, Zhou X. PredTAD: a machine learning framework that models 3D chromatin organization alterations leading to oncogene dysregulation in breast cancer cell lines. Comput Struct Biotechnol J. 2021;19:2870–80.

Cieslik M, Chinnaiyan AM. Global genomics project unravels cancer's complexity at unprecedented scale. Nature. 2020;578:39–40.

Colaprico A, Silva TC, Olsen C, Garofano L, Cava C, Garolini D, Sabedot TS, Malta TM, Pagnotta SM, Castiglioni I, Ceccarelli M, Bontempi G, Noushmehr H. TCGAbiolinks: an R/Bioconductor package for integrative analysis of TCGA data. Nucleic Acids Res. 2016;44: e71.

Das T, Andrieux G, Ahmed M, Chakraborty S. Integration of online omics-data resources for cancer research. Front Genet. 2020;11:578345.

De Groot PM, Wu CC, Carter BW, Munden RF. The epidemiology of lung cancer. Transl Lung Cancer Res. 2018;7:220–33.

De Sousa VML, Carvalho L. Heterogeneity in lung cancer. Pathobiology. 2018;85:96–107.

Dienstmann R, Vermeulen L, Guinney J, Kopetz S, Tejpar S, Tabernero J. Consensus molecular subtypes and the evolution of precision medicine in colorectal cancer. Nat Rev Cancer. 2017;17: 79–92.

Dimitrakopoulos C, Hindupur SK, Colombi M, Liko D, Ng CKY, Piscuoglio S, Behr J, Moore AL, Singer J, Ruscheweyh HJ, Matter MS, Mossmann D, Terracciano LM, Hall MN, Beerenwinkel N. Multi-omics data integration reveals novel drug targets in hepatocellular carcinoma. BMC Genomics. 2021;22:592.

Domon B, Aebersold R. Mass spectrometry and protein analysis. Science. 2006;312:212–7.

Du W, Elemento O. Cancer systems biology: embracing complexity to develop better anticancer therapeutic strategies. Oncogene. 2015;34:3215–25.

Eicher T, Kinnebrew G, Patt A, Spencer K, Ying K, Ma Q, Machiraju R, Mathe AEA. Metabolomics and multi-omics integration: a survey of computational methods and resources. Meta. 2020;10:202.

Fearon ER. "Molecular genetics of colorectal cancer," in *Cancer prevention: from the laboratory to the clinic: implications of genetic, molecular, and preventive research,* eds. H.L. Bradlow, M.P. Osborne & U. Veronesi.), 1995 101–110.

Gao Q, Zhu HW, Dong LQ, Shi WW, Chen R, Song ZJ, Huang C, Li JQ, Dong XW, Zhou YT, Liu Q, Ma LJ, Wang XY, Zhou J, Liu YS, Boja E, Robles AI, Ma WP, Wang P, Li YZ, Ding L, Wen B, Zhang B, Rodriguez H, Gao DM, Zhou H, Fan J. Integrated proteogenomic characterization of HBV-related hepatocellular carcinoma. Cell. 2019;179:561.

Ghaffari S, Hanson C, Schmidt RE, Bouchonville KJ, Offer SM, Sinha S. An integrated multiomics approach to identify regulatory mechanisms in cancer metastatic processes. Genome Biol. 2021;22:19.

Ghosh D, Bernstein JA, Khurana Hershey GK, Rothenberg ME, Mersha TB. Leveraging multilayered "omics" data for atopic dermatitis: a road map to precision medicine. Front Immunol. 2018;9:2727.

Graw S, Chappell K, Washam CL, Gies A, Bird J, Robeson MS 2nd, Byrum SD. Multi-omics data integration considerations and study design for biological systems and disease. Mol Omics. 2021;17:170–85.

Hao Y, Li D, Xu Y, Ouyang J, Wang Y, Zhang Y, Li B, Xie L, Qin G. Investigation of lipid metabolism dysregulation and the effects on immune microenvironments in pan-cancer using multiple omics data. BMC Bioinformatics. 2019;20:195.

Hausman DM. What is cancer? Perspect Biol Med. 2019;62:778–84.

Herbst RS, Morgensztern D, Boshoff C. The biology and management of non-small cell lung cancer. Nature. 2018;553:446–54.

Holowatyj AN, Haffa M, Lin T, Scherer D, Gigic B, Ose J, Warby CA, Himbert C, Abbenhardt-Martin C, Achaintre D, Boehm J, Boucher KM, Gicquiau A, Gsur A, Habermann N, Herpel E, Kauczor HU, Keski-Rahkonen P, Kloor M, Von Knebel-Doeberitz M, Kok DE, Nattenmuller J, Schirmacher P, Schneider M, Schrotz-King P, Simon T, Ueland PM, Viskochil R, Weijenberg MP, Scalbert A, Ulrich A, Bowers LW, Hursting SD, Ulrich CM. Multi-omics analysis reveals adipose-tumor crosstalk in patients with colorectal cancer. Cancer Prev Res (Phila). 2020;13: 817–28.

Hu W, Yang Y, Li X, Huang M, Xu F, Ge W, Zhang S, Zheng S. Multi-omics approach reveals distinct differences in left- and right-sided colon cancer. Mol Cancer Res. 2018;16:476–85.

Hu Z, Bi G, Sui Q, Bian Y, Du Y, Liang J, Li M, Zhan C, Lin Z, Wang Q. Analyses of multi-omics differences between patients with high and low PD1/PDL1 expression in lung squamous cell carcinoma. Int Immunopharmacol. 2020;88:106910.

Huang L, Brunell D, Stephan C, Mancuso J, Yu X, He B, Thompson TC, Zinner R, Kim J, Davies P, Wong STC. Driver network as a biomarker: systematic integration and network modeling of multi-omics data to derive driver signaling pathways for drug combination prediction. Bioinformatics. 2019;35:3709–17.

Huang S, Chaudhary K, Garmire LX. More is better: recent progress in multi-omics data integration methods. Front Genet. 2017;8:84.

Huang Y, Duanmu J, Liu Y, Yan M, Li T, Jiang Q. Analysis of multi-omics differences in left-side and right-side colon cancer. PeerJ. 2021;9:e11433.

Jahanafrooz Z, Mosafer J, Akbari M, Hashemzaei M, Mokhtarzadeh A, Baradaran B. Colon cancer therapy by focusing on colon cancer stem cells and their tumor microenvironment. J Cell Physiol. 2020;235:4153–66.

Jones GS, Baldwin DR. Recent advances in the management of lung cancer. Clin Med (Lond). 2018;18:s41–6.

Kel AE, Stegmaier P, Valeev T, Koschmann J, Poroikov V, Kel-Margoulis OV, Wingender E. Multi-omics "upstream analysis" of regulatory genomic regions helps identifying targets against methotrexate resistance of colon cancer. EuPA Open Proteom. 2016;13:1–13.

Kim TH, Dekker J. ChIP-seq. Cold Spring Harb Protoc; 2018.

Koch A, De Meyer T, Jeschke J, Van Criekinge W. MEXPRESS: visualizing expression, DNA methylation and clinical TCGA data. BMC Genomics. 2015;16:636.

Koch A, Jeschke J, Van Criekinge W, Van Engeland M, De Meyer T. MEXPRESS update 2019. Nucleic Acids Res. 2019;47:W561–5.

Kong Y, Qiao Z, Ren Y, Genchev GZ, Ge M, Xiao H, Zhao H, Lu H. Integrative analysis of membrane proteome and MicroRNA reveals novel lung cancer metastasis biomarkers. Front Genet. 2020;11:1023.

Lam K, Pan K, Linnekamp JF, Medema JP, Kandimalla R. DNA methylation based biomarkers in colorectal cancer: a systematic review. Biochim Biophys Acta. 2016;1866:106–20.

Lee TY, Huang KY, Chuang CH, Lee CY, Chang TH. Incorporating deep learning and multi-omics autoencoding for analysis of lung adenocarcinoma prognostication. Comput Biol Chem. 2020;87:107277.

Li L, Wang H. Heterogeneity of liver cancer and personalized therapy. Cancer Lett. 2016;379:191–7.

Li P, Guo M, Sun B. Integration of multi-omics data to mine cancer-related gene modules. J Bioinforma Comput Biol. 2019;17:1950038.

Lin S, Yin YA, Jiang X, Sahni N, Yi S. Multi-OMICs and genome editing perspectives on liver cancer signaling networks. Biomed Res Int. 2016;2016:6186281.

Liu CY, Chen KF, Chen PJ. Treatment of liver cancer. Cold Spring Harb Perspect Med. 2015;5: a021535.

Liu SH, Shen PC, Chen CY, Hsu AN, Cho YC, Lai YL, Chen FH, Li CY, Wang SC, Chen M, Chung IF, Cheng WC. DriverDBv3: a multi-omics database for cancer driver gene research. Nucleic Acids Res. 2020a;48:D863–70.

Liu X, Qin J, Gao T, Li C, Chen X, Zeng K, Xu M, He B, Pan B, Xu X, Pan Y, Sun H, Xu T, Wang S. Analysis of METTL3 and METTL14 in hepatocellular carcinoma. Aging (Albany NY). 2020b;12:21638–59.

Liu XN, Cui DN, Li YF, Liu YH, Liu G, Liu L. Multiple "omics" data-based biomarker screening for hepatocellular carcinoma diagnosis. World J Gastroenterol. 2019;25:4199–212.

Llabata P, Mitsuishi Y, Choi PS, Cai D, Francis JM, Torres-Diz M, Udeshi ND, Golomb L, Wu Z, Zhou J, Svinkina T, Aguilera-Jimenez E, Liu Y, Carr SA, Sanchez-Cespedes M, Meyerson M, Zhang X. Multi-omics analysis identifies MGA as a negative regulator of the MYC pathway in lung adenocarcinoma. Mol Cancer Res. 2020;18:574–84.

Lu Y, Li Q, Zheng K, Fu C, Jiang C, Zhou D, Xia C, Ma S. Development of a high efficient promoter finding method based on transient transfection. Gene X. 2019;2:100008.

Luan M, Song F, Qu S, Meng X, Ji J, Duan Y, Sun C, Si H, Zhai H. Multi-omics integrative analysis and survival risk model construction of non-small cell lung cancer based on The Cancer Genome Atlas datasets. Oncol Lett. 2020;20:58.

Lv J, Wang J, Shang X, Liu F, Guo S. Survival prediction in patients with colon adenocarcinoma via multi-omics data integration using a deep learning algorithm. Biosci Rep. 2020;40: BSR20201482.

Mantini G, Pham TV, Piersma SR, Jimenez CR. Computational analysis of Phosphoproteomics data in multi-omics cancer studies. Proteomics. 2021;21:e1900312.

Marengo A, Rosso C, Bugianesi E. Liver cancer: connections with obesity, fatty liver, and cirrhosis. Annu Rev Med. 2016;67:103–17.

Mármol I, Sánchez-De-Diego C, Pradilla Dieste A, Cerrada E, Rodriguez Yoldi MJ. Colorectal carcinoma: a general overview and future perspectives in colorectal cancer. Int J Mol Sci. 2017;18:197.

Matthiesen R, Jensen ON. Analysis of mass spectrometry data in proteomics. Methods Mol Biol. 2008;453:105–22.

Mcglynn KA, Tsao L, Hsing AW, Devesa SS, Fraumeni JF. International trends and patterns of primary liver cancer. Int J Cancer. 2001;94:290–6.

Mckillop IH, Schrum LW. Alcohol and liver cancer. Alcohol. 2005;35:195–203.

Meng C, Zeleznik OA, Thallinger GG, Kuster B, Gholami AM, Culhane AC. Dimension reduction techniques for the integrative analysis of multi-omics data. Brief Bioinform. 2016;17:628–41.

Miao R, Luo H, Zhou H, Li G, Bu D, Yang X, Zhao X, Zhang H, Liu S, Zhong Y, Zou Z, Zhao Y, Yu K, He L, Sang X, Zhong S, Huang J, Wu Y, Miksad RA, Robson SC, Jiang C, Zhao Y, Zhao H. Identification of prognostic biomarkers in hepatitis B virus-related hepatocellular carcinoma and stratification by integrative multi-omics analysis. J Hepatol. 2014;61:840–9.

Minna JD, Roth JA, Gazdar AF. Focus on lung cancer. Cancer Cell. 2002;1:49–52.

Nakano D, Kawaguchi T, Iwamoto H, Hayakawa M, Koga H, Torimura T. Effects of canagliflozin on growth and metabolic reprograming in hepatocellular carcinoma cells: multi-omics analysis of metabolomics and absolute quantification proteomics (iMPAQT). PLoS One. 2020;15: e0232283.

Netanely D, Stern N, Laufer I, Shamir R. PROMO: an interactive tool for analyzing clinically-labeled multi-omic cancer datasets. BMC Bioinformatics. 2019;20:732.

Nicora G, Vitali F, Dagliati A, Geifman N, Bellazzi R. Integrated multi-omics analyses in oncology: a review of machine learning methods and tools. Front Oncol. 2020;10:1030.

Nusinow DP, Szpyt J, Ghandi M, Rose CM, Mcdonald ER 3rd, Kalocsay M, Jane-Valbuena J, Gelfand E, Schweppe DK, Jedrychowski M, Golji J, Porter DA, Rejtar T, Wang YK, Kryukov GV, Stegmeier F, Erickson BK, Garraway LA, Sellers WR, Gygi SP. Quantitative proteomics of the cancer cell line Encyclopedia. Cell. 2020;180(387–402):e316.

O'connell JB, Maggard MA, Livingston EH, Yo CK. Colorectal cancer in the young. Am J Surg. 2004;187:343–8.

Ouyang X, Fan Q, Ling G, Shi Y, Hu F. Identification of diagnostic biomarkers and subtypes of liver hepatocellular carcinoma by multi-omics data analysis. Genes (Basel). 2020;11:1051.

Park M, Kim D, Moon K, Park T. Integrative analysis of multi-omics data based on blockwise sparse principal components. Int J Mol Sci. 2020;21:8202.

Pastushenko I, Blanpain C. EMT transition states during tumor progression and metastasis. Trends Cell Biol. 2019;29:212–26.

Patz EF, Goodman PC, Bepler G. Current concepts—screening for lung cancer. N Engl J Med. 2000;343:1627–33.

Potter JD. Colorectal cancer: molecules and populations. J Natl Cancer Inst. 1999;91:916–32.

Preisser F, Cooperberg MR, Crook J, Feng F, Graefen M, Karakiewicz PI, Klotz L, Montironi R, Nguyen PL, D'amico AV. Intermediate-risk prostate cancer: stratification and management. Eur Urol Oncol. 2020;3:270–80.

Rhee EP. How omics data can be used in nephrology. Am J Kidney Dis. 2018;72:129–35.

Salgia R, Skarin AT. Molecular abnormalities in lung cancer. J Clin Oncol. 1998;16:1207–17.

Shen H, Yang J, Huang Q, Jiang MJ, Tan YN, Fu JF, Zhu LZ, Fang XF, Yuan Y. Different treatment strategies and molecular features between right-sided and left-sided colon cancers. World J Gastroenterol. 2015;21:6470–8.

Shen M, Xu M, Zhong F, Crist MC, Prior AB, Yang K, Allaire DM, Choueiry F, Zhu J, Shi H. A multi-omics study revealing the metabolic effects of Estrogen in liver cancer cells HepG2. Cell. 2021a;10:455.

Shen Y, Xiong W, Gu Q, Zhang Q, Yue J, Liu C, Wang D. Multi-omics integrative analysis uncovers molecular subtypes and mRNAs as therapeutic targets for liver cancer. Front Med (Lausanne). 2021b;8:654635.

Sia D, Villanueva A, Friedman SL, Llovet JM. Liver cancer cell of origin, molecular class, and effects on patient prognosis. Gastroenterology. 2017;152:745–61.

Siegel R, Desantis C, Jemal A. Colorectal cancer statistics, 2014. CA-a Cancer J Clin. 2014;64: 104–17.

Song J, Yang J, Lin R, Cai X, Zheng L, Chen Y. Molecular heterogeneity of guanine nucleotide binding-protein gamma subunit 4 in left- and right-sided colon cancer. Oncol Lett. 2020;20:334.

Spiro SG, Silvestri GA. One hundred years of lung cancer. Am J Respir Crit Care Med. 2005;172: 523–9.

Srivatanakul P, Sriplung H, Deerasamee S. Epidemiology of liver cancer: an overview. Asian Pac J Cancer Prev. 2004;5:118–25.

Stintzing S. Management of colorectal cancer. F1000Prime Rep. 2014;6:108.

Subramanian I, Verma S, Kumar S, Jere A, Anamika K. Multi-omics data integration, interpretation, and its application. Bioinform Biol Insights. 2020;14:1–24.

Sun YV, Hu YJ. Integrative analysis of multi-omics data for discovery and functional studies of complex human diseases. Adv Genet. 2016;93:147–90.

Sung H, Ferlay J, Siegel RL, Laversanne M, and Bray, F.J.C.a.C.J.F.C.. Global Cancer Statistics 2020: GLOBOCAN Estimates of Incidence and Mortality Worldwide for 36 Cancers in 185 Countries. 2021, 71.

Tomczak K, Czerwinska P, Wiznerowicz M. The cancer genome atlas (TCGA): an immeasurable source of knowledge. Contemp Oncol (Pozn). 2015;19:A68-77.

Tong D, Tian Y, Zhou T, Ye Q, Li J, Ding K, Li J. Improving prediction performance of colon cancer prognosis based on the integration of clinical and multi-omics data. BMC Med Inform Decis Mak. 2020;20:22.

Torre LA, Siegel RL, Jemal A. Lung cancer statistics. Adv Exp Med Biol. 2016a;893:1–19.

Torre LA, Siegel RL, Ward EM, Jemal A. Global cancer incidence and mortality rates and trends-an update. Cancer Epidemiol Biomark Prev. 2016b;25:16–27.

Valencia AM, Kadoch C. Chromatin regulatory mechanisms and therapeutic opportunities in cancer. Nat Cell Biol. 2019;21:152–61.

Van Meerbeeck JP, Fennell DA, De Ruysscher DKM. Small-cell lung cancer. Lancet. 2011;378: 1741–55.

Vasaikar SV, Straub P, Wang J, Zhang B. LinkedOmics: analyzing multi-omics data within and across 32 cancer types. Nucleic Acids Res. 2018;46:D956–63.

Wang C, Yang F, Chen T, Dong Q, Zhao Z, Liu Y, Chen B, Liang H, Yang H, Gu Y. RHPCG: a database of the regulation of the hippo pathway in cancer genome. Database (Oxford). 2019;2019:baz135.

Wang Z, Jensen MA, Zenklusen JC. A practical guide to the cancer genome atlas (TCGA). Methods Mol Biol. 2016;1418:111–41.

Wang Z, Wei Y, Zhang R, Su L, Gogarten SM, Liu G, Brennan P, Field JK, Mckay JD, Lissowska J, Swiatkowska B, Janout V, Bolca C, Kontic M, Scelo G, Zaridze D, Laurie CC, Doheny KF, Pugh EK, Marosy BA, Hetrick KN, Xiao X, Pikielny C, Hung RJ, Amos CI, Lin X, Christiani DC. Multi-omics analysis reveals a HIF network and hub gene EPAS1 associated with lung adenocarcinoma. EBioMedicine. 2018;32:93–101.

Weinberg DS, Schoen RE. Screening for colorectal cancer. Ann Intern Med. 2014;160

Weinberg RA. How cancer arises. Sci Am. 1996;275:62–70.

Wistuba I, Gazdar AF. Lung cancer preneoplasia. Annu Rev Pathol. 2006;1:331–48.

Woo HG, Choi JH, Yoon S, Jee BA, Cho EJ, Lee JH, Yu SJ, Yoon JH, Yi NJ, Lee KW, Suh KS, Kim YJ. Integrative analysis of genomic and epigenomic regulation of the transcriptome in liver cancer. Nat Commun. 2017;8:839.

Wu B, Wang Z, Lin N, Yan XB, Lv ZC, Ying ZM, Ye ZM. A panel of eight mRNA signatures improves prognosis prediction of osteosarcoma patients. Medicine. 2021a;100:e24118.

Wu L, Liu F, Cai H. IOAT: an interactive tool for statistical analysis of omics data and clinical data. BMC Bioinformatics. 2021b;22:326.

Wu Y, Yang Y, Gu H, Tao B, Zhang E, Wei J, Wang Z, Liu A, Sun R, Chen M, Fan Y, Mao R. Multi-omics analysis reveals the functional transcription and potential translation of enhancers. Int J Cancer. 2020;147:2210–24.

Xie B, Yuan Z, Yang Y, Sun Z, Zhou S, Fang X. MOBCdb: a comprehensive database integrating multi-omics data on breast cancer for precision medicine. Breast Cancer Res Treat. 2018;169: 625–32.

Xie Q, Fan F, Wei W, Liu Y, Xu Z, Zhai L, Qi Y, Ye B, Zhang Y, Basu S, Zhao Z, Wu J, Xu P. Multi-omics analyses reveal metabolic alterations regulated by hepatitis B virus core protein in hepatocellular carcinoma cells. Sci Rep. 2017;7:41089.

Xie Y. Hepatitis B virus-associated hepatocellular carcinoma. Adv Exp Med Biol. 2017;1018:11–21.

Xu Q, Zhai JC, Huo CQ, Li Y, Dong XJ, Li DF, Huang RD, Shen C, Chang YJ, Zeng XL, Meng FL, Yang F, Zhang WL, Zhang SN, Zhou YM, Zhang Z. OncoPDSS: an evidence-based clinical decision support system for oncology pharmacotherapy at the individual level. BMC Cancer. 2020a;20:740.

Xu X, Gong C, Wang Y, Hu Y, Liu H, Fang Z. Multi-omics analysis to identify driving factors in colorectal cancer. Epigenomics. 2020b;12:1633–50.

Xu Y, She Y, Li Y, Li H, Jia Z, Jiang G, Liang L, Duan L. Multi-omics analysis at epigenomics and transcriptomics levels reveals prognostic subtypes of lung squamous cell carcinoma. Biomed Pharmacother. 2020c;125:109859.

Yamashita T, Wang XW. Cancer stem cells in the development of liver cancer. J Clin Investig. 2013;123:1911–8.

Yang H, Jin W, Liu H, Wang X, Wu J, Gan D, Cui C, Han Y, Han C, Wang Z. A novel prognostic model based on multi-omics features predicts the prognosis of colon cancer patients. Mol Genet Genomic Med. 2020a;8:e1255.

Yang Z, Xu J, Li L, Li R, Wang Y, Tian Y, Guo W, Wang Z, Tan F, Ying J, Jiao Y, Gao S, Wang J, Gao Y, He J. Integrated molecular characterization reveals potential therapeutic strategies for pulmonary sarcomatoid carcinoma. Nat Commun. 2020b;11:4878.

Yi H, Li G, Long Y, Liang W, Cui H, Zhang B, Tan Y, Li Y, Shen L, Deng D, Tang Y, Mao C, Tian S, Cai Y, Zhu Q, Hu Y, Chen W, Fang L. Integrative multi-omics analysis of a colon cancer cell line with heterogeneous Wnt activity revealed RUNX2 as an epigenetic regulator of EMT. Oncogene. 2020;39:5152–64.

Yi T, Zhang Y, Ng DM, Xi Y, Ye M, Cen L, Li J, Fan X, Li Y, Hu S, Rong H, Xie Y, Zhao G, Chen L, Chen C, Ni S, Mi J, Dai X, Liao Q. Regulatory network analysis of mutated genes based on multi-omics data reveals the exclusive features in tumor immune microenvironment between left-sided and right-sided colon cancer. Front Oncol. 2021;11:685515.

Yildiz G. Integrated multi-omics data analysis identifying novel drug sensitivity-associated molecular targets of hepatocellular carcinoma cells. Oncol Lett. 2018;16:113–22.

Yin Z, Yan X, Wang Q, Deng Z, Tang K, Cao Z, Qiu T. Detecting prognosis risk biomarkers for colon cancer through multi-omics-based prognostic analysis and target regulation simulation Modeling. Front Genet. 2020;11:524.

Yoo BC, Kim KH, Woo SM, Myung JK. Clinical multi-omics strategies for the effective cancer management. J Proteome. 2018;188:97–106.

Yu C, Qi X, Lin Y, Li Y, Shen B. iODA: an integrated tool for analysis of cancer pathway consistency from heterogeneous multi-omics data. J Biomed Inform. 2020;112:103605.

Yuan Y, Bao J, Chen Z, Villanueva AD, Wen W, Wang F, Zhao D, Fu X, Cai Q, Long J, Shu XO, Zheng D, Moreno V, Zheng W, Lin W, Guo X. Multi-omics analysis to identify susceptibility genes for colorectal cancer. Hum Mol Genet. 2021;30:321–30.

Zeng D, Ye Z, Shen R, Yu G, Wu J, Xiong Y, Zhou R, Qiu W, Huang N, Sun L, Li X, Bin J, Liao Y, Shi M, Liao W. IOBR: multi-omics immuno-oncology biological research to decode tumor microenvironment and signatures. Front Immunol. 2021;12:687975.

Zhang H, Jin Z, Cheng L, Zhang B. Integrative analysis of methylation and gene expression in lung adenocarcinoma and squamous cell lung carcinoma. Front Bioeng Biotechnol. 2020a;8:3.

Zhang J, Bajari R, Andric D, Gerthoffert F, Lepsa A, Nahal-Bose H, Stein LD, Ferretti V. The international cancer genome consortium data portal. Nat Biotechnol. 2019;37:367–9.

Zhang Y, Yang M, Ng DM, Haleem M, Yi T, Hu S, Zhu H, Zhao G, Liao Q. Multi-omics data analyses construct TME and identify the immune-related prognosis signatures in human LUAD. Mol Ther Nucleic Acids. 2020b;21:860–73.

Zochbauer-Muller S, Minna JD. The biology of lung cancer including potential clinical applications. Chest Surg Clin N Am. 2000;10:691–708.

Zou Y, Ruan S, Jin L, Chen Z, Han H, Zhang Y, Jian Z, Lin Y, Shi N, Jin H. CDK1, CCNB1, and CCNB2 are prognostic biomarkers and correlated with immune infiltration in hepatocellular carcinoma. Med Sci Monit. 2020;26:e925289.

Chapter 6
Multi-Omics Data Analysis for Inflammation Disease Research: Correlation Analysis, Causal Analysis and Network Analysis

Maozhen Han, Na Zhang, Zhangjie Peng, Yujie Mao, Qianqian Yang, Yiyang Chen, Mengfei Ren, and Weihua Jia

In recent decades, an enormous amount of research on the human gut microbiota has established that it is strongly associated with human health and is involved in the occurrence and development of a variety of diseases, including inflammation diseases. However, current research has concentrated on the relationship between disease and the human gut microbiota, as well as the interactions between microorganisms. The absence of a more detailed understanding of the mechanism and causation of diseases associated with the gut microbiota restricts clinical diagnosis and treatment. Due to the advancement of sequencing and mass spectrometry techniques, numerous approaches have been used in microbiome research to generate multi-omics datasets that can provide a comprehensive view of the compositions and changes in microbial communities' genetic, metabolic, and biochemical processes, as well as an in-depth understanding of the gut microbiome and diseases. Nonetheless, the absence of systematic reviews of multi-omics approaches and their application to diseases restricts their application to microbiome research. As such, we took a holistic view of multi-omics approaches in the gut microbiome and discussed how multi-omics approaches could aid in disease diagnosis and treatment in this review. To be clear, we used inflammation disease as a model disease to introduce multi-omics approaches, integrated analysis methods for multi-omics datasets, and their application to inflammation diseases, particularly in terms of

M. Han (✉)
School of Life Sciences, Anhui Medical University, Hefei, Anhui, China

Key Laboratory of Molecular Biophysics of the Ministry of Education, Hubei Key Laboratory of Bioinformatics and Molecular-imaging, Department of Bioinformatics and Systems Biology, College of Life Science and Technology, Huazhong University of Science and Technology, Wuhan, Hubei, China
e-mail: hanmz@ahmu.edu.cn

N. Zhang · Z. Peng · Y. Mao · Q. Yang · Y. Chen · M. Ren · W. Jia
School of Life Sciences, Anhui Medical University, Hefei, Anhui, China

K. Ning (ed.), *Methodologies of Multi-Omics Data Integration and Data Mining*,
Translational Bioinformatics 19, https://doi.org/10.1007/978-981-19-8210-1_6

treatment methods involving microbiome approaches. Without a doubt, our comprehensive review of multi-omics approaches in inflammation disease and the bioinformatics tools for integrating multi-omics datasets may help identify clades for clinical diagnosis and treatment of inflammation diseases.

6.1 Introduction

The human microbiota is a collection of microorganisms that coexist in symbiotic communities with various human body sites (Manzo and Bhatt 2015). The human microbiota contains over 100 trillion bacteria, the majority of which colonized the gut, evolved alongside the host, and became an inseparable part of the host (Glasner 2017). In recent decades, mounting evidence has established that the human gut microbiome is associated with human health, contributes to host physiology, metabolism, and the development of the host immune system (Sharon et al. 2016), and is involved in the development of inflammation diseases and many other diseases, such as hypertension (Li et al. 2017), cholestatic liver disease (Isaacs-Ten et al. 2020), rheumatoid arthritis (RA) (Manasson et al. 2020), systemic lupus erythematosus (SLE) (Chen et al. 2020), cancers (Slowicka et al. 2020; Amy et al. 2020), etc., which suggested that human gut microbiome play an essential and important role in keeping host health. Hence, more and more researchers focused on the research of gut microbiota to uncover the characters of the gut microbiome, explore the dynamic changes of gut microbiota during the development of disease, and investigate the detailed mechanism between the gut microbiome and disease using various omics approaches.

Researchers have examined the characteristics of the human gut microbiome as a result of the success of their investigation into gut microbial communities. For example, the human gut microbiome is normally in homeostasis and can be classified into a specific enterotype that is conserved across disparate populations (Arumugam et al. 2011) based on the composition of gut microbial communities. Three enterotypes have been identified in the human gut microbiome to date: enterotype 1 (predominant genus *Bacteroides*), enterotype 2 (predominant genus *Prevotella*), and enterotype 3 (the predominant genus is *Ruminococcus*). We previously demonstrated that human microbial communities are bidirectionally plastic and resilient, as well as the eliciting effects on clinical practice for gut microbiome-associated diseases (Liu et al. 2019). With the increasing number of human gut microbiome studies, researchers gained a better understanding of the gut microbiota and identified factors that affect the composition of gut microbial communities, which may be associated with the development of human diseases. Among these studies, Peer Bork and colleagues reported that enterotypes are unaffected by age, gender, body weight, or national population (Arumugam et al. 2011), and the following study suggested that alternative enterotype states are influenced by long-term diet (Wu et al. 2011). Our studies have suggested that the human gut microbial compositions, including taxonomical composition and functional composition, are affected by dietary factors (Liu et al. 2019), physical characteristics, and sport-related features (Han et al. 2020a). Specifically, the correlation analysis between the dynamic changes of human gut microbiota and the dietary shifts showed that the resilience of human gut microbiota is largely mediated by dietary changes. The variation partitioning analysis between the compositions of gut microbial communities and various factors, which was mainly divided into three different groups, including dietary factors, physical characteristics, and sport-related indices from three cohorts of athletes showed that these factors can in concert explain 41% of the inter-person human gut microbiome (Han et al. 2020a). Besides, a growing number of studies paid attention to the correlation between factors and human gut microbiome composition, and between the development of diseases and human gut microbiome (Ghaisas et al. 2016; Gaulke and Sharpton 2018; Bäckhed et al. 2015). The findings of the gut microbiome studies have been demonstrated that the environment and host factors can influence the composition of the human gut microbiome with large-scale association analyses (Kurilshikov et al. 2021; Spor et al. 2011). However, the results of these studies are limited to correlation analysis between disease and human gut microbiota, the interactions among microbiota, and lack of more detailed mechanism analysis and causality analysis (Cani 2018; Walter et al. 2020; Harley and Karp 2012), although researchers have conducted a lot of human gut microbiome studies.

Additionally, it should be noted that these studies examined the correlation between disease and the human gut microbiota using a single microbiome approach, including 16S rRNA amplicon sequencing, metagenomic sequencing, metaproteomic data, and meta-metabolome data. For example, a previous systematic review summarized sixteen articles that used 16S rRNA-targeted sequencing data from 777 patients with irritable bowel syndrome (IBS) and 461 healthy controls (HCs) to examine the differences in gut microbiota between IBS patients and HCs and found that the results were inconsistent, if not contradictory (Duan et al. 2019). Notably, the Human Microbiome Project (HMP), which consists of two phases, HMP1 and iHMP, was established to generate resources for the purpose of characterizing the human microbiota (Turnbaugh et al. 2007; Integrative et al. 2019). Regarding the first phase of HMP, HMP1 focused on the microbiome collected from five major body sites of healthy humans using 16S rRNA amplicon sequencing and metagenomic shotgun sequencing to generate resources for the healthy human microbiome, including 600 microbial reference genomes, 70 million 16S sequences, 700 metagenomes, and 60 million predicted genes and proteins (Proctor 2011; Gevers et al. 2012), which provides baseline taxonomic and functional dives (Lloyd-Price et al. 2017). While HMP1 remains the largest resource for the human microbiome to date, the lack of longitudinal datasets, including host genetic and human microbiome datasets, and the limitation of single omics data make it difficult to gain a comprehensive understanding of the microbiome's causal relationship with disease. To address this, researchers proposed the second phase of HMP, dubbed iHMP, which collects host and microbiome samples from three distinct studies (pregnancy and preterm birth, inflammatory bowel disease (IBD), and type 2 diabetes (T2D), integrates longitudinal datasets using multiple omics technologies, and enables researchers to conduct dynamic analyses between host disease and microbiome using these multi-omics datasets during periods of huma (Integrative 2014).

Thus, to gain a thorough understanding of disease and develop a rational microbiome approach for treating gut microbiome-related diseases in clinical practice, we believe that it is critical to employ multiple approaches, particularly the combination of multi-omics methods, to collect multi-omics data and conduct big data analysis to uncover correlation, causality, and network analysis.

Numerous studies have demonstrated that multi-omics approaches in human gut microbiome research have the potential to provide a comprehensive picture of the compositions and changes in genetic, metabolic, and biochemical processes. Additionally, an increasing number of studies have used multi-omics approaches to generate multi-omics datasets, demonstrating a thorough understanding of the human gut microbiome (Lloyd-Price et al. 2019; Metwaly and Haller 2019). Nonetheless, there is a dearth of systematic reviews of multi-omics approaches and their application to diseases involving the gut microbiome, which limits multi-omics' application in the field of the gut microbiome, particularly for clinical diagnosis and treatment of diseases. As a result, we took a holistic view of multi-omics approaches in the gut microbiome and how multi-omics approaches may aid in disease diagnosis and treatment in this review. To clarify, we used inflammation disease as a model disease to demonstrate correlation analysis, causal analysis, and network analysis using multi-omics datasets derived from multi-omics approaches. To begin, we discussed the human gut microbiota and microbiome, inflammation diseases, and the relationship between the human gut microbiota and inflammation diseases. Second, we discussed the benefits of multi-omics approaches and summarized the methodology for integrating multi-omics datasets generated through multi-omics approaches. Thirdly, we discussed the applications of multi-omics approaches to diseases, particularly inflammation diseases, and briefly discussed how microbiome approaches can be used to treat inflammation diseases clinically. We summarized the application of multi-omics approaches to inflammation disease, the bioinformatics tools for integrating multi-omics datasets, and identified clades for clinical diagnosis and treatment of inflammation disease.

6.2 Human Gut Microbiota and Gut Microbiome

Humans have always been subjected to the natural microbial environment, and the human body is inhabited by an enormous number of microorganisms, forming a complex ecological community and symbiont of the human and bacteria (Dekaboruah et al. 2020). The human microbiota can be classified into bacteria, archaea, fungi, viruses, and eukaryotes and it can be colonized in the skin, the oral cavity, intestinal tract, and so on (Clemente et al. 2012). Numerous studies conducted over the last two decades have established that the human microbiota is primarily found in the intestinal tract, and these microorganisms have been dubbed the human gut microbiota (Thursby and Juge 2017), and tremendous of studies focused on human gut microbiota have demonstrated that the human gut microbiota is greatly impacting human health and physiology (Fan and Pedersen 2020) and plays an essential role in maintaining host health (Marchesi et al. 2016; Valdes et al. 2018).

Specifically, increasing evidence indicates that *Firmicutes*, *Bacteroidetes*, *Proteobacteria*, and *Actinobacteria* are the four most abundant bacterial phyla in the human gut microbial community (Han et al. 2020a; Zhang et al. 2015a; Shin et al. 2015) and the ratio of *Firmicutes* to *Bacteroidetes* was reported to be positively correlated with the production of total short-chain fatty acids (SCFAs) (Fernandes et al. 2014) and suggested as a potential indicator for monitoring the health of host (Chen et al. 2016), especially for solving the obesity problem (Sutoyo et al. 2020). Nevertheless, whether the *Firmicutes/Bacteroidetes* ratio really important is still a hotly debated spot (Magne et al. 2020; Schwiertz et al. 2010). Additionally, *Bacteroides*, *Prevotella*, *Ruminococcus*, and *Faecalibacterium* are the dominant bacterial genera in the human gut microbial community (Gorvitovskaia et al. 2016) and the previous three genera are famous as the dominant genus for three enterotypes (Romo-Vaquero et al. 2019).

Furthermore, several bacteria have been discovered and selected as biomarkers, which might be associated with the occurrence, development, and treatment of diseases, and the performance of athletes (Sandhu and McBride 2018; Scheiman et al. 2019; Marietta et al. 2016; Mithieux 2018). For example, numerous studies have demonstrated that *Clostridioides difficile* colonizes the colon and produces toxins, which can inhibit action polymerization in host cells and lead to cell death, causing the most common healthcare-related infection in American (called *C. difficile* infection, CDI) (Sandhu and McBride 2018; Mushtaq 2018; Abt et al. 2016). Additionally, previous research has established that *Prevotella copri* is the most prevalent pathogen in rheumatoid arthritis (RA) and that *P. copri* is strongly associated with the occurrence and development of RA (Tong et al. 2020; Scher et al. 2013; Pianta et al. 2017). Interestingly, a previous study demonstrated that enteral exposure to *Prevotella histicola* suppresses arthritis via mucosal regulation and proposed that *P. histicola* is a novel commensal that can be used to treat rheumatoid arthritis with few or no adverse effects (Marietta et al. 2016; Maeda and Takeda 2017). Additionally, several studies have demonstrated that *Veillonella atypica* can increase the metabolic conversion of exercise-induced lactate to propionate, thereby increasing the running time of mice, implying that it could be used to improve athlete performance (Scheiman et al. 2019; Han et al. 2020b; Kulecka et al. 2020).

To date, an increasing number of studies have focused on the relationship between human gut microbiota and disease, and with the advancement of sequencing techniques and the innovation of research technologies, studies of human gut microbiota have advanced to the omics level, namely the human gut microbiome (Cani 2018), revealing additional patterns or dynamic changes in human gut microbial communities. For example, several microbial taxonomic and functional signatures have been associated with metabolic traits in multiple cases of type 1 diabetes mellitus (T1DM) using metagenomic, metatranscriptomic, and metaproteomic datasets (Heintz-Buschart et al. 2016). These studies demonstrated that the human gut microbiome can be used as a potential controller of diseases (Kho and Lal 2018; Benítez-Páez et al. 2019). Undoubtedly, studies of the human gut microbiome have emphasized the benefits of studying species or strain levels to gain a better understanding of the relationship between human gut microbiota and diseases, as well as to identify clades for diagnosing and treating diseases using microbiome approaches.

6.3 Relationship Between Inflammation Diseases and Human Gut Microbiota

The human immune system is currently subjected to a variety of stresses, including emotional and physical strain, as well as exposure to environmental pollutants. Inflammation occurs naturally during this process as the human body's immune system responds to illness and infection. However, inflammation can sometimes be misdirected, resulting in the immune system attacking body tissues and resulting in inflammation diseases. Inflammatory diseases encompass more than 100 distinct conditions. Clinically prevalent inflammatory diseases include rheumatoid arthritis (RA), systemic lupus erythematosus (SLE) (Santos and Morand 2009), Crohn's disease (CD), ulcerative colitis (UC), and irritable bowel syndrome (IBS). And RA and SLE are two famous examples of inflammation diseases. Although many studies have focused on the occurrence, development, and molecular mechanism of RA and SLE (Alamanos and Drosos 2005; Yong et al. 2008), researchers continue to lack a clear understanding of the etiology of these two diseases, limiting their treatment. Fortunately, as evidence accumulates, the complex etiology of rheumatoid arthritis and systemic lupus erythematosus is becoming clear.

Specifically, RA is an autoimmune disease characterized by a chronic inflammatory disorder and a variety of inflammatory symptoms, including joint pain, swelling, and stiffness, as well as damage to other organs and tissues throughout the body, including the skin, eyes, lungs, and heart (Majithia and Geraci 2007; Bullock et al. 2018). Although the etiology of RA remains unknown, numerous previous studies have demonstrated that the disease is associated with genetic and environmental factors (Majithia and Geraci 2007; Aletaha and Smolen 2018), including human gut microbiota (Maeda and Takeda 2017; Horta-Baas et al. 2017). To begin, the results of experiments in germ-free mice demonstrated that the gut microbiota can shape the intestinal immune system, implying that the gut microbiota has a strong positive correlation with immune-mediated diseases (Round and Mazmanian 2009; Torres et al. 2020), including RA. Second, the increasing number of animal and human studies have demonstrated that gut microbiota plays an important role in the occurrence, development, and treatment of RA (Maeda and Takeda 2017). For an instance, several studies have revealed that the dysbiosis of human gut microbial communities can cause the disorder of immune system in RA patients. *P. copri* was dominant in RA patients and found to be correlated with an absence of human leukocyte antigen-DRB1 (Scher et al. 2013). In our laboratory, we used metagenome sequencing to compare the changes in taxonomic composition of the gut microbial communities between healthy SD mice and SD mice with RA (CIA model). Our analysis of metagenome datasets from six healthy SD mice and five SD mice with rheumatoid arthritis revealed significant differences in taxonomic and functional compositions compositions between SD healthy mice and SD mice with rheumatoid arthritis. Eight species were identified as biomarkers, including *Lactobacillus sp*. ASF360, *Akkermansia muciniphila*, *Prevotella copri*, *Parabacteroides goldsteinii*, *Parabacteroides johnsonii*, *Lactobacillus reuteri*, and *Bacteroides dorei*. Furthermore, our findings suggested that these species may be associated with the occurrence and development of rheumatoid arthritis. Naturally, the precise functional mechanism of these species is being verified.

Similarly, systemic lupus erythematosus (SLE) is an autoimmune disease that is characterized by a type I interferon gene signature (Guerrini et al. 2018). In this chronic disease, the immune system of the host mistakenly attacks its healthy tissues, causing inflammation and tissue damage. The most common type of SLE is lupus and it can affect the joints, skin, lungs, kidneys, brains, and other tissues (Borchers et al. 2010). Although the complex etiology of SLE is not fully understood, numerous studies have suggested that the interactions between host genetics and environmental factors contributing to the occurrence and development of SLE (Guerrini et al. 2018; Neuman and Koren 2017). Particularly, several studies results have proposed that an association between gut microbiota and SLE (Neuman and Koren 2017), the disorder of gut microbiota have been identified in SLE cohorts and candidate biomarkers, such as *Ruminococcus gnavus* and *Enterococcus gallinarum*, maybe contribute to the immune pathogenesis (Guerrini et al. 2018; Silverman et al. 2019). These studies suggested that dysbiosis of the gut microbiota, particularly dynamic alteration of specific bacteria, may be associated with the occurrence and progression of SLE. However, the causal relationship between gut microbiota and SLE is unknown, and additional research is required to fully understand the detailed mechanism of SLE.

Together, these previous studies suggested that the gut microbiota play a critical role in the onset and progression of inflammatory diseases and established a strong association between gut microbiota and inflammation diseases. However, the lack of causality between gut microbiota and inflammation diseases makes it difficult to gain a thorough understanding of the pathogenesis of inflammation diseases and to develop a clinically effective microbiome approach for treating inflammation diseases.

6.4 The Advantages of Multi-Omics Approaches and the Methodology for Integrating the Multi-Omics Datasets

Over the last two decades, researchers have recognized that the microbial community is composed of a diverse range of microorganisms, the majority of which are unculturable, and the critical role of the gut microbiota in maintaining host health. To elucidate which microbiota exist in a microbial community (who), what a microbial community is capable of (what), and how a microbial community functions (how), a variety of techniques have been used in microbiome research, including microarrays, high-throughput technologies, and liquid chromatography-mass spectrometry (LC-MS)(Hamady and Knight 2009; Zhou et al. 2015; Hansen et al. 2019). With the advancement of high-throughput technologies, 16S rRNA amplicon sequencing, whole genome sequencing, and transcriptome sequencing have all been widely applied in microbiome research, demonstrating the dynamic pattern of microbial communities. For instance, in the last decade, 16S rRNA amplicon sequencing has become a standard method for verifying the taxonomic compositions of microbial communities and shed light on the structure of numerous microbial ecosystems (Prodan et al. 2020; Caporaso et al. 2011). Due to the limitation of 16S rRNA amplicon sequencing in terms of identification level, whole genomic sequencing of metagenomic DNA was developed and is now widely used in microbiome research, specifically metagenomics. The taxonomic composition of the microbial community can be determined at the species level using metagenome datasets, and the functional composition can be profiled against various databases, including the CAZyme database (Huang et al. 2018), nitrogen cycle database (Tu et al. 2019), CARD database (Jia et al. 2016), and so on. Hence, researchers can analyze and investigate the pattern of the microbial ecosystems with different perspectives based on omics datasets obtained from different omics approaches.

To date, an enormous amount of microbiome research has been conducted, yielding massive omics datasets. The metagenome, meta-transcriptome, meta-proteome, meta-metabolome, and culturomics are examples of these omics datasets (Sarangi et al. 2019). Researchers discovered a link between diseases and microbiota, particularly in the gut microbiota, using these single omics datasets. For instance, the metagenome is enhancing our understanding of the taxonomic composition of species and the gene content or functional characteristics of microbial communities. Researchers developed a protocol for conducting a metagenome-wide association study using datasets of gut microbial communities obtained from 345 Chinese individuals and identified approximately 60,000 markers associated with type 2 diabetes (Qin et al. 2012). A previous study has demonstrated that the dysbiosis of gut microbiota contributed to the development of hypertension based on 196 gut microbiota metagenome datasets, including 41 healthy controls, 56 individuals with pre-hypertension, and 99 individuals with primary hypertension (Li et al. 2017). Similarly, we expanded the analysis depth based on these 196 metagenome datasets and investigated the function of viruses in the development of hypertension

(Han et al. 2018). As a supplementary, meta-transcriptome can provide the information of active bacteria and a wealth of knowledge about the expression of their genes in a microbial community (White et al. 2016). A meta-transcriptomic analysis of feces from 12 macaques with idiopathic chronic diarrhea and 12 matched healthy controls revealed that the expression of genes involved in inflammation and transcripts from several bacterial pathogens, including *Campylobacter*, *Helicobacter*, and the protozoan *Trichomonas*, were increased (Westreich et al. 2019). Additionally, meta-proteomics aims to profile the metabolic activities of the microbiota within a microbial community, which complements other omics approaches and can establish a direct link between microbial communities' genetic potential and functional metabolism (Abraham et al. 2014; Wang et al. 2020a). Numerous studies have concentrated on the human gut microbiota's meta-proteome, identifying and quantifying several peptides, and associating several peptides with diseases (Xiong et al. 2015). For example, a previous study identified and quantified 91,902 peptides and 341 proteins as biomarkers for colorectal cancer patients compared to healthy controls (Long et al. 2020). Additionally, meta-metabolomics aims to characterize the composition and dynamics of metabolites in biological samples, and has emerged as a technique for defining host-microbial relationships (Lee et al. 2019). Microbiome studies focusing on the gut microbiota's meta-metabolome have increased over the last decade, demonstrating that several metabolites can be used as biomarkers for disease clinical diagnosis (Vernocchi et al. 2016; Zhang et al. 2019).

In summary, the development of sequencing and mass spectrometry techniques resulted in the generation of numerous omics datasets and, in most cases, a single omics dataset for microbiome research. While researchers were able to gain a thorough understanding of the gut microbiome using these single omics datasets, the result was not comprehensive and did not resolve the causality of diseases, limiting the accuracy of clinical diagnosis and treatment. Thus, it is critical to choose two or more microbiome approaches in order to obtain multi-omics datasets and conduct a comprehensive analysis to examine the correlations, causal relationships, and network relationships between gut microbiota and diseases.

In general, multi-omics approaches to the microbiome are a synthesis of several different omics approaches used in microbiome research, and its numerous omics datasets are used to deduce the pattern of a biological process. Additionally, multi-omics datasets encompass a variety of high-dimensional biological datasets, such as 16S rRNA amplicon sequencing data, metagenomic data, metagenome data, meta-transcriptome data, meta-proteome data, meta-metabolome data, and culturomics data (Jiang et al. 2019). Multi-omics data analysis can currently shed light on the relationship between gut microbiota and disease, provide an in-depth understanding of disease occurrence and progression, and may even resolve disease causality.

At the moment, many researchers are working on a method to obtain multi-omics datasets associated with host diseases and gut microbiota, to conduct data mining on multi-omics datasets from a variety of perspectives in order to reveal correlations between diseases and gut microbiota, between microorganisms, and to demonstrate disease causality. For instance, the application of multi-omics approaches to fecal and serum samples collected from 110 healthy individuals in two locations in India, including 16S rRNA amplicon sequencing, whole-genome shotgun metagenomics sequencing, and metabolomics, revealed a unique composition of the Indian gut microbiome, particularly for the gut microbial gene catalog, and provided novel insight into the gut-microbe-metabolic axis (Dhakan et al. 2019). Multi-omics analysis of 1640 samples from 16S rRNA gene datasets and 26 samples from metagenomics datasets from patients with chronic obstructive pulmonary disease (COPD) revealed that 12 microbial genera can be used as COPD biomarkers. Additionally, researchers inferred the metabolic potential of the airway microbiome, linked these biomarkers to host targets, and investigated the effects of COPD patient separation, demonstrating the feasibility of integrating multi-omics datasets to probe disease biology (Wang et al. 2020b). Totally, in the past decade, multi-omics datasets analyses have revealed significant associations between gut microbiota and diseases, including obesity, diabetes, IBD, and so on. Particularly, several studies have demonstrated the causative roles for the gut microbiome in the occurrence and development of diseases, such as dietary fibers alleviate type 2 diabetes (Zhao et al. 2018) and obesity (Fei and Zhao 2013). Thus, we believe that integrating multi-omics datasets and conducting correlation, causal, and network analyses is critical for gaining a better understanding of the gut microbiota, resolving disease causation, and providing clues for clinical diagnosis and treatment. In the future, integration multi-omics data analysis will become increasingly popular in microbiome-disease research.

Besides, the growing number of articles focused on the methodology for integrating the multi-omics datasets and provided the methods or tools to conduct the integration analysis, which propose promising solutions to multi-omics datasets and speed up its application in microbiome research. Specifically, firstly, the tools developed for analyzing the single-omics data, such as for 16S rRNA amplicon sequencing data, QIIME (Caporaso et al. 2010), QIIME2 (Bolyen et al. 2019), Mothur (Schloss et al. 2009), Vsearch (Rognes et al. 2016), DADA2 (Callahan et al. 2016), Deblur (Amir et al. 2017), Parallel-META3 (Jing et al. 2017), etc., were developed for profiling the taxonomical composition of the microbial community. Similarly, PICRUSt (Langille et al. 2013), PICRUSt2 (Douglas et al. 2020), Tax4Fun (Aßhauer et al. 2015) were developed for predicting the functional composition, while Bugbase was developed for predicting the phenotypic composition based on 16S rRNA amplicon sequencing data. As to metagenomic data, Kraken (Wood and Salzberg 2014), Prokka (Seemann 2014), MOCAT2 (Kultima et al. 2016), and MetaPhlAn2 (Truong et al. 2015) was designed and developed for profiling the taxonomical composition, while HUMANn2 was used for predicting the functional composition of a microbial community based on metagenome and metatranscriptomes (Franzosa et al. 2018). Moreover, the tools for analyzing the meta-metabolome data were also established and developed, such as MetaboAnalyst (Xia et al. 2015), MetaBox (Wanichthanarak et al. 2017), MetaCoreTM, InCroMAP, and 3Omics (Cambiaghi et al. 2017). Besides, several powerful bioinformatics tools, such as Meta-Proteome-Analyzer (Muth et al. 2012), MPA portable (Muth et al. 2018), MetaProteomeAnalyzer (Muth et al. 2015), have been developed for analyzing the meta-proteome data and providing the interpretation of proteins of microbial communities. Secondly, several computational frameworks were established and developed to integrate high-throughput omics datasets and obtain an in-depth understanding of the gut microbiome with different perspectives. For example, a previous study has integrated human gut microbiome data (WGS data), untargeted serum metabolome data, and measures of host physiology provided a framework, which was established in R platform and can be applied to other investigations and demonstrated the potential mechanistic links (Pedersen et al. 2018). As to the results of integrated multi-omics datasets, several articles conducted the network analysis and given the interpretations with network perspective (Ramos et al. 2019; Misra et al. 2019; Tuncbag et al. 2016; Yan et al. 2018), which provide great potential to solve the interactions among themselves and with their environment.

In comparison to traditional single-omics data analysis, multi-omics datasets can provide additional information about microbial communities, allow for in-depth exploration of the interactions between microbiota and other features, and demonstrate the relationship between microbiota and diseases, particularly when investigating disease causation. We believe that as high-throughput technologies are developed and methods and tools for the joint analysis of multi-omics datasets are established, multi-omics datasets will be obtained from an increasing number of microbiome researches. Correlation analysis, causal analysis, and network analysis can all be used to gain a better understanding of the gut microbiota and to aid in the clinical diagnosis and treatment of diseases using these multi-omics datasets.

6.5 The Application of Multi-Omics Approaches to Inflammation Diseases and its Clinical Treatment with Microbiome Approaches

Owing to the advantages of multi-omics approaches, it has been widely applied to gut microbiome research in inflammation diseases. For example, based on multi-omics datasets of inflammatory bowel diseases (IBD), including metagenomes, metatranscriptomes, proteomes, metabolomes, and viromes of IBD patients and healthy controls, researchers demonstrated the abundance of facultative anaerobes increased, as well as an increase in molecular disruptions in microbial transcription, such as *Clostridia*, and several metabolites, such as acylcarnitines, bile acids, and short-chain fatty acids, which provide a comprehensive description of the host and gut microbial activities in IBD (Lloyd-Price et al. 2019). In the following study, the authors proposed cutting-edge methodologies with a unique computational toolbox for multi-omics datasets analysis, which benefits the biomarker discovery (Metwaly and Haller 2019). As to RA, previous studies have been revealed that the gut microbiota plays an important role in the occurrence and development of RA (Kishikawa et al. 2020; Xu et al. 2020; Van de Wiele et al. 2016), and the alterations of the gut microbiome and oral microbiome between RA patients and healthy controls (Zhang et al. 2015b). Specifically, for the purpose of examining the relationship between gut microbiota and rheumatoid arthritis, a previous review summarized the primary and most recent applications of multi-omics approaches in human rheumatoid arthritis research, which uncovers new avenues for preventing and managing rheumatoid arthritis (Cassotta et al. 2021). Concerning SLE, a previous article collected fecal samples from 117 untreated SLE patients and 52 post-treated SLE patients and compared them to 115 matched healthy controls to obtain metagenome datasets for these samples. The results of this study indicated that the composition of the gut microbiota of individuals with SLE differs significantly from that of healthy controls, and that several species, including *Clostridium* ATCC BAA-442, *Clostridium leptum*, *Atopobium rimae*, *Shuttleworthia satelles*, *Actinomyces massiliensis*, and *Bacteroides fragilis*, were enriched in SLE patients and decreased after treatment, suggesting that they could be used as biomarkers. Additionally, the authors assessed and validated the findings using the mouse fecal metagenome (Chen et al. 2021). In total, various inflammation diseases have been associated with changes in the composition and function of the gut microbiota, and emerging evidence suggests that the distribution of the gut microbiota plays a role in the occurrence and development of inflammation diseases.

Thus, based on their understanding of the gut microbiota's role in inflammation diseases, researchers have proposed several microbiome approaches, including fecal microbiota transplantation (FMT), probiotic bacteria, and dietary habit modification, for shaping the taxonomical and functional compositions of the human gut microbiome and treating inflammation diseases. For example, numerous case reports and cohort studies on inflammatory bowel disease have described the clinical use of FMT in inflammatory bowel disease patients (Lopez and Grinspan 2016) and this method has gained interest as a novel treatment approach for IBD (Colman and Rubin 2014). Additionally, an increasing number of articles examined the use of probiotic bacteria in the treatment of inflammatory diseases and evaluated the method's therapeutic efficacy. For example, previous research demonstrated that restoring the gut microbiome of a single bacterium, *Lactobacillus casei* (ATCC334), significantly suppresses the induction of adjuvant-induced arthritis and protects the bones in rheumatoid arthritis mice, indicating that probiotic bacteria may be a promising treatment option for rheumatoid arthritis (Pan et al. 2019). Similarly, a randomized double-blind clinical trial of RA revealed that the supplementation of *L. casei* (LC01) can affect the disease activity and inflammatory cytokines in RA (Alipour et al. 2014). However, as to the efficacy of probiotic bacteria, many researchers hold the opposite opinion (Aqaeinezhad Rudbane et al. 2018; Mohammed et al. 2017; Pan et al. 2017) and whether the effect of probiotic bacteria in inflammation diseases, especially for RA is still under verification. So, more microbiome researches should be paid attention to verify it in the future.

Together, multi-omics data analysis for correlation analysis, causal analysis, and network analysis of inflammation diseases is critical to elucidate the underlying mechanism and is urgently needed to identify novel and precise diagnostic biomarkers and clinical treatment strategies utilizing microbiome approaches.

6.6 Conclusion

Numerous studies have established that the human gut microbiota is critical for human health maintenance and that dysbiosis of the human gut microbiome contributes to the occurrence and development of human diseases. Due to the importance of the human gut microbiome, many microbiome scientists refer to it as the body’s second genome. Although numerous microbiome studies have been conducted, the detailed mechanisms of disease and the causal relationship between disease and gut microbiota remain elusive. Due to the advancement of sequencing and mass spectrometry techniques, multi-omics approaches have been applied to microbiome research, and integrated analysis of multi-omics datasets has been performed to gain a better understanding of the gut microbiota and disease. We chose inflammation disease as a model disease in this review to illustrate the relationship between inflammation disease and gut microbiota, particularly for multi-omics approaches and bioinformatics methods for integrating multi-omics datasets and their applications in inflammation diseases. Our findings summarized the roles of gut microbiota in the occurrence and development of inflammation diseases such as inflammatory bowel disease, rheumatoid arthritis, and systemic lupus erythematosus, discussed the benefits of multi-omics approaches, introduced bioinformatics methods for dealing with multi-omics datasets, and described clinical strategies utilizing microbiome approaches for treating inflammation diseases. Our comprehensive review of multi-omics approaches and their application may shed light on the clinical relevance of the human gut microbiome and inflammation diseases, as well as identify clades for developing an inflammation disease therapeutic schedule.

References

Abraham PE, et al. Metaproteomics: extracting and mining proteome information to characterize metabolic activities in microbial communities. Curr Protoc Bioinformatics. 2014;46(1):13.26. 1-13.26. 14

Abt MC, McKenney PT, Pamer EG. Clostridium difficile colitis: pathogenesis and host defence. Nat Rev Microbiol. 2016;14(10):609–20.

Alamanos Y, Drosos AA. Epidemiology of adult rheumatoid arthritis. Autoimmun Rev. 2005;4(3): 130–6.

Aletaha D, Smolen JS. Diagnosis and Management of Rheumatoid Arthritis: a review. JAMA. 2018;320(13):1360–72.

Alipour B, et al. Effects of L actobacillus casei supplementation on disease activity and inflammatory cytokines in rheumatoid arthritis patients: a randomized double-blind clinical trial. Int J Rheum Dis. 2014;17(5):519–27.

Amir A, et al. Deblur rapidly resolves single-nucleotide community sequence patterns. MSystems. 2017;2:2.

Amy IY, et al. Gut microbiota modulate CD8 T cell responses to influence colitis-associated tumorigenesis. Cell Rep. 2020;31(1):107471.

Aqaeinezhad Rudbane SM, et al. The efficacy of probiotic supplementation in rheumatoid arthritis: a meta-analysis of randomized, controlled trials. Inflammopharmacology. 2018;26(1):67–76.

Arumugam M, et al. Enterotypes of the human gut microbiome. Nature. 2011;473(7346):174–80.

Aßhauer KP, et al. Tax4Fun: predicting functional profiles from metagenomic 16S rRNA data. Bioinformatics. 2015;31(17):2882–4.

Bäckhed F, et al. Dynamics and stabilization of the human gut microbiome during the first year of life. Cell Host Microbe. 2015;17(5):690–703.

Benítez-Páez A, et al. A multi-omics approach to unraveling the microbiome-mediated effects of arabinoxylan oligosaccharides in overweight humans. mSystems. 2019;4(4):e00209–19.

Bolyen E, et al. Reproducible, interactive, scalable and extensible microbiome data science using QIIME 2. Nat Biotechnol. 2019;37(8):852–7.

Borchers AT, et al. The geoepidemiology of systemic lupus erythematosus. Autoimmun Rev. 2010;9(5):A277–87.

Bullock J, et al. Rheumatoid arthritis: a brief overview of the treatment. Med Princ Pract. 2018;27 (6):501–7.

Callahan BJ, et al. DADA2: high-resolution sample inference from Illumina amplicon data. Nat Methods. 2016;13(7):581–3.

Cambiaghi A, Ferrario M, Masseroli M. Analysis of metabolomic data: tools, current strategies and future challenges for omics data integration. Brief Bioinform. 2017;18(3):498–510.

Cani PD. Human gut microbiome: hopes, threats and promises. Gut. 2018;67(9):1716–25.

Caporaso JG, et al. QIIME allows analysis of high-throughput community sequencing data. Nat Methods. 2010;7(5):335–6.

Caporaso JG, et al. Global patterns of 16S rRNA diversity at a depth of millions of sequences per sample. Proc Natl Acad Sci. 2011;108(Supplement 1):4516–22.

Cassotta M, et al. Nutrition and rheumatoid arthritis in the 'omics' era. Nutrients. 2021;13(3):763.

Chen B, et al. The gut microbiota of non-treated patients with SLE defines an autoimmunogenic and proinflammatory profile. Hoboken NJ: Arthritis Rheumatol; 2020.

Chen BD, et al. An autoimmunogenic and proinflammatory profile defined by the gut microbiota of patients with untreated systemic lupus erythematosus. Arthritis Rheumatol. 2021;73(2):232–43.

Chen S, et al. Linkages of firmicutes and Bacteroidetes populations to methanogenic process performance. J Ind Microbiol Biotechnol. 2016;43(6):771–81.

Clemente JC, et al. The impact of the gut microbiota on human health: an integrative view. Cell. 2012;148(6):1258–70.

Colman RJ, Rubin DT. Fecal microbiota transplantation as therapy for inflammatory bowel disease: a systematic review and meta-analysis. J Crohn's Colitis. 2014;8(12):1569–81.

Dekaboruah E, et al. Human microbiome: an academic update on human body site specific surveillance and its possible role. Arch Microbiol. 2020;202(8):2147–67.

Dhakan DB, et al. The unique composition of Indian gut microbiome, gene catalogue, and associated fecal metabolome deciphered using multi-omics approaches. GigaScience. 2019;8:3.

Douglas GM, et al. PICRUSt2 for prediction of metagenome functions. Nat Biotechnol. 2020;38 (6):685–8.

Duan R, et al. Alterations of gut microbiota in patients with irritable bowel syndrome based on 16S rRNA-targeted sequencing: a systematic review. Clin Transl Gastroenterol. 2019;10:2.

Fan Y, Pedersen O. Gut microbiota in human metabolic health and disease. Nat Rev Microbiol. 2020:1–17.

Fei N, Zhao L. An opportunistic pathogen isolated from the gut of an obese human causes obesity in germfree mice. ISME J. 2013;7(4):880–4.

Fernandes J, et al. Adiposity, gut microbiota and faecal short chain fatty acids are linked in adult humans. Nutr Diabetes. 2014;4(6):e121.

Franzosa EA, et al. Species-level functional profiling of metagenomes and metatranscriptomes. Nat Methods. 2018;15(11):962–8.

Gaulke CA, Sharpton TJ. The influence of ethnicity and geography on human gut microbiome composition. Nat Med. 2018;24(10):1495–6.

Gevers D, et al. The human microbiome project: a community resource for the healthy human microbiome. PLoS Biol. 2012;10(8):e1001377.

Ghaisas S, Maher J, Kanthasamy A. Gut microbiome in health and disease: linking the microbiome–gut–brain axis and environmental factors in the pathogenesis of systemic and neurodegenerative diseases. Pharmacol Ther. 2016;158:52–62.

Glasner ME. Finding enzymes in the gut metagenome. Science. 2017;355(6325):577–8.

Gorvitovskaia A, Holmes SP, Huse SM. Interpreting Prevotella and Bacteroides as biomarkers of diet and lifestyle. Microbiome. 2016;4(1):1–12.

Guerrini MM, Vogelzang A, Fagarasan S. A hen in the wolf Den: a pathobiont tale. Immunity. 2018;48(4):628–31.

Hamady M, Knight R. Microbial community profiling for human microbiome projects: tools, techniques, and challenges. Genome Res. 2009;19(7):1141–52.

Han M, et al. The human gut virome in hypertension. Front Microbiol. 2018;9:3150.

Han M, et al. Stratification of athletes' gut microbiota: the multifaceted hubs associated with dietary factors, physical characteristics and performance. Gut Microbes. 2020a;12(1):1–18.

Han M, et al. Comparative genomics uncovers the genetic diversity and characters of Veillonella atypica and provides insights into its potential applications. Front Microbiol. 2020b;11:1219.

Hansen RL, et al. Nanoparticle microarray for high-throughput microbiome metabolomics using matrix-assisted laser desorption ionization mass spectrometry. Anal Bioanal Chem. 2019;411 (1):147–56.

Harley IT, Karp CL. Obesity and the gut microbiome: striving for causality. Mol Metab. 2012;1 (1–2):21–31.

Heintz-Buschart A, et al. Integrated multi-omics of the human gut microbiome in a case study of familial type 1 diabetes. Nat Microbiol. 2016;2(1):16180.

Horta-Baas G, et al. Intestinal dysbiosis and rheumatoid arthritis: a link between gut microbiota and the pathogenesis of rheumatoid arthritis. J Immunol Res. 2017;2017

Huang L, et al. dbCAN-seq: a database of carbohydrate-active enzyme (CAZyme) sequence and annotation. Nucleic Acids Res. 2018;46(D1):D516–21.

Integrative H. The Integrative human microbiome project: dynamic analysis of microbiome-host omics profiles during periods of human health and disease. Cell Host Microbe. 2014;16(3): 276–89.

Integrative H, et al. The integrative human microbiome project. Nature. 2019;569(7758):641–8.

Isaacs-Ten A, et al. Intestinal microbiome-macrophage crosstalk contributes to cholestatic liver disease by promoting intestinal permeability in mice. Hepatology (Baltimore, Md). 2020;72(6): 2090.

Jia B, et al. CARD 2017: expansion and model-centric curation of the comprehensive antibiotic resistance database. Nucleic Acids Res. 2016:gkw1004.

Jiang D, et al. Microbiome multi-omics network analysis: statistical considerations, limitations, and opportunities. Front Genet. 2019;10:995.

Jing G, et al. Parallel-META 3: comprehensive taxonomical and functional analysis platform for efficient comparison of microbial communities. Sci Rep. 2017;7(1):1–11.

Kho ZY, Lal SK. The human gut microbiome—a potential controller of wellness and disease. Front Microbiol. 2018;9:1835.

Kishikawa T, et al. Metagenome-wide association study of gut microbiome revealed novel aetiology of rheumatoid arthritis in the Japanese population. Ann Rheum Dis. 2020;79(1): 103–11.

Kulecka M, et al. The composition and richness of the gut microbiota differentiate the top polish endurance athletes from sedentary controls. Gut Microbes. 2020;11(5):1374–84.

Kultima JR, et al. MOCAT2: a metagenomic assembly, annotation and profiling framework. Bioinformatics. 2016;32(16):2520–3.

Kurilshikov A, et al. Large-scale association analyses identify host factors influencing human gut microbiome composition. Nat Genet. 2021;53(2):156–65.

Langille MG, et al. Predictive functional profiling of microbial communities using 16S rRNA marker gene sequences. Nat Biotechnol. 2013;31(9):814–21.

Lee H-J, et al. Meta-analysis of targeted metabolomics data from heterogeneous biological samples provides insights into metabolite dynamics. bioRxiv. 2019:509372.

Li J, et al. Gut microbiota dysbiosis contributes to the development of hypertension. Microbiome. 2017;5(1):14.

Liu H, et al. Resilience of human gut microbial communities for the long stay with multiple dietary shifts. Gut. 2019;68(12):2254–5.

Lloyd-Price J, et al. Strains, functions and dynamics in the expanded human microbiome project. Nature. 2017;550(7674):61–6.

Lloyd-Price J, et al. Multi-omics of the gut microbial ecosystem in inflammatory bowel diseases. Nature. 2019;569(7758):655–62.

Long S, et al. Metaproteomics characterizes human gut microbiome function in colorectal cancer. npj Biofilms Microbiomes. 2020;6(1):14.

Lopez J, Grinspan A. Fecal microbiota transplantation for inflammatory bowel disease. Gastroenterol Hepatol. 2016;12(6):374–9.

Maeda Y, Takeda K. Role of gut microbiota in rheumatoid arthritis. J Clin Med. 2017;6(6):60.

Magne F, et al. The firmicutes/Bacteroidetes ratio: a relevant marker of gut dysbiosis in obese patients? Nutrients. 2020;12(5):1474.

Majithia V, Geraci SA. Rheumatoid arthritis: diagnosis and management. Am J Med. 2007;120 (11):936–9.

Manasson J, Blank RB, Scher JU. The microbiome in rheumatology: where are we and where should we go? Ann Rheum Dis. 2020;79(6):727–33.

Manzo VE, Bhatt AS. The human microbiome in hematopoiesis and hematologic disorders. Blood. 2015;126(3):311–8.

Marchesi JR, et al. The gut microbiota and host health: a new clinical frontier. Gut. 2016;65(2): 330–9.

Marietta EV, et al. Suppression of inflammatory arthritis by human gut-derived Prevotella histicola in humanized mice. Arthritis Rheumatol. 2016;68(12):2878–88.

Metwaly A, Haller D. Multi-omics in IBD biomarker discovery: the missing links. Nat Rev Gastroenterol Hepatol. 2019;16(10):587–8.

Misra BB, et al. Integrated omics: tools, advances and future approaches. J Mol Endocrinol. 2019;62(1):R21–45.

Mithieux G. Does Akkermansia muciniphila play a role in type 1 diabetes? Gut. 2018;67(8): 1373–4.

Mohammed AT, et al. The therapeutic effect of probiotics on rheumatoid arthritis: a systematic review and meta-analysis of randomized control trials. Clin Rheumatol. 2017;36(12):2697–707.

Mushtaq A. New clinical recommendations for Clostridium difficile. Lancet Infect Dis. 2018;18(4): 384.

Muth T, et al. *Meta-Proteome-Analyzer: A software tool specifically developed for the functional and taxonomic characterization of metaproteome data*. in *GCB2012: German conference on bioinformatics*. 2012.

Muth T, et al. The MetaProteomeAnalyzer: a powerful open-source software suite for metaproteomics data analysis and interpretation. J Proteome Res. 2015;14(3):1557–65.

Muth T, et al. MPA portable: a stand-alone software package for analyzing metaproteome samples on the go. Anal Chem. 2018;90(1):685–9.

Neuman H, Koren O. The gut microbiota: a possible factor influencing systemic lupus erythematosus. Curr Opin Rheumatol. 2017;29(4):374–7.

Pan H, et al. Whether probiotic supplementation benefits rheumatoid arthritis patients: a systematic review and meta-analysis. Engineering. 2017;3(1):115–21.

Pan H, et al. A single bacterium restores the microbiome dysbiosis to protect bones from destruction in a rat model of rheumatoid arthritis. Microbiome. 2019;7(1):107.

Pedersen HK, et al. A computational framework to integrate high-throughput '-omics' datasets for the identification of potential mechanistic links. Nat Protoc. 2018;13(12):2781–800.

Pianta A, et al. Evidence of the immune relevance of Prevotella copri, a gut microbe, in patients with rheumatoid arthritis. Arthritis Rheumatol. 2017;69(5):964–75.

Proctor LM. The human microbiome project in 2011 and beyond. Cell Host Microbe. 2011;10(4): 287–91.

Prodan A, et al. Comparing bioinformatic pipelines for microbial 16S rRNA amplicon sequencing. PLoS One. 2020;15(1):e0227434.

Qin J, et al. A metagenome-wide association study of gut microbiota in type 2 diabetes. Nature. 2012;490(7418):55–60.

Ramos PIP, et al. Leveraging user-friendly network approaches to extract knowledge from high-throughput omics datasets. Front Genet. 2019;10:1120.

Rognes T, et al. VSEARCH: a versatile open source tool for metagenomics. PeerJ. 2016;4:e2584.

Romo-Vaquero M, et al. Deciphering the human gut microbiome of urolithin metabotypes: association with enterotypes and potential cardiometabolic health implications. Mol Nutr Food Res. 2019;63(4):1800958.

Round JL, Mazmanian SK. The gut microbiota shapes intestinal immune responses during health and disease. Nat Rev Immunol. 2009;9(5):313–23.

Sandhu BK, McBride SM. Clostridioides difficile. Trends Microbiol. 2018;26(12):1049–50.

Santos LL, Morand EF. Macrophage migration inhibitory factor: a key cytokine in RA, *SLE and atherosclerosis*. Clin Chimica Acta. 2009;399(1–2):1–7.

Sarangi AN, Goel A, Aggarwal R. Methods for studying gut microbiota: a primer for physicians. J Clin Exp Hepatol. 2019;9(1):62–73.

Scheiman J, et al. Meta-omics analysis of elite athletes identifies a performance-enhancing microbe that functions via lactate metabolism. Nat Med. 2019;25(7):1104–9.

Scher JU, et al. Expansion of intestinal Prevotella copri correlates with enhanced susceptibility to arthritis. elife. 2013;2:e01202.

Schloss PD, et al. Introducing mothur: open-source, platform-independent, community-supported software for describing and comparing microbial communities. Appl Environ Microbiol. 2009;75(23):7537–41.

Schwiertz A, et al. Microbiota and SCFA in lean and overweight healthy subjects. Obesity. 2010;18 (1):190–5.

Seemann T. Prokka: rapid prokaryotic genome annotation. Bioinformatics. 2014;30(14):2068–9.

Sharon G, et al. The central nervous system and the gut microbiome. Cell. 2016;167(4):915–32.

Shin N-R, Whon TW, Bae J-W. Proteobacteria: microbial signature of dysbiosis in gut microbiota. Trends Biotechnol. 2015;33(9):496–503.

Silverman GJ, Azzouz DF, Alekseyenko AV. Systemic lupus erythematosus and dysbiosis in the microbiome: cause or effect or both? Curr Opin Immunol. 2019;61:80–5.

Slowicka K, et al. Zeb2 drives invasive and microbiota-dependent colon carcinoma. Nat Cancer. 2020;1(6):620–34.

Spor A, Koren O, Ley R. Unravelling the effects of the environment and host genotype on the gut microbiome. Nat Rev Microbiol. 2011;9(4):279–90.

Sutoyo DA, Atmaka DR, Sidabutar LMG. Dietary factors affecting firmicutes and Bacteroidetes ratio in solving obesity problem: a literature review. Media Gizi Indonesia. 2020;15(2):94–109.

Thursby E, Juge N. Introduction to the human gut microbiota. Biochem J. 2017;474(11):1823–36.

Tong Y, et al. Oral microbiota perturbations are linked to high risk for rheumatoid arthritis. Front Cell Infect Microbiol. 2020;9:475.

Torres J, et al. Infants born to mothers with IBD present with altered gut microbiome that transfers abnormalities of the adaptive immune system to germ-free mice. Gut. 2020;69(1):42–51.

Truong DT, et al. MetaPhlAn2 for enhanced metagenomic taxonomic profiling. Nat Methods. 2015;12(10):902–3.

Tu Q, et al. NCycDB: a curated integrative database for fast and accurate metagenomic profiling of nitrogen cycling genes. Bioinformatics. 2019;35(6):1040–8.

Tuncbag N, et al. Network-based interpretation of diverse high-throughput datasets through the omics integrator software package. PLoS Comput Biol. 2016;12(4):e1004879.

Turnbaugh PJ, et al. The human microbiome project. Nature. 2007;449(7164):804–10.

Valdes AM, et al. Role of the gut microbiota in nutrition and health. BMJ. 2018;361:k2179.

Van de Wiele T, et al. How the microbiota shapes rheumatic diseases. Nat Rev Rheumatol. 2016;12 (7):398.

Vernocchi P, Del Chierico F, Putignani L. Gut microbiota profiling: metabolomics based approach to unravel compounds affecting human health. Front Microbiol. 2016;7:1144.

Walter J, et al. Establishing or exaggerating causality for the gut microbiome: lessons from human microbiota-associated rodents. Cell. 2020;180(2):221–32.

Wang Y, et al. Metaproteomics: a strategy to study the taxonomy and functionality of the gut microbiota. J Proteome. 2020a;219:103737.

Wang Z, et al. Multi-omic meta-analysis identifies functional signatures of airway microbiome in chronic obstructive pulmonary disease. ISME J. 2020b;14(11):2748–65.

Wanichthanarak K, et al. Metabox: a toolbox for metabolomic data analysis, interpretation and integrative exploration. PLoS One. 2017;12(1):e0171046.

Westreich ST, et al. Fecal metatranscriptomics of macaques with idiopathic chronic diarrhea reveals altered mucin degradation and fucose utilization. Microbiome. 2019;7(1):41.

White RA, et al. The past, present and future of microbiome analyses. Nat Protoc. 2016;11(11): 2049–53.

Wood DE, Salzberg SL. Kraken: ultrafast metagenomic sequence classification using exact alignments. Genome Biol. 2014;15(3):1–12.

Wu GD, et al. Linking long-term dietary patterns with gut microbial enterotypes. Science. 2011;334 (6052):105–8.

Xia J, et al. MetaboAnalyst 3.0—making metabolomics more meaningful. Nucleic Acids Res. 2015;43(W1):W251–7.

Xiong W, et al. Microbial metaproteomics for characterizing the range of metabolic functions and activities of human gut microbiota. Proteomics. 2015;15(20):3424–38.

Xu H, et al. Interactions between gut microbiota and immunomodulatory cells in rheumatoid arthritis. Mediat Inflamm. 2020;2020

Yan J, et al. Network approaches to systems biology analysis of complex disease: integrative methods for multi-omics data. Brief Bioinform. 2018;19(6):1370–81.

Yong P, et al. Management of hypogammaglobulinaemia occurring in patients with systemic lupus erythematosus. Rheumatology. 2008;47(9):1400–5.

Zhang J, et al. A phylo-functional core of gut microbiota in healthy young Chinese cohorts across lifestyles, geography and ethnicities. ISME J. 2015a;9(9):1979–90.

Zhang X, et al. The oral and gut microbiomes are perturbed in rheumatoid arthritis and partly normalized after treatment. Nat Med. 2015b;21(8):895–905.

Zhang X, et al. Advancing functional and translational microbiome research using meta-omics approaches. Microbiome. 2019;7(1):154.

Zhao L, et al. Gut bacteria selectively promoted by dietary fibers alleviate type 2 diabetes. Science. 2018;359(6380):1151–6.

Zhou J, et al. High-throughput metagenomic technologies for complex microbial community analysis: open and closed formats. MBio. 2015;6:1.

Chapter 7
Microbiome Data Analysis and Interpretation: Correlation Inference and Dynamic Pattern Discovery

Kang Ning and Yuxue Li

Microbial communities are everywhere within our bodies and in the environments (Byrd et al. 2018), which play a key role in human health and all critical nutrient cycles on earth. The microbiome refers to the entire micro-environment, including microorganisms, genomes, and the surrounding environment. With the development of high-throughput sequencing (HTS) technology and data analysis methods, the role of the microbiome in humans, animals, plants, and the environment has become increasingly clear in recent years.

7.1 Microbiome and its Importance

Microbes are everywhere. The origin of life on earth began with microorganisms, which greatly promoted the evolution of the earth. Environmental microorganisms tend to vary from place to place, and there is an enrichment phenomenon in a specific environment, and the composition of human microorganisms also has similar characteristics. So far, scientists have mainly studied microorganisms that cause diseases.

In early research, scientists discovered that the human body needs to coexist peacefully with trillions of microbes, and called it the "microbiome". Later, microbial community refers to the collection of all microbes and their genetic information in a certain environment or ecosystem. Figure 7.1 shows the timeline of the microbiome research.

K. Ning (✉) · Y. Li
Key Laboratory of Molecular Biophysics of the Ministry of Education, Hubei Key Laboratory of Bioinformatics and Molecular-imaging, Center of Artificial Intelligence Biology, Department of Bioinformatics and Systems Biology, College of Life Science and Technology, Huazhong University of Science and Technology, Wuhan, China
e-mail: ningkang@hust.edu.cn

K. Ning (ed.), *Methodologies of Multi-Omics Data Integration and Data Mining*,
Translational Bioinformatics 19, https://doi.org/10.1007/978-981-19-8210-1_7

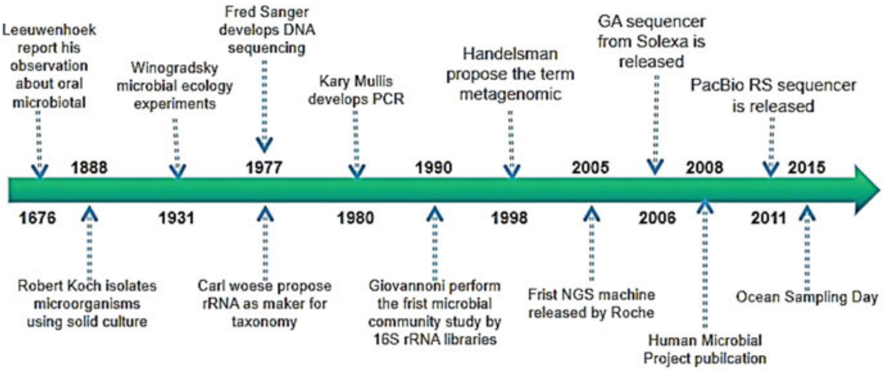
Leeuwenhoek report his observation about oral microbiotal
Winogradsky microbial ecology experiments
Fred Sanger develops DNA sequencing
Kary Mullis develops PCR
Handelsman propose the term metagenomic
GA sequencer from Solexa is released
PacBio RS sequencer is released
1888
1977
1990
2005
2008
2015
1676
1931
1980
1998
2006
2011
Robert Koch isolates microorganisms using solid culture
Carl woese propose rRNA as maker for taxonomy
Giovannoni perform the frist microbial community study by 16S rRNA libraries
Frist NGS machine released by Roche
Human Microbial Project pubilcation
Ocean Sampling Day

Fig. 7.1 A brief timeline of microbiome research. Microbiome studies could be traced back to hundreds of years, but only by use of high throughput sequencing these studies have discovered millions of genes and species that have never been identified before

In addition, there are many professional terms in microbiome research. Table 7.1 explains some professional terms.

Microbiome is one of the emerging omics studies with profound biomedical applications. It is now already known that many biomedical applications are related to the microbiome (Fig. 7.2).

7.2 Experimental and Analytical Approaches for Microbiome Researches

7.2.1 Metagenomics

Metagenome was proposed by Handelsman et al. (Di Bella et al. 2013) in 1998 and defined as “the genomes of the total microbiota found in nature”, which refers to the sum of the genetic material of all microbiota found in the environment (Nowrotek et al. 2019). It contains the genes of cultivable and non-cultivable microbiota, and currently mainly refers to the sum of the genomes of bacteria and fungi in environmental samples.

7.2.1.1 The Differences Between 16S and Metagenome (Ruairi Robertson 2020)

The Sequencing Principles

16S rDNA is the most common “molecular clock” in bacterial taxonomy and is highly conservative. The sequence contains 9 hypervariable regions and

Table 7.1 Terms used in microbiome studies

Basic terms	Introduction
Microbiota	A collection of microorganisms is present in a special environment. First, defined by Lederberg and McCray in 2001, the emphasis is on the importance of human health and disease-related microbes. The microbial composition is studied mainly through molecular methods, based on 16S, 18S rRNA or other marker genes or genome regions for analysis. Biological samples are amplified, sequenced, and finally divided into different classes according to the sequences.
Metataxonomics	The entire microbial population is described through a high-throughput sequencing process and a phylogenetic tree, called a macrotaxonomic tree, which can be used to represent the relationships between all obtained sequences. Although viruses are part of the microbial community, there are no universal viral marker genes that can be used to classify them.
Metagenome	Metagenome refers to the genomes and genes associated with members of a microbial community. The sample DNA is collected by shotgun sequencing and compared with the reference database for annotation.
Microbiome	The microbiome refers to the entire genome (genes) of microorganisms (bacteria, archaea, lower or higher eukaryotes, and viruses) as well as their surrounding environment.
Metabolomics	All metabolites present in a given strain and a single tissue are collectively called the metabolome. Metabolomics is used to characterize the metabolites of a given strain or individual tissue.
Metabonomics	This is a variant of the method metabolomics, which describes the methods used to generate metabolic profiles from complex systems.
Metatranscriptomics	This concept refers to the use of high-throughput sequencing to analyze the corresponding meta-cDNA of RNAs (meta-RNAs) expressed in the microbial community.
Metaproteomics	Refers to a large-scale description of the entire proteome in a specific environment/clinical sample at a specific time. This approach does not distinguish between a protein derived from a microbial population or a host or environmen.

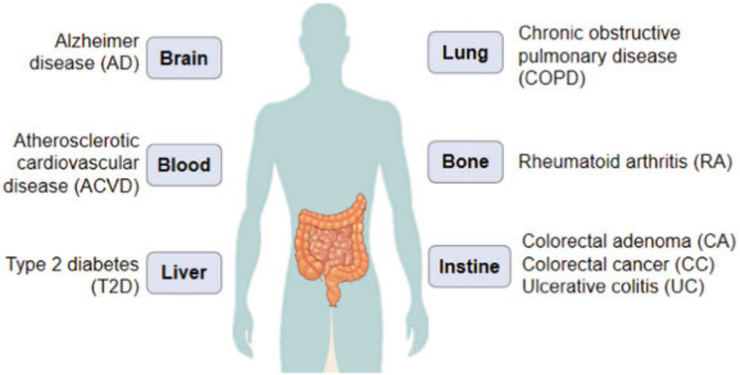

Fig. 7.2 The microbiome-related biomedical applications. The human gut microbiome is tightly connected with brain, blood, liver, lung, bone and intestine systems

10 conserved regions. A sequence of a certain hypervariable region (such as V4 or V3-V4) is amplified by PCR and then sequenced to obtain a sequence of about 200–400 bps. Metagenomic sequencing, like conventional DNA libraries, randomly breaks microbial genome DNA into small fragments, and then adds adapters at both ends of the fragments for high-throughput sequencing.

Different Fields of Study

16S sequencing mainly studies the species composition, the evolutionary relationship among species, and the diversity of communities. Metagenomic sequencing can also be used for further research at the genetic and functional levels.

Different Degrees of Species Identification

Many of the sequences obtained by 16S sequencing are not annotated at the species level, while metagenomic sequencing can identify microbiota to the species level or even the strain level.

Application Fields of Metagenomics

(a) Environmental microbial diversity (Wang et al. 2019).
(b) Gene mining (Vakhlu et al. 2008).
(c) Disease association analysis (Kishikawa et al. 2020).
(d) Drug development (Chiu and Miller 2019).

The Process of Metagenomics Research

Informatics and bioinformatics technologies are needed in all aspects of metagenomics big data analysis. The first is the storage of big data (Papageorgiou et al. 2018). After classification and sorting, the data needs to be stored in a standardized database for subsequent analysis. The second is the pre-processing of big data. The pre-processing of big data is the basis of metagenomics research, and its speed and accuracy will have a great impact on the progress of the experiment and the conclusion. Finally, the data after basic analysis needs further information analysis, comparison, and refinement. The community composition and microbial diversity, community function and genetic variation, community structure and species correlation, community and environment interaction will be analysed (Sangwan et al. 2016). Figure 7.3 briefly summarizes the various processes of metagenomics analysis.

Metagenomics usually focuses on the analysis of microbial diversity, population structure, species evolutionary relationships, gene function activity, community collaboration, and the relationship with the environment (Oulas et al. 2015). The analysis process, database, and software can refer to the following Fig. 7.4:

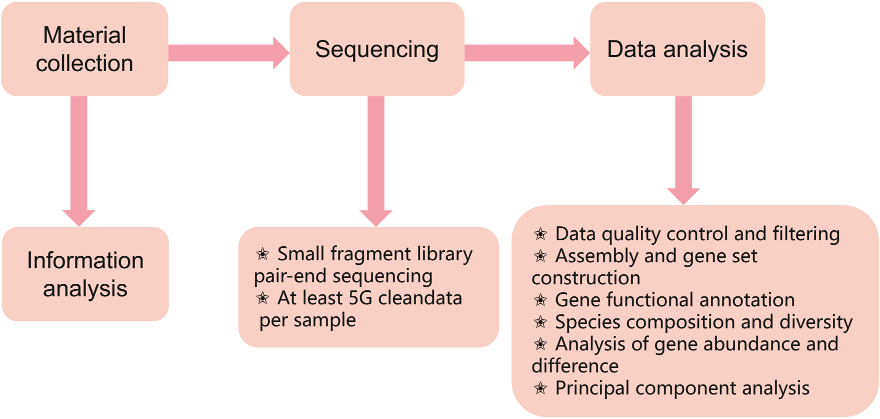
Material collection
Sequencing
Data analysis
Information analysis
✫ Small fragment library pair-end sequencing
✫ At least 5G cleandata per sample
✫ Data quality control and filtering
✫ Assembly and gene set construction
✫ Gene functional annotation
✫ Species composition and diversity
✫ Analysis of gene abundance and difference
✫ Principal component analysis

Fig. 7.3 The processes of metagenomics analysis. Select appropriate research materials, conduct genome sequencing, and perform analysis based on the obtained sequencing data

7.2.2 High-Throughput Sequencing Technology

In 1977, the first generation of DNA sequencing technology (Sanger) came out and has a history of more than 30 years.

High-throughput sequencing (also known as Next Generation Sequencing, NGS) is to randomly fragment DNA (or cDNA) and add adapters to prepare sequencing libraries. Through the extension reaction of tens of thousands of clones in the library, the corresponding signal is detected, and the sequence information is finally obtained. The current main sequencing technology platforms include the Solexa,454, and solid.

7.2.2.1 Application of High-Throughput Sequencing Technology to Species Identification

High-throughput sequencing represents several emerging technologies that are being developed for species identification, but the way they record nucleotide variations are fundamentally different. In addition, these methods have significant differences in throughput, read length, accuracy, and technical deviation.

7.2.2.2 Application of High-Throughput Sequencing Technology to Individual Identification

When targeting SNPs in rapidly evolving sites, partial genomes, or full genomes, the HTS method can be used to distinguish individuals. Early HTS population genomics methods mainly studied the distribution of SNPs in certain variable regions. With the

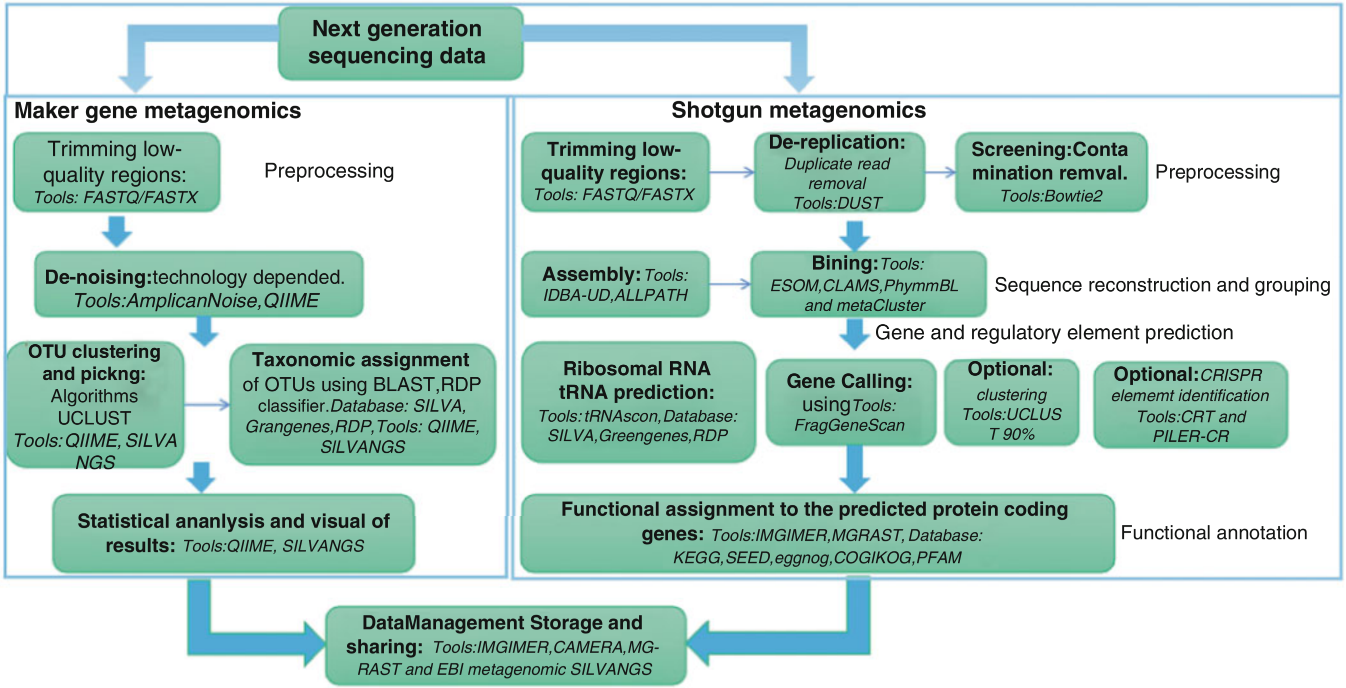
Next generation sequencing data
Maker gene metagenomics
Trimming low-quality regions: Tools: FASTQ/FASTX
Preprocessing
De-noising:technology depended. Tools:AmplicanNoise,QIIME
OTU clustering and pickng: Algorithms UCLUST Tools:QIIME, SILVA NGS
Taxonomic assignment of OTUs using BLAST,RDP classifier.Database: SILVA, Grangenes,RDP,Tools: QIIME, SILVANGS
Statistical ananlysis and visual of results: Tools:QIIME, SILVANGS
Shotgun metagenomics
Trimming low-quality regions: Tools: FASTQ/FASTX
De-replication: Duplicate read removal Tools:DUST
Screening:Contamination remval. Tools:Bowtie2
Preprocessing
Assembly: Tools: IDBA-UD,ALLPATH
Bining: Tools: ESOM,CLAMS,PhymmBL and metaCluster
Sequence reconstruction and grouping
Gene and regulatory element prediction
Ribosomal RNA tRNA prediction: Tools: tRNAscon,Database: SILVA,Greengenes,RDP
Gene Calling: using Tools: FragGeneScan
Optional: clustering Tools:UCLUST 90%
Optional: CRISPR elememt identification Tools:CRT and PILER-CR
Functional assignment to the predicted protein coding genes: Tools:IMGIMER,MGRAST, Database: KEGG,SEED,eggnog,COGIKOG,PFAM
Functional annotation
DataManagement Storage and sharing: Tools:IMGIMER,CAMERA,MG-RAST and EBI metagenomic SILVANGS

Fig. 7.4 Flow chart of the basic steps and tools of metagenomics analysis. The next generation sequencing can obtain marker gene metagenomics and shotgun metagenomics, and followed analysis based on these data can obtain useful information (Oulas et al. 2015)

sharp drop in the cost of HTS, it is easy to determine part and the entire genome of a pathogenic organism from pure cultures, host tissues, and soil environments. Whole-genome sequencing (WGS) can produce orders of magnitude more information about polymorphic sites, and a better understanding of their connection and occurrence in exons and introns (Tedersoo et al. 2019).

7.2.2.3 Technical Deviations of High-Throughput Sequencing Technology

All molecular-based analysis methods have specific analysis biases. Labeling bias may select organisms with high copy numbers. Primer bias will discriminate against targets whose primer-template does not match (Tedersoo et al. 2019). PCR bias is manifested as uneven amplification of target species due to differences in AT:CG ratio, DNA secondary structure, and marker length.

Label confusion and chimeric molecule formation are common technical products in HTS. Chimeric molecules are usually formed during PCR, and when the amplification is incomplete (due to low processability, short amplification time, or nucleotide depletion), these short fragments are initiated as templates in subsequent cycles. As the PCR cycle and the community complexity increase, the formation of chimeric molecules between two closely related organisms has become more common. In essence, tag-switch artifacts are also chimeric molecules that are formed between multi-component samples during the post-PCR library preparation step (Tedersoo et al. 2019).

7.2.3 Optimizing Microbiome Research Methods to Avoid Misunderstandings

Using high-throughput sequencing methods to study the microbiome, we can obtain the composition of the microbiome from different sources and compare them to reveal the relevant patterns of the microbiome. However, there are often many deviations during the experiment. Therefore, we must carefully design the experiment to ensure that it can answer the questions raised, and the statistical analysis methods involved in the research must be designed from the beginning, to realize a power analysis.

7.2.3.1 Influencing Factors

Antibiotic, diet, age, gender, longitudinal instability (time gradient), cage effect in animal experiments.

7.2.3.2 Precautions during Sample Collection and Processing

Sample Storage Conditions

The most important thing in the preservation of microbiome samples is to reduce the variation of the original microflora from sample collection to treatment and to maintain the same preservation conditions for all samples (Kim et al. 2017).

Set Negative Control

Negative control samples are collected to assess the background of contamination. We usually set three types of negative control samples in each 16S rRNA sequencing reaction. "blank swab", a sterile cotton swab was opened from the package and subjected to a complete sequencing protocol. "blank extraction", DNA extraction and all subsequent steps are performed without the addition of additional material. "blank library", the extraction scheme is not implemented. DNA-free water is used in the subsequent steps of the extraction of the protocol, starting with library construction, to characterize contamination in the downstream steps.

Set Positive Control

Positive control samples can verify that the sample preparation and a sequencing process is proceeding smoothly. When samples are purified on a porous plate, the samples are aligned at specific locations on the plate, which allows any sample mix-ups to be tracked and detected in the sequencing results. Ideally, positive and negative control samples would be placed asymmetrically on the extraction plate to determine the orientation of the plate.

7.3 Microbiome Big Data and Challenges

Microbiomics data integration and data analysis have uniqueness. A microbial community is usually composed of more than one species, thus making microbiome data more complicated to understand. In recent years, the analysis of microbiome data has become a research hotspot, and 16S sequencing is an important breakthrough in the field of microbial ecology.

7.3.1 *Main Methods of Microbiome Analysis*

In the past 10 years, with the development and application of high-throughput sequencing technology, the relevant analytical methods and tools in the field of microbiome research have also made rapid progress. A large number of excellent software, processes, and visualization tools have been released, further promoting the development of this field (Liu et al. 2019).

7.3.1.1 Amplicon Analysis Software

Amplicon analysis technology is widely used in microbiology and can quickly learn the microbial diversity in the research object. Mothur, QIIME, and USEARCH are three important amplicon analysis software, published and cited 10,000 times in the past 10 years. Figure 7.5 integrates these widely used works and methods in the field of microbiome research in recent years.

Mothur It integrates the previously published OTU definition software DOTUR, OTU difference comparison tool SONS, and other available tools, and realized the first set of relatively complete analysis procedures, making it possible for researchers to carry out amplicon analysis (Schloss et al. 2020).

QIIME Compared with mothur, it has more advantages, mainly including (1) integrating more than 200 pieces of related software and packages to achieve more choices for each step; (2) providing more than 150 scripts to achieve various personalized analyses; (3) the process is highly open and easy to integrate new software and methods; (4) enhance statistics and visualization, realize diversity, species composition, difference comparison, network, and many other methods and publication-level charting. Since QIIME allows researchers in the same field to carry out personalized analysis and visualization of amplicon data more

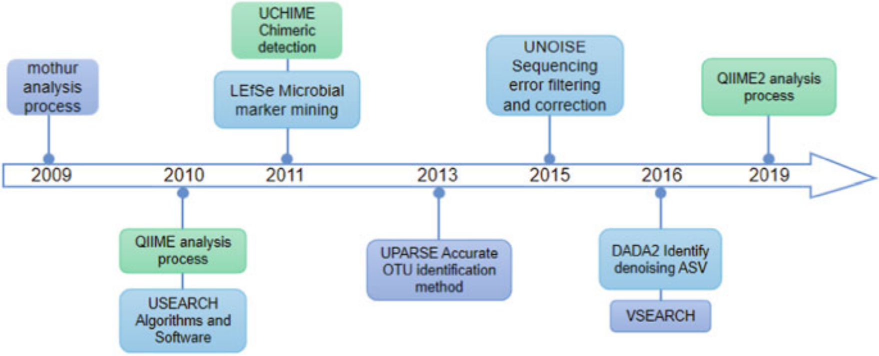
UCHIME Chimeric detection
mothur analysis process
LEfSe Microbial marker mining
UNOISE Sequencing error filtering and correction
QIIME2 analysis process
2009
2010
2011
2013
2015
2016
2019
QIIME analysis process
USEARCH Algorithms and Software
UPARSE Accurate OTU identification method
DADA2 Identify denoising ASV
VSEARCH

Fig. 7.5 Important software and algorithms used in microbiome research. These software and algorithms are organized according to their invention time along the timeline

autonomously, it has gradually become the most popular software in this field (Caporaso et al. 2010; Kuczynski et al. 2012).

USEARCH-Based Amplicon Analysis Process Based on previous algorithms and software, Robert gradually developed USEARCH into a complete amplicon analysis process including nearly 200 commands. It is also cross-platform, compact, free of dependencies, and easy to install.

7.3.1.2 Metagenomics Analysis Software

Compared with amplicon sequencing, metagenomics sequencing can not only obtain unbiased species composition, but also obtain the functional composition of the research object, and even splice the draft genome of some microorganisms. In areas with many studies such as the human gut microbiome, the quantitative analysis of metagenomics species and functional composition can be quickly realized through comparison based on reference databases, such as MetaPhlAn2 and Kraken2 for sequence species classification, and HUMAnN2 for functional composition quantification. For areas lacking high-quality metagenomics reference databases, it is necessary to splice metagenomics data from the beginning (de novo) and perform gene prediction. Commonly used metagenomics splicing software includes MEGA-HIT and metaSPAdes, and gene annotation software such as Prokka and GeneMarkS-2. For the combined analysis of multi-sample or multi-batch metagenomics data, CD-HIT is usually used to construct a non-redundant gene set. The obtained gene set is compared to a variety of protein function annotation databases, such as CAZy, CARD, and VFDB.

More, it is also possible to assemble a single bacterial genome through the binning method. At present, the commonly used binning tools include MetaBAT 2, MaxBin 2, and CONCOCT, but the results are quite different. Two binning purification tools metaWRAP and DAS_Tool were published last year, which solved the problem of difficult selection of binning tools and large differences in results. They usually integrate the results of 3 ~ 5 binning tools, further screening, and comprehensive utilization, to obtain higher quality a single-bacterial genome, at the same time provides a series of common analysis functions such as quantification and annotation of bins.

7.3.1.3 Statistics and Visualization Tools

Taxonomic tables and functional tables obtained by amplicon and metagenomics analysis are collectively referred to as feature tables, which are a common format in the analysis results of second-generation sequencing data. In downstream analysis, data can be converted and presented by selecting a variety of R packages, graphical interfaces, command line, or web version tools. The Bioconductor website provides thousands of R packages for biological data analysis. STAMP can realize principal component analysis and multiple statistical methods to compare two or more groups. LEfSe can implement a command-line tool based on linear discriminant analysis to find feature vectors.

7.3.2 The Basic Process of Microbiome Analysis

On May 23, 2018, Nature Microbiology Review (2017 IF: 31.851) published a review of research methods in the field of the microbiome, which not only systematically summarized the past but also provided a clear picture of the research methods in the field in the next 3–5 years. This review aims to provide direct guidance for the design and execution of microbiome experiments and the analysis of the resulting data, with particularly emphasis on the human, model, and environmental microbiome. Next, we will guide readers to a more professional commentary on the specific topics that exist (Knight et al. 2018).

7.3.2.1 Experimental Design

Designing experiments that produce meaningful data is the first step in the analysis. The general methods applicable to microbiome analysis are independent of the source of the sample. However, the specific details of the analysis may depend on the source of the sample. When evaluating different sample types, other main considerations are experimental design and sample collection. Careful experimental design is essential to obtain accurate and meaningful results from microbiome research.

Different methods of investigating microbial communities, including marker gene, metagenome, and metatranscriptome sequencing, may produce different results. We outline the best workflow for each method in Fig. 7.6.

After careful design and sample collection, microbiome data generation consisted primarily of 16S, metagenomic, or macrotranscriptome sequencing. The advantages and disadvantages of the three commonly used microbial community research methods are compared and analyzed in Table 7.2:

Marker gene sequencing (amplicons): The primers used in marker gene sequencing are usually designed for a specific field to determine the phylogenetic relationship of microorganism in the sample (Chen et al. 2011).

Metagenomic analysis: Metagenomic analysis is a method of sequencing all microbial genomes in a sample (Chen et al. 2011).

Macrotranscriptome analysis: Metatranscriptome analysis is to analyze the transcription process of the microbiome by using RNA sequencing to provide information about gene expression and microbiome functional activity.

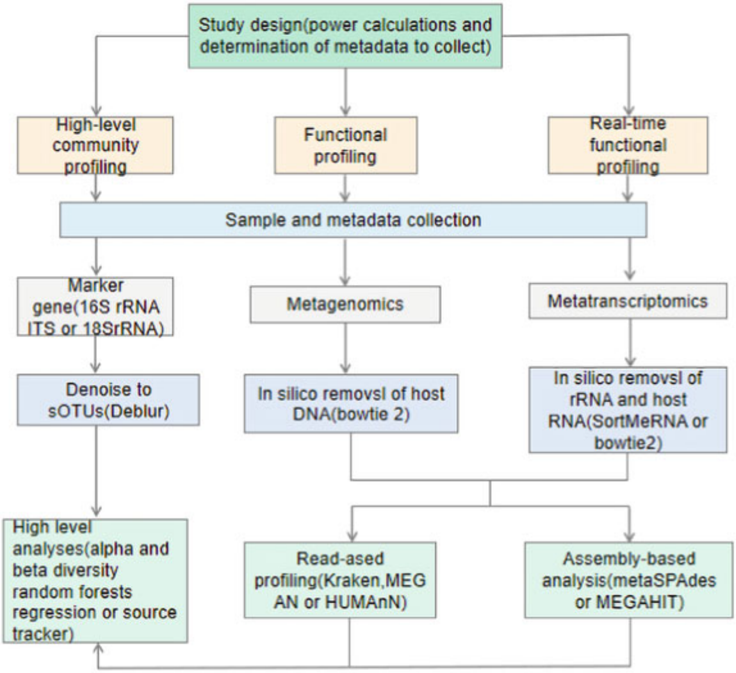
Study design(power calculations and determination of metadata to collect)
High-level community profiling
Functional profiling
Real-time functional profiling
Sample and metadata collection
Marker gene(16S rRNA ITS or 18SrRNA)
Metagenomics
Metatranscriptomics
Denoise to sOTUs(Deblur)
In silico removsl of host DNA(bowtie 2)
In silico removsl of rRNA and host RNA(SortMeRNA or bowtie2)
High level analyses(alpha and beta diversity random forests regression or source tracker)
Read-ased profiling(Kraken,MEGAN or HUMAnN)
Assembly-based analysis(metaSPAdes or MEGAHIT)

Fig. 7.6 The practical workflow for analyzing 16S, metagenomics, and metatranscription sequencing data. The analysis mainly divided into high-level community profiling and functional profiling and real-time functional profiling. And the analysis of different sequencing data in three modules have corresponding methods and tools

7.3.2.2 Higher Level Analysis

Processing the microbiome data generates a matrix that correlates feature richness (taxa or genes) with samples. The overall pattern of microbiome variation is usually assessed by alpha and beta diversity.

The software used to perform Alpha and Beta diversity calculations include QIIME, Mothur, and the R software package vegan. The non-parametric permutation tests PERMANOVA and ANOSIM are used to assess the significant β diversity clustering between groups, but PERMANOVA may perform better on data sets with changes in dispersion within groups. To visualize beta diversity data, sorting techniques such as principal coordinate analysis (PCoA) or principal component analysis (PCA) are usually used.

7.3.2.3 Integrating Other Omics Data

For a given study, the integration of other data types (including marker gene sequencing, metagenomics, metatranscriptomics, metaproteomics, metabolomics,

Table 7.2 Three approaches for microbial community research (Chen et al. 2011)

Method	Advantage	Disadvantage
Marker gene analysis	• Fast, simple and inexpensive sample preparation and analysis • Closely related to genome content • Suitable for samples with low biomass and high host contamination • Compare with existing large public data sets	• No discrimination in death or alive • Affected by amplification bias • Selection of primers and variable regions will amplify the deviation • Need prior knowledge of the microbial community • Resolution is usually only to genus • Need for proper negative control • Limited functional information
Metagenomic analysis	• The relative abundance of microbial functional genes can be directly inferred; for known organisms, microbial classification and phylogenetic identity can be achieved at the species and strain level • It is not assumed to understand the microbial community • No biases associated with PCR • The in-situ growth rate of target organisms with sequenced genomes can be estimated • It is possible to assemble a population-average microbial genome • Can be used for new gene families	• Relatively expensive, laborious and complicated sample preparation and analysis • Contamination of DNA and organelles from the host may obscure microbial characteristics • The default pipeline usually does not annotate viruses and plasmids well • No discrimination in death or alive • Due to assembly artifacts, population average microbial genomes are often inaccurate
Macrotranscriptome analysis	• When paired with a marker gene, it is possible to estimate which microorganisms in the community are actively transcribing • Inherently distinguish between active living organisms and dormant or dead microorganisms and extracellular DNA • Capture dynamic internal differences • Direct assessment of microbial activity, including response to interventions and event exposure	• The most expensive, laborious and complicated sample preparation and analysis • Host mRNA contamination and rRNA must be removed • Need for careful sample collection and storage • Data biased towards organisms with high transcription rates • Paired DNA sequencing is required to decouple the transcription rate from changes in bacterial abundance

and other technologies) is essential for a comprehensive understanding of the composition and function of microbial communities.

There are many difficulties in integrating multiple omics data. For example, gene expression and metabolism run on different time scales, and microorganisms usually produce many metabolites only in response to molecular signals from other species.

Simply finding correlations in various omics data is only the first step. Establishing causality and correlation across data sets is the next challenge. In multi-omics analysis, it is important to correct for multiple comparisons. Despite these challenges, the future potential of omics data integration is promising. Some examples have successfully integrated metabolome, metabolome, and metabolome data, clarified gene regulation in the microbiome and correlated the presence of microorganisms with metabolites.

7.3.3 The Basic Flow of Microbiome Data Analysis

The development of high-throughput sequencing technology has led to a series of microbiome research technologies, such as amplicon, metagenome, and macrotranscriptome, which have rapidly promoted the development of microbiome (Liu et al. 2019). The next-generation sequencing (NGS) technology makes it possible to study the microbial composition based on the non-culture method, which greatly promotes the study of the microbiome. Current studies on microbiome samples mainly focus on three levels:(1) microbe culture level, (2) DNA level, and (3) mRNA level.

Microbiome research can be divided into four stages (Fig. 7.7b): (1) Microbiome sample preparation: based on scientific experimental design, microbiome samples from people, animals and plants or the environment are collected, and DNA or RNA is selected to extract according to the purpose of the research. (2) Meta-omics data production: Meta-omics data were obtained by constructing a sequencing library and high-throughput sequencing after DNA or RNA extraction of samples. (3) Data processing (quality control quantitative): obtaining microbiome data, quality control should be carried out first, including removing the primers and connectors added artificially in the process of sequencing and database building, as well as the low-quality sequences generated in the process of sequencing. In addition, the sequencing results of the host-associated microbiome contain a large number of host sequences, which need to be removed by comparing the host genome. The

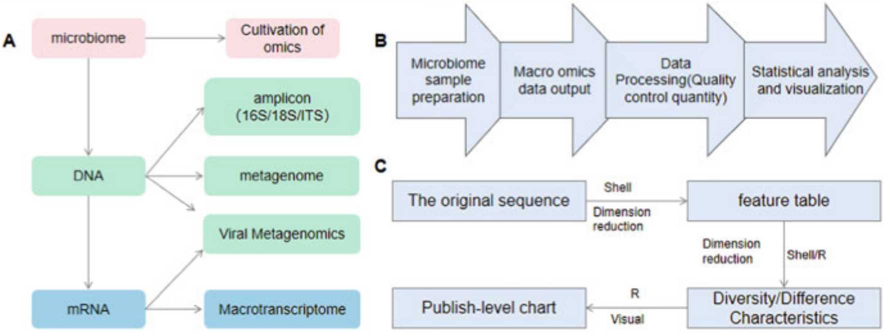

Fig. 7.7 Overview of microbiome research procedure (Liu et al. 2019). (**a**) The main three levels on microbiome samples. (**b**) The four stages of microbiome research. (**c**) The main three steps of microbiome data analysis

obtained clean data were then compared to the reference database or the reference gene set assembled from De novo, and the quantitative value was the feature table. According to the sequence annotation type, the feature table could be divided into species or functional gene composition tables. (4) Statistical analysis and visualization: the feature table should be further combined with the sample metadata for statistical analysis, and appropriate graphs should be selected for visualization, which is conducive to the observation and summary of biological laws, and improve the readability and dissemination of results (Fig. 7.7b).

After obtaining the original microbiome data, how to analyze it to a highly readable publication-level chart? To facilitate understanding, this article divides the microbiome data analysis process into three main steps (Fig. 7.7c).

7.4 Representative Microbiome Databases and Analysis Tools

The development of the next-generation sequencing (NGS) technology makes it possible to study the microbial composition based on non-culture methods and promotes the study of the microbiome (Liu et al. 2019).

The accumulation and integration of microbiome data, the development of microbiome has greatly promoted the study of microbial communities. In the past decade, a large number of microbiology-related literature has been published. In terms of microbiome data analysis and mining, with the accumulation of microbiome data, a large number of microbiome data analysis methods, platforms and database resources have emerged. Table 7.3 is the current popular common algorithm and analysis platform.

High-quality reference databases are the foundation for efficient analysis of microbiome data, and progress in this field depends on the application of large-scale culturomics and the publication of more high-quality reference genomes. At the same time, it is also necessary to classify the published data, improve its usability and further mining.

Metagenomics is a method of studying the microbiome, which involves sequencing the genetic material of all microbes in a given environment, such as soil, water, or the human gut. One of the most important databases for microbiome researches is EBI MGnify. EBI Metagenomics is one of EMBL-EBI's fastest-growing data resources, has changed its name to MGnify, provides a freely available automated process for assembly, analysis, and archiving of microbiome data.

With the development of the microbiome, a large numbers of microbiome analysis tools have emerged, greatly improving the efficiency of analysis. Table 7.4 illustrates the widely used analysis tools.

In 2017, in a review titled "A review of methods and databases for metagenomic classification and assembly" published in Briefings in Bioinformatics, there are also many ideas and software summaries that can be referred to (Breitwieser et al. 2019).

Table 7.3 Representative microbiome analysis platform and database as online resources

Software (platform)	Analysis data object	Analysis result	Reference
MOCAT	Metagenome	Species structure, abundance and function classification, and comparison between species	Kultima et al. (2016)
MEGAN	16S rRNA	Species structure, abundance and function classification, and comparison between species	Huson et al. (2007)
MetaPhlAn	Metagenome	Species structure, abundance	Segata et al. (2012)
PICRUSt	Metagenome, 16S rRNA	Species structure and function classification	Douglas et al. (2020); Langille et al. (2013)
antiSMASH	Metagenome	BGC analysis	Blin et al. (2019); Medema et al. (2011)
Sort-ITEMS	16S rRNA	Species structure and function classification	Monzoorul Haque et al. (2009)
UniFrac	16S rRNA	Species structure, abundance and function classification, and comparison between species	Lozupone and Knight (2005)
QIIME	16S rRNA	Species structure, abundance and function classification,	Caporaso et al. (2010); Kuczynski et al. (2012)
MG-RAST	Metagenome, 16S rRNA	Species structure, abundance and function classification, and comparison between species	Keegan et al. (2016)
IBDsite	Metagenome, 16S rRNA	Species structure, abundance and function classification, and comparison between species	Merelli et al. (2012)

1. Figure 7.8 showed the classic process of metagenomics.
2. Table 7.5 represented the commonly used quality control tools.
3. Table 7.6 represented the commonly used Classification annotation tools.
4. Table 7.7 represented the assembly and binning tools are summarized.

7.5 Representative Microbiome Analysis Applications

Based on currently available microbiome analysis databases and tools, many prominent biological problems have largely been solved, or at least have gained deeper insights. Typical examples of microbiome analysis are outlined below.

Table 7.4 Representative microbiome analysis tools

Tool name	Type	Web link	Reference
Uparse	OTU clustering tool	https://drive5.com/uparse	Edgar (2013)
Usearch	Integrated sequence analysis tool for amplicon	https://www.drive5.com/usearch/	Rognes et al. (2016)
Vsearch	Alternative implementation of Vsearch	https://github.com/torognes/vsearch	Rognes et al. (2016)
DADA2	Amplicon sequence variant (ASVs)tools	https://benjjnebgithub.io/dada2/	Callahan et al. (2016)
Deblur	Amplicon sequence variant (ASVs)tools	https://github.com/biocore/deblur	Schuler et al. (2015)
PICRUSt/ PICRUSt2	Functional profiles prediction from amplified marker genes	http://picrust.github.io/picrust/	Douglas et al. (2018); Douglas et al. (2020)
Tax4Fun	Functional profiles prediction from amplified marker genes	http://tax4fun.gobicsde/	Aßhauer et al. (2015)
QIIME/ QIIME2	Integrated microbiome bioinformatics workflow	http://qiime.org/ https://qiime2.org/	Caporaso et al. (2010); Kuczynski et al. (2012)
Mothur	Integrated microbiome bioinformatics workflow	https://mothur.org/	Schloss (2020)
Kraken	Taxonomical annotation of WGS short reads	http://ccb.jhu.edu/software/kraken/	Wood et al. (2014)
MetaphlAn2	Taxonomical annotation of WGS short reads	https://huttenhower.sph.harvard.edu/	Truong et al. (2015)
HUMANn2	Funcctional annotation of WGS short reads	https://huttenhower.sph.harvard.edu/human	Franzosa et al. (2018)
metaSPAdes	Assembling of WGS short reads	https://github.com/ablab/spades	Nurk et al. (2017)

7.5.1 Biogeographical Characteristics of the Intestinal Flora

Spatial structure is essential for natural ecosystems and many microbial communities (including gut flora), exhibit complex spatial organization. Biogeographic maps of bacteria can reveal the interaction of community functions, but existing methods cannot accommodate the hundreds of species found in the natural microbiome.

In July 2019, "Spatial metagenomic characterization of microbial biogeography in the gut" published by Nature Biotechnology, proposes a new sequencing method, MAPS-Seq, which is a multiplex sequencing technology that analyzes microbial cells in their natural geographic environment to statistically reconstruct the local spatial organization of the microbiome.

MAPS-Seq is a physical fixation of the input sample followed by in situ fixation of the microbiota by perfusion and polymerization on an acrylamide polymer matrix. The embedded sample is then broken by cold bead pulping, cell lysis is performed, and size selection is performed through a nylon screen to produce cell clusters or

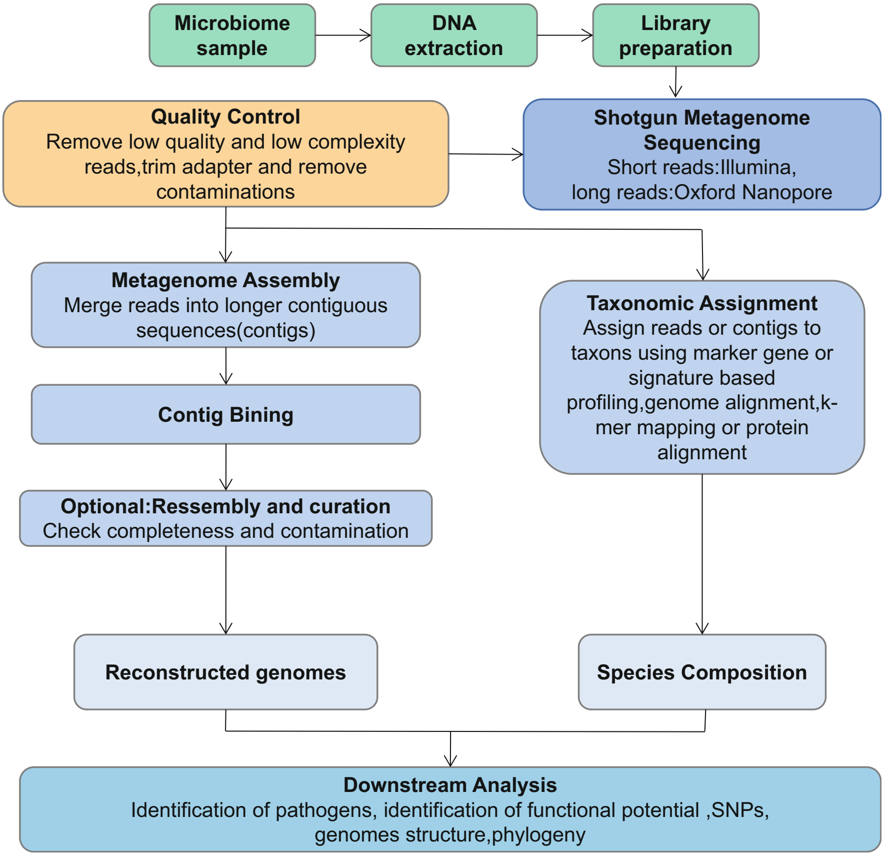
Microbiome sample
DNA extraction
Library preparation
Quality Control
Remove low quality and low complexity reads,trim adapter and remove contaminations
Shotgun Metagenome Sequencing
Short reads:Illumina,
long reads:Oxford Nanopore
Metagenome Assembly
Merge reads into longer contiguous sequences(contigs)
Taxonomic Assignment
Assign reads or contigs to taxons using marker gene or signature based profiling,genome alignment,k-mer mapping or protein alignment
Contig Bining
Optional:Ressembly and curation
Check completeness and contamination
Reconstructed genomes
Species Composition
Downstream Analysis
Identification of pathogens, identification of functional potential ,SNPs, genomes structure,phylogeny

Fig. 7.8 Overview of the metagenomic analysis process. Microbial sample acquisition, Total DNA extraction, library construction, on-machine sequencing, data quality control (removal of low quality and adapters, etc., removal of interference information such as host genome), metagenomic assembly, Contig Binning, genome reconstruction, taxonomic annotation (based on reads, contig, bins, the restored genome for species annotation), and other downstream analysis

particles with the desired physical size and adjustable physical size. The resulting cluster contains genomic DNA fixed in the original arrangement, and local spatial information is preserved. These clusters were then co-encapsulated with gel beads using a microfluidic apparatus, each containing a unique barcode positive 16S rRNA amplification primer. The primers were photolyzed from the beads and clusters, and the genomic DNA was released from the clusters by triggering the degradation of the polymer matrix within the droplets, and PCR amplification in the 16S V4 region was performed. The droplets were then broken up and the resulting library was deeply sequenced. Sequenced reads are filtered and grouped according to their unique barcodes, resulting in bacterial operational taxonomic units (OTUs) identity and relative abundance (RA) in a single cell cluster of defined size.

Table 7.5 Representative quality control tools (Luo et al. 2010)

Tool	Introduction	Web link	Reference
FastQC	Quality control tool showing statics such as quality value, sequence length distribution and GC content distribution	http://www.bioinformatics.babraham.ac.uk/projects/fastqc/	de Sena Brandine (2019)
FastQ Screen	Screen a library against sequence databases to see if composition of library matches expectations	http://www.bioinformatics.babraham.ac.uk/projects/fastq_screen	Wingett and Andrews (2018)
BBtools	BBDuk trims and filters reads using k-mers and entropy information, BBNorm normalizes coverage by down sampling reads	http://jgi.doe.gov/data-and-tools/bbtools/bb-tools-user-guide/	Institute., D.J.G (2021)
Trimmomatic	Flexible read trimming tool for Illumina data	http://www.usadellab.org/cms/?page=trimmomatic	Bolger et al. (2014)
Cutadapt	Find and remove adapter sequences, primers, poly-A tails and other types of unwanted sequence	https://cutadapt.readthedocs.io	Martin (2011)
Khmer/ diginorm	Tools for k-mer error trimming of reads and digital normalization of samples	http://khmer.readthedocs.io	Crusoe et al. (2015)
MultiQC	Summarize results from different analysis (such as FastQC) into one report	https://multiqc.info/	Sangwan et al. (2016)

The isolated clusters of homogeneous mouse fecal bacteria or Escherichia coli were analyzed by MAP-seq. The results showed that most of the detected barcodes were uniquely mapped to their respective initial communities with minimal mixing and that the introduction of contaminants during sample processing was negligible, meaning that MAP-Seq could accurately measure bacterial properties and abundances within a single spatially limited cluster of cells. MAP-Seq was applied to the colon microbiome of mice. The authors generated and characterized clusters of cells from the distal colon of mice fed a plant-polysaccharide diet, including epithelial and digestive tissue, producing 1406 clusters. A total of 236 OTUs were identified, and their prevalence in the entire cluster was highly correlated with the high abundance obtained by standard 16S sequencing, meaning that richer taxa were also physically dispersed over more space. The spatial distribution of taxa across groups appears to be mixed (median of 9 OTUs per group), but some clusters contain only a few OTUs, indicating spatial aggregation or clumping in parts of the community. In addition, the distribution of OTU in each cluster was observed to be significantly lower than clusters of the same size produced by homogeneous fecal bacteria, which was a control for a well-mixed community. These results suggest that the various taxa in the gut microbiome are neither completely mixed nor highly structured on the scale of tens of microns, but are heterographically distributed in mixed plaques.

Table 7.6 Representative Classification annotation tools

Tool	Introduction	Web link	Reference
Kraken	Fast taxonomic classifier using in-memory k-mer search of metagenomics reads against a database built from multiple genomes	https://ccb.jhu.edu/software/kraken/	Wood et al. (2014)
Kraken-HLL	Extension of kraken counting unique k-mers for taxa and allowing multiple database	https://github.com/fbreitwieser/kraken-hllCLARK	
CLARK(-S)	Fast taxonomic classifier using in-memory k-mer search of metagenomics reads against a database built from completed genomes. Extension uses spaced k-mer seeds for better classification.	http://clark.cs.ucr.edu	Ounit and Lonardi (2016)
Kallisto	Taxonomic profiler using pseudo-alignment with k-mers using techniques based on transcript quantification.	https://github.com/pachterlab/kallisto	Bray et al. (2016)
DIAMOND	Protein homology search using spaced seeds with a reduced amino acid alphabet,2000–20,000 times faster than BLASTX	https://github.com/bbuchfink/diamond	Buchfink et al. (2015)
BLAST+	Highly sensitive nucleotide translated-nucleotide protein alignment	https://blast.ncbi.nlm.nih.gov	Camacho et al. (2009)
MetaPhlAn2	Marker gene-based taxonomic profiler	https://bitbucket.org/biobakery/metaphlan2	Truong et al. (2015)
mOTU	Taxonomic profiler based on a set of 40 prokaryotic marker genes	http://www.bork.embl.de/software/mOTU/	Sunagawa et al. (2013)

The research found that diet plays an important role in the variation of intestinal flora between individuals. The authors divided co-reared mice into two cohorts, one on a plant-based polysaccharide-based diet (LF, the same as the previous cohorts) and the other on a high-fat, high-sugar diet (HF, commonly used in diet-induced obesity studies), to investigate changes in the microbiome representative of the diet-related changes. After 10 days, a significant decrease in species richness in the cecum and colon was observed in HF-fed mice compared to LF-fed mice.

To determine whether dietary changes could alter the spatial organization of the microbiota in a way that would contribute to the observed changes in species diversity, the authors performed MAP-seq on distal colon samples from mice fed LF or HF diets. The authors found that the distribution of unique OTU per 20 μm cluster was similar between the two diets. This implies that the distribution of species in the local ~20 μm range is controlled by factors that are common or not influenced by both diets. However, assessment of diversity at higher taxonomic levels showed significantly higher diversity in HF clusters, suggesting that while both LF and HF clusters contain a similar number of OTUs, taxa within a single HF cluster are more

Table 7.7 Representative the assembly and binning tools

Tool	Synopsis	Web link	Reference
Megahit	Co-assembly of metagenomic reads with variable k-mer lengths and low memory usage	https://github.com/voutcn/megahit	Gleeson et al. (2011); Wosinska et al. (2019)
SPAdes	Dbg assembler using multiple k-mers, works also for simple metagenomes	http://cab.spbu.ru/software/spades	Bankevich et al. (2012)
MetaSPAdes	Extension of SPADES with better assemblies with different abundances, conserved regions and strain mixtures	http://cab.spbu.ru/software/spades/	Nurk et al. (2017)
Ray Meta	DBG assembler with fixed k-mer size	http://denovoassemler.sourceforge.net/	Boisvert et al. (2012)
IDBA-UD	DBG assembler using multiple k-mer sizes, analyzes coverages between paths to give better assemblies in complex metagenomes with uneven coverage	http://i.cs.hku.hk/~alse/hkubrg/projects/idba_ud/	Peng et al. (2012)
MetAMOS	Framework for metagenomic assembly, analysis and validation	http://metamos.reasthedocs.io	Treangen et al. (2013)
MOCAT2	Pipeline for read filtering taxonomic profiling assembly, gene prediction and functional analysis	http://mocat.embl.de/	Kultima et al. (2016)
CONCOCT	Bins contigs using nucleotide composition, coverage data in multiple samples and paired-end read information	https://github.com/BinPro/CONCOCT	Alneberg et al. (2014)
COCACOLA	Binning contigs in using read coverage, correlation, sequence composition and paired-end read linkage	https://github.com/younglululu/COCACOLA	Lu et al. (2017)

phylogenetically diverse. This further emphasizes the usefulness of map-seq for examining spatial organization at a higher classification resolution.

7.5.2 *Plasticity of Intestinal Flora (Dynamic Pattern)*

“Unaccustomed to water and soil” is a very common symptom that people produce during travel. This is actually a reaction caused by changes in the intestinal flora caused by dietary changes. How does diet affect gut microbes? Scientists have conducted many studies on this question. They selected samples of patients and healthy people and adjusted their diets over a period of time. The results shown that short-term diet changes can change the composition of the intestinal flora of patients and healthy people. However, the conclusions of these studies are only a few points in time, and the respondents have only experienced one dietary change. The dynamic pattern of the microbial community on a longer time scale and various dietary changes remain unclear.

In 2018, the Ningkang team published an article "Resilience of human gut microbial communities for the long stay with multiple dietary shifts" on Gut, observing the changes of human intestinal flora under long-term dietary changes, and clarifying the principle of plasticity of intestinal flora. The authors recruited a team of ten Chinese volunteers. They set off from Beijing, stayed in Trinidad and Tobago for 6 months, and then returned to Beijing. By using a high-density longitudinal sampling strategy, fecal samples of volunteers were collected and their detailed dietary information was recorded. High-throughput sequencing and correlation analysis of stool samples revealed that the human intestinal flora changes dynamically due to dietary changes over a long period of time. The research found that the microbial community in the intestine has two-way plasticity and elasticity during the long-term stay, and has a variety of dietary changes.

7.5.3 Gene Mining

Our body host a vast array of microorganisms that encode about 100 times more unique genes than our own genome. These microorganisms have a great impact on human healthy, especially those living in the intestines. The intestinal microbes are essential for digesting food. It will be a high chance of suffering intestinal disease or obesity once the intestinal microbiota is changed. Therefore, it is necessary to decipher the content, diversity and function of the gut microbial community based on the gene diversity since genes are the basic units that control functions and traits.

Jun Wang et al. established a human gut microbial gene catalog through metagenomic sequencing. They collected 124 fecal samples of European individuals and recovered 3.3 million non-redundant microbial genes from the Illumina-based metagenomic sequencing data of these samples (Qin et al. 2010). The number of non-redundant microbial genes is 150 times the number of human gene complement. These microbial genes were largely shared among individuals in this cohort, and they covered the vast majorities of (more common) microbial genes in the cohort, contained the majorities of human gut microbial genes. Interestingly, more than 99 percent of these genes were bacterial, the entire cohort consisted of 1000 to 1150 endemic bacterial species, with individual containing at least 160 such species, and that they were also largely shared. The minimum intestinal metagenome and the minimum intestinal bacterial genome in terms of the functions of all individuals and most bacteria were defined and described by the authors, respectively.

The research found that most of the microorganisms living in the gut have a great impact on human physiology and nutrition and are crucial to human life. Researchers aimed to understand and exploit the impact of the gut microbiome on human health in terms of its content, diversity and function based on the 16S ribosomal RNA gene (rRNA) sequencing. They found that two bacterial families, the Bacteroidaceae and the Antimicrobiaceae, account for more than 90% of known phylogenetic categories and dominate the distal intestinal flora, and they also found large differences in the gut microbiome between healthy individuals.

To generate a broad catalog of microbial genes from the sequencing data of human gut, the authors first assembled short Illumina reads into longer overlapping clusters (contigs), which could then be analyzed and annotated using standard methods. They used SoapDeNovo to assemble all Illumina GA sequence data from scratch, and they assembled a total of 6.58 million overlapping groups from up to 42.7% of Illumina GA reads. They were surprised to find that nearly 35% of reads from any one sample could be mapped to overlap groups from the other samples, indicating a common sequence core. The authors then combined the unassembled reads from all 124 samples and repeated the de novo assembly process to complete the overlapping group setup, and they obtained bout 400,000 overlapping groups with a length of 370 Mb and a length of N50 939 bp. Therefore, the authors assembled total of 10.7 GB overlapping groups.

To establish a non-redundant human gut microbiome genome, the authors first used the Metagene program to predict ORFs in contigs, and they found 14,048,045 ORFs over 100 bp in length which accounted for 86.7% of the contigs, comparable to the 86 percent found in fully sequenced genomes. The authors found that about wo-thirds of the ORFs appeared to be incomplete, and they attributed it to the size of the assembled overlapping groups (N50 is 2.2 KB). In order to avoid dataset bloat due to possible sequencing errors, the authors used very strict criteria (95% conformance exceeds 90% of the shorter ORF length) to remove the excess ORFs.

The authors defined the “epidemic genes” as the genes in the non-redundant set because they are encoded on an overlapping group assembled from the richest read segments. The authors examined the number of prevalent genes found across all individuals, which is a function of the sequencing range, and they set the minimum number of genes to support reads to two. Through estimating the coverage richness (ICE) based on incidence determined by 100 people (the maximum number that can be accommodated by the Evaluations21 program), the authors concluded that they captured 85.3% of the prevalence genes, suggesting that the catalog contains the vast majority of the prevalent genes in this cohort, even through this may be an underestimate.

Furthermore, there are already studies on associating microbiome and other omics for the in-depth understanding of the regulation principles and dynamic patterns. These multi-omics studies include: Multi-omics analysis reveals the changes of metabolome, gut microbiome and immune indicators with age (Zhang et al. 2021). Human gut microbiota from Autism Spectrum Disorder promote(ASD) behavioral symptoms in mice (Sharon et al. 2019). Gut microbiome structure and metabolic activity in inflammatory bowel disease (Franzosa et al. 2019).

7.6 Conclusion and Perspectives

In summary, the microbiome has become one of the key omics that could assist the in-depth understanding of the biological system, and the microbiome data analysis has already helped for the interpretation of patterns and dynamics of the microbial communities at every niche of our body.

References

Alneberg J, et al. Binning metagenomic contigs by coverage and composition. Nat Methods. 2014;11(11):1144–6.

Aßhauer KP, et al. Tax4Fun: predicting functional profiles from metagenomic 16S rRNA data. Bioinformatics. 2015;31(17):2882–4.

Bankevich A, et al. SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. J Comput Biol. 2012;19(5):455–77.

de Sena Brandine G, Smith AD. Falco: high-speed FastQC emulation for quality control of sequencing data. F1000Res. 2019b;8:1874.

Blin K, et al. antiSMASH 5.0: updates to the secondary metabolite genome mining pipeline. Nucleic Acids Res. 2019;47(W1):W81–7.

Boisvert S, et al. Ray Meta: scalable de novo metagenome assembly and profiling. Genome Biol. 2012;13(12):R122.

Bolger AM, Lohse M, Usadel B. Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics. 2014;30(15):2114–20.

Bray NL, et al. Near-optimal probabilistic RNA-seq quantification. Nat Biotechnol. 2016;34(5): 525–7.

Breitwieser FP, Lu J, Salzberg SL. A review of methods and databases for metagenomic classification and assembly. Brief Bioinform. 2019;20(4):1125–36.

Buchfink B, Xie C, Huson DH. Fast and sensitive protein alignment using DIAMOND. Nat Methods. 2015;12(1):59–60.

Byrd AL, Belkaid Y, Segre JA. The human skin microbiome. Nat Rev Microbiol. 2018;16(3): 143–55.

Callahan BJ, et al. DADA2: high-resolution sample inference from Illumina amplicon data. Nat Methods. 2016;13(7):581–3.

Camacho C, et al. BLAST+: architecture and applications. BMC Bioinformatics. 2009;10(1):421.

Caporaso JG, et al. QIIME allows analysis of high-throughput community sequencing data. Nat Methods. 2010;7(5):335–6.

Chen C, et al. Removing batch effects in analysis of expression microarray data: an evaluation of six batch adjustment methods. PLoS One. 2011;6(2):–e17238.

Chiu CY, Miller SA. Clinical metagenomics. Nat Rev Genet. 2019;20(6):341–55.

Crusoe MR, et al. The khmer software package: enabling efficient nucleotide sequence analysis. F1000Res. 2015;4:900.

Di Bella JM, et al. High throughput sequencing methods and analysis for microbiome research. J Microbiol Methods. 2013;95(3):401–14.

Douglas GM, Beiko RG, Langille MGI. Predicting the functional potential of the microbiome from marker genes using PICRUSt. In: Beiko RG, Hsiao W, Parkinson J, editors. Microbiome analysis: methods and protocols. New York, NY: Springer New York; 2018. p. 169–77.

Douglas GM, et al. PICRUSt2 for prediction of metagenome functions. Nat Biotechnol. 2020;38 (6):685–8.

Edgar RC. UPARSE: highly accurate OTU sequences from microbial amplicon reads. Nat Methods. 2013;10(10):996–8.

Franzosa EA, et al. Gut microbiome structure and metabolic activity in inflammatory bowel disease. Nat Microbiol. 2019;4(2):293–305.

Franzosa EA, et al. Species-level functional profiling of metagenomes and metatranscriptomes. Nat Methods. 2018;15(11):962–8.

Gleeson M, et al. The anti-inflammatory effects of exercise: mechanisms and implications for the prevention and treatment of disease. Nat Rev Immunol. 2011;11(9):607–15.

Huson DH, et al. MEGAN analysis of metagenomic data. Genome Res. 2007;17(3):377–86.

Institute., D.J.G. *BBDuk guide, 2021.*

Keegan KP, Glass EM, Meyer F. MG-RAST, a metagenomics service for analysis of microbial community structure and function. Methods Mol Biol. 2016;1399:207–33.

Kim D, et al. Optimizing methods and dodging pitfalls in microbiome research. Microbiome. 2017;5(1):52.

Kishikawa T, et al. Metagenome-wide association study of gut microbiome revealed novel aetiology of rheumatoid arthritis in the Japanese population. Ann Rheum Dis. 2020;79(1): 103–11.

Knight R, et al. Best practices for analysing microbiomes. Nat Rev Microbiol. 2018;16(7):410–22.

Kuczynski J, et al. Using QIIME to analyze 16S rRNA gene sequences from microbial communities. Curr Protoc Microbiol. 2012;27(1):1E.5.1–1E.5.20.

Kultima JR, et al. MOCAT: a metagenomics assembly and gene prediction toolkit. PLoS One. 2012;7(10):e47656.

Kultima JR, et al. MOCAT2: a metagenomic assembly, annotation and profiling framework. Bioinformatics. 2016;32(16):2520–3.

Langille MGI, et al. Predictive functional profiling of microbial communities using 16S rRNA marker gene sequences. Nat Biotechnol. 2013;31(9):814–21.

Liu Y, et al. Methods and applications for microbiome data analysis. Yi chuan = Hereditas. 2019;41 (9):845–62.

Lozupone C, Knight R. UniFrac: a new phylogenetic method for comparing microbial communities. Appl Environ Microbiol. 2005;71(12):8228–35.

Lu YY, et al. COCACOLA: binning metagenomic contigs using sequence COmposition, read CoverAge, CO-alignment and paired-end read LinkAge. Bioinformatics (Oxford, England). 2017;33(6):791–8.

Luo J, et al. A comparison of batch effect removal methods for enhancement of prediction performance using MAQC-II microarray gene expression data. Pharmacogenomics J. 2010;10 (4):278–91.

Martin M. Cutadapt removes adapter sequences from high-throughput sequencing reads. EMBnetjournal. 2011.

Medema MH, et al. antiSMASH: rapid identification, annotation and analysis of secondary metabolite biosynthesis gene clusters in bacterial and fungal genome sequences. Nucleic Acids Res. 2011;39(suppl_2):W339–46.

Merelli I, Viti F, Milanesi L. IBDsite: a Galaxy-interacting, integrative database for supporting inflammatory bowel disease high throughput data analysis. BMC Bioinformatics. 2012;13(14): S5.

Monzoorul Haque M, et al. SOrt-ITEMS: sequence orthology based approach for improved taxonomic estimation of metagenomic sequences. Bioinformatics (Oxford, England). 2009;25 (14):1722–30.

Nowrotek M, et al. Culturomics and metagenomics: in understanding of environmental resistome. Front Environ Sci Eng. 2019;13(3):40.

Nurk S, et al. metaSPAdes: a new versatile metagenomic assembler. Genome Res. 2017;27(5): 824–34.

Oulas A, et al. Metagenomics: tools and insights for analyzing next-generation sequencing data derived from biodiversity studies. Bioinf Biol Insights. 2015;9:BBI.S12462.

Ounit R, Lonardi S. Higher classification sensitivity of short metagenomic reads with CLARK-S. Bioinformatics. 2016;32(24):3823–5.

Papageorgiou L, et al. Genomic big data hitting the storage bottleneck. EMBnetjournal. 2018;24: e910.

Peng Y, et al. IDBA-UD: a de novo assembler for single-cell and metagenomic sequencing data with highly uneven depth. Bioinformatics. 2012;28(11):1420–8.

Qin J, et al. A human gut microbial gene catalogue established by metagenomic sequencing. Nature. 2010;464(7285):59–65.

Rognes T, et al. VSEARCH: a versatile open source tool for metagenomics. PeerJ. 2016;4:e2584.

Ruairi Robertson P. *16S rRNA Gene Sequencing vs. Shotgun Metagenomic Sequencing* https://blog.microbiomeinsights.com/16s-rrna-sequencing-vs-shotgun-metagenomic-sequencing. 2020.7.20.

Sangwan N, Xia F, Gilbert JA. Recovering complete and draft population genomes from metagenome datasets. Microbiome. 2016;4(1):8.

Schloss PD. Reintroducing mothur: 10 Years Later. Appl Environ Microbiol. 2020;86(2).

Schuler CJ, Fau-Hirsch M, et al. *Learning to Deblur, 2015.*

Segata N, et al. Metagenomic microbial community profiling using unique clade-specific marker genes. Nat Methods. 2012;9(8):811–4.

Sharon G, et al. Human gut microbiota from autism spectrum disorder promote behavioral symptoms in mice. Cell. 2019;177(6):1600–18.e17

Sunagawa S, et al. Metagenomic species profiling using universal phylogenetic marker genes. Nat Methods. 2013;10(12):1196–9.

Tedersoo L, et al. High-throughput identification and diagnostics of pathogens and pests: overview and practical recommendations. Mol Ecol Resour. 2019;19(1):47–76.

Treangen TJ, et al. MetAMOS: a modular and open source metagenomic assembly and analysis pipeline. Genome Biol. 2013;14(1):R2.

Truong DT, et al. MetaPhlAn2 for enhanced metagenomic taxonomic profiling. Nat Methods. 2015;12(10):902–3.

Vakhlu J, Sudan AK, Johri BN. Metagenomics: future of microbial gene mining. Indian J Microbiol. 2008;48(2):202–15.

Wang Z, et al. Time-course relationship between environmental factors and microbial diversity in tobacco soil. Sci Rep. 2019;9(1):19969.

Wingett SW, Andrews S. FastQ Screen: a tool for multi-genome mapping and quality control. F1000Res. 2018;7:1338.

Wood DE, Salzberg SL. Kraken: ultrafast metagenomic sequence classification using exact alignments. Genome Biol. 2014;15(3):R46.

Wosinska L, et al. The potential impact of probiotics on the gut microbiome of athletes. Nutrients. 2019;11(10):2270.

Zhang X, et al. Age-related compositional changes and correlations of gut microbiome, serum metabolome, and immune factor in rats. GeroScience. 2021;43(2):709–25.

Chapter 8
Current Progress of Bioinformatics for Human Health

Jin Zhao, Shu Zhang, Shunyao Wu, Wenke Zhang, and Xiaoquan Su

8.1 Introduction

Massive biological data provides a broad view for understanding the dynamics of human health status and disease from multiple aspects. During the past decade, the tremendous volume of biological data has been produced in different ways. How to analyze the high-volume data precisely and efficiently, and take advantage from it has become one of the most essential bottlenecks for precision medicine. Newly developed bioinformatics tools are bringing opportunities for these challenges, from sequence-based algorithms such as genome assembly and genome comparison, to disease classifiers like regular machine learning and neural network. In this chapter, we summarize the widely-used state-of-the-art computational approaches of multi-omics data to study human health and diseases, including bioinformatics methods and tools for genomics, transcriptomics, metagenomics, and single-cell data, as well as machine learning algorithms and strategies.

8.2 Genome Comparison and Analysis Expands our Understanding of Genetic Diseases and Treatments

Sequencing technology enables the survey of genetic information on a large scale. Using high throughput sequencing and bioinformatical analysis, studies like the human genome project (HGP) (Randal 1991) have highly promoted our understanding of human health and diseases (Hood and Rowen 2013; Lander 2011). Genome comparison and analysis are the basis for elucidating the evolutionary relationships

J. Zhao · S. Zhang · S. Wu · W. Zhang · X. Su (✉)
College of Computer Science and Technology, Qingdao University, Qingdao, China
e-mail: suxq@qdu.edu.cn

K. Ning (ed.), *Methodologies of Multi-Omics Data Integration and Data Mining*,
Translational Bioinformatics 19, https://doi.org/10.1007/978-981-19-8210-1_8

of species, as well as revealing gene functions and molecular mechanisms. Previous research revealed the occurrence between genetic diseases and changes in human genome structure that known as genome variants (Brandler et al. 2016; Sudmant et al. 2015; Feuk et al. 2006), which can be detected by genome comparison. Specifically, according to the distinct areas between two genomes, it would be possible to determine the pathogenic genes, which brings new approaches to predict the possibility of certain diseases, as well as suggest the best treatment strategy (Sankoff et al. 1992; Rasko et al. 2011; Ozery-Flato and Shamir 2009; Chen et al. 2008; Nalbantoglu et al. 2010).

Genome Assembly A prerequisite for genome variants detection is the reconstruction of complete genomes. The length of genomes varies between species: human genome is about 2.9G base pairs (bp), while that of bacteria is only 3–10 M bp. However, next generation sequencing (NGS; e.g. Illumina platforms) cannot directly produce DNA reads with adequate length to cover complete genomes. Thus, the complete genetic information restored from short sequence reads is always relied on sequence assembly. Currently, available sequence assembly algorithms fall into two categories: de novo sequence assembly and reference-based assembly (Pop 2009). The reference-based assembly approaches like E-RGA (Vezzi et al. 2011) usually employ existing genomes as templates, and then locate and sort short reads by sequence alignment. However, it can also miss short reads that exhibit significant structural differences to the references. The de novo genome assembly strategies, which assemble short reads into long sequences with the use of overlap between them instead of a reference genome, have been used in a wide range such as Celera Assembler (Venter et al. 2001), Arachne (Batzoglou et al. 2002), CAP3 (Huang and Madan 1999), Canu (Koren et al. 2017), miniasm (Li 2015) and wtdbg (Ruan and Li 2020).

Scaffold Filling The assembly algorithms merge short reads into "long sequences" of contigs or scaffolds, which are considered as draft genomes (Huson et al. 2002). Such drafted genomes often introduce errors due to their incompleteness. Thus, the scaffold filling problem is proposed to restore the whole genomes via computation across assembled long sequences (Muoz et al. 2010). This problem can be solved in two scenarios: a one-sided mode that fill a single scaffold from other sequences with missing genes, and two-sided mode can fill two scaffolds from each other. Usually, the scaffold filling strategy aims to minimize the double cut and join (DCJ) distance among different sequences, which is a polynomial-time algorithm for whole-genome reconstruction (Muoz et al. 2010). On the other side, some alternative sequence distance metrics are also used as optimization objectives with better performance in accuracy or reliability for scaffolds filling, such as the breakpoint distance and adjacency number distance, while they can introduce additional computing complexity. For example, the adjacency number-based scaffold filling is NP-Hard (Jiang et al. 2010). Thus, approximated solutions of such algorithms are then developed to reduce the high time-complexity, e.g. with adjacency number distance, one-sided scaffold filling can achieve a 1.2-approximation local search algorithm (Jiang et al.

2011; Jiang et al. 2012; Ma and Jiang 2016; Liu et al. 2013) and two-sided scaffold filling reached a performance ratio 1.4 + ε (Ma et al. 2021).

Genome Variation Detection Since the aforementioned software and algorithms have highly promoted genome variant detection, studies on related human diseases such as cancer (Norris et al. 2016; Macintyre et al. 2016), mendelian disorders (Sanchis-Juan et al. 2018), autism (Hedges et al. 2012), and Alzheimer (Qiang et al. 2017) have been consequently improved. Generally, genome variants can be divided into three types, single nucleotide variant (SNV) (Mackeh et al. 2018; Poirion et al. 2018; Kleftogiannis et al. 2019), small insertion/deletion (indel) (Ratan et al. 2015; Ferlaino et al. 2017), and structural variant (Nakagawa and Fujita 2018; Piazza and Heyer 2019). Among them, structural variant is most frequently associated with genetic diseases (Carvalho and Lupski 2016; Gonzalez-Garay 2014). Based on next-generation sequencing data, Some robust algorithms have been developed for detecting structural variation, for example, Hydra (Sindi et al. 2009) (based on pair-end mapping), CNVnator (Abyzov et al. 2011) (based on read depth), AGE (Abyzov and Gerstein 2011) (based on split read), SOAPdenovo2 (Luo et al. 2012) (based on de novo assembly). In recent years, the third-generation sequencing technology (PacBio or Oxford Nanopore) lays the foundation to improve the detection of structural variants (English et al. 2015; Wenger et al. 2019), for it extends the length of the sequence to several thousand of base pairs. In this case, some tools or software packages have been implemented to detect structural variants based on long DNA reads from sequencers, such as Sniffles (Sedlazeck et al. 2018), SVIM (Heller and Vingron 2019) and SVLR (Gu et al. 2021). Therefore, the emergence of new technology makes genome variant detection more accurate and efficient and brings new opportunities for genetic disease treatments.

8.3 Transcriptome Analysis Enables the in-Depth Elucidation of Disease Mechanisms

Alternative splicing occurs as a normal phenomenon in eukaryotes, and it allows a gene to express multiple transcripts, which plays an important role in regulating gene expression and producing the diversity of proteins (Mo and M.J. L 2009; Baralle and Jimena 2017; Kelemen et al. 2013). Numerous researches have revealed that a great number of human diseases, especially cancer, are related to abnormal splicing (Climente-González et al. 2017; A, S et al. 2016; Giuseppe et al. 2019; André and C.T. A 2003). The identification and analysis of expressed transcripts play an important role in diseases researches. In this section, we focus on three key steps of transcriptome studies, including transcriptome assembly, differential expression analysis, and abnormal splicing detection.

Transcriptome Assembly Advances in High-throughput RNA sequencing (RNA-seq) have opened the door to transcriptome reconstruction (Fatih and

M.P. M 2011; Peng et al. 2016). The RNA-seq protocol takes the expressed transcripts as input and outputs millions of short reads. In principle, such short reads can allow us to recover all expressed transcripts. However, this task is complicated by various alternative splicing variants, highly similar paralogs, different expression levels of transcripts of the same gene, and sequencing errors and bias (Jin et al. 2021). Like genome assembly, existing transcriptome assembly strategies can also be generally divided into two categories: reference-based and de novo assembly approach. On a high level, both reference-based and de novo strategies generally first construct graphs to represent splicing variants and then traverse the graph to extract paths as recovered transcripts. However, the reference-based assemblers such as StringTie (Mihaela et al. 2015; Sam et al. 2019), TransRef (Ting et al. 2021), Scallop (Mingfu and Carl 2017), TransComb (Juntao et al. 2016a), Cufflinks (Cole et al. 2012), Bayesember (Lasse et al. 2014), CIDANE (Stefan et al. 2016), iReckon (M. M.A et al. 2013a), and Scripture (Mitchell et al. 2010) usually construct graphs based on the alignments between reference genomes and RNA-seq reads. De novo assemblers such as TransLiG (Juntao et al. 2019), rnaSPAdes (Elena et al. 2019), MultiTrans (Jin et al. 2021), BinPacker (Juntao et al. 2016b), SOAPdenovo-trans (Yinlong et al. 2014), IDBA-Tran (Yu et al. 2013), Oases (H, S.M. 2012), and Trinity (Segata et al. 2011) construct graphs according to the overlaps between reads. The reference-based assemblers usually achieve higher accuracy than de novo assemblers when a high-quality reference genome is available; on the other hand, de novo assemblers are desired when the reference genome is unknown, incomplete, or substantially altered as in cancer tissues.

Differential Expression Analysis Differential expression analysis aims to find genes or transcripts differentially expressed between conditions, which is an integral part of understanding the molecular basis of phenotypic variation and diseases (D, R. M., et al. 2010). Such analysis is always performed in the following steps. Firstly, estimating the expression levels for each gene or transcript, which are usually measured by RPKM (the number of reads per kilobase of the gene/transcript per million reads mapped to the gene) and FPKM (the expected number of fragments per kilobase of gene/transcript per million fragments sequenced). Then, normalizing and identifying differentially expressed genes. There are three types of approaches for this step: (*i*) data-driven reference normalization such as GRSN (Carl et al. 2008), KDWL (Wen-Ping et al. 2011), KDQ (Qi et al. 2005), IRON (A, W.E et al. 2013b), BSN (Aanes et al. 2014) and SVR (Meng et al. 2019), (*ii*) extra-control reference normalization such as Spike-in Controls (Lovén et al. 2012) and wcloess normalization (Di et al. 2013), and (*iii*) all-gene reference normalization such as CrossNorm (Lixin et al. 2016a), ICN (Lixin et al. 2016b) and GPA normalization (Huiling et al. 2008). Finally, evaluate the relevance of the produced data with biological phenotypes of interest. Usually, two alternative methods are widely used: parametric and non-parametric. Parametric methods such as edgeR (D, R.M., et al. 2010) and baySeq (J, H.T., and K.K. A 2010) are applied to the normalized data (each expression for a given gene is mapped into a particular distribution, such as Poisson and negative binomial) while non-parametric methods including NOIseq (Sonia et al. 2012) and SAMseq (Li and Tibshirani 2013) take into consideration that data distribution cannot be defined from a finite set of parameters, such that non-parametric methods can capture more details about the data distribution.

Abnormal Splicing Detection Recent studies have shown that disruption of normal programs of splicing regulated by different splicing factors can lead to human diseases. For example, Eglė et al. revealed that alternative splicing is related to Alzheimer's and Parkinson's diseases (Eglė and Arvydas 2021). Recognition of aberrations in splicing events that are associated with disease has contributed to our understanding of disease pathogenesis. Methods for detecting alternative splicing can be categorized into two quantification schemas: count-based and isoform resolution strategies (Ruolin et al. 2014). For differential splicing, the count-based models such as DEXSeq (Simon et al. 2012), SDGseq (Wang et al. 2013), SplcingCompass (Moritz et al. 2013), rMATs (Shihao et al. 2014), rDiff-parametric (Philipp et al. 2013) and SeqGSEA (Xi and C.M. J 2014) usually configure each gene into a single representation consisting of counting units. Counting units can be full or truncated exonic regions or junction regions. Instead of transforming the question into detecting differential usage of counting units, isoform resolutions models (e.g. Cufflinks (Cole et al. 2012) and DiffSplice (Carl et al. 2008)) usually directly compare the relative isoform abundance across samples or conditions. These two methods have their own advantages and disadvantages, that is the count-based models can accurately discover the local differences while isoform resolution models detect the aberrations in isoform levels.

8.4 A New Sight on Human-Microbe Associations by Data Mining of Microbiome

The human microbiome can provide a novel sight to study on the relationship between microbial communities and their hosts (Blaser et al. 2016). In the past decade, massive volume of microbiome samples has been collected to discover the microbial- associations with human health (Forslund et al. 2015; Halfvarson et al. 2017; Poore et al. 2020; Qin et al. 2010). Meta-analysis on multiple cohorts can produce reliable and reproducible results for further applications (Wirbel et al. 2019; Bisanz et al. 2019; Armour et al. 2019). Profiling for taxonomic or functional features is the basis to study the microbiome, which mainly relies on DNA sequencing (Knight et al. 2018). In general, two sequencing approaches have been widely adopted: amplicon sequencing using marker genes (e.g. 16S rRNA, 18S rRNA or ITS) for taxonomy recognition, and shotgun metagenomic whole-genome sequencing (WGS) that survey genome-wide sequences of all species in a specimen.

Microbiome Profiling For short sequence fragments of marker genes, a series tools have been developed for taxonomy classification by reads clustering and OTU (Operational Taxonomic Units) picking based on sequence similarity such as

UPARSE (Edgar 2013) and Usearch (Edgar 2010). To further improve the accuracy of marker-gene-based analysis on single-nucleotide level, amplicon sequence variants (ASVs) denoising algorithms like DADA2 (Callahan et al. 2016), Deblur (Amir et al. 2017) and UNOISE3 (Edgar 2016) are then used in the latest works, which provide higher reliability, reproducibility and comprehensiveness than regular OTUs (Callahan et al. 2017). PICRUSt (Langille et al. 2013; Douglas et al. 2020) and Tax4Fun (Asshauer et al. 2015) even predict the functional profiles from amplicons using the pre-processed information between marker genes and reference whole genomes. In addition, comprehensive pipelines including QIIME (Caporaso et al. 2010; Bolyen et al. 2019), Mothur (Schloss et al. 2009) and Parallel-Meta (Jing et al. 2017; Chen et al. 2022) integrate most of these profiling methods with additional statistical analysis on alpha diversity and beta diversity. As a cost-efficient approach, amplicon-based analysis has been widely used in large number of microbiome studies, however, the precision is also challenged due to PCR bias (Jones et al. 2015), limited resolution of short-read markers and insufficient marker-genome associations. For instance, taxonomy annotation by amplified regions of 16S rRNA gene fragments is always on genus level (Edgar 2018; Yarza et al. 2014), and function inference might not be reliable for shortage of reference genomes (Langille et al. 2013).

With higher scale of sequence numbers, WGS is more informative as well. Thus, some approaches employ unassembled WGS short reads for species or strain level taxonomy identification (Ye et al. 2019; Scholz et al. 2016) (e.g. Karken (Wood and Salzberg 2014), mOTUs (Sunagawa et al. 2013), and MetaPhlAn2 (Segata et al. 2012)) and function parsing (e.g. HUMANn2 (Franzosa et al. 2018)). On the other side, binning- or assembling-based method (e.g. metaSPAdes (Bankevich et al. 2012), meta-IDBA (Peng et al. 2012) and MetaWRAP (Uritskiy et al. 2018)) are suitable for genome re-construction, de novo gene prediction with unknown species, as well as single nucleotide polymorphism (SNP) analysis. However, the broad-range application of WGS is also suffered from the 3–10 folds higher overall cost including sequencing, data storage and sharing, bioinformatics processing of reads quality control (Zhou et al. 2014; Zhou et al. 2018), profiling (Ye et al. 2019; Franzosa et al. 2018) than those of amplicons (Langille et al. 2013; Jing et al. 2017; Rognes et al. 2016; Lu and Salzberg 2020). Recently, a new library preparation protocol named ‘shallow shotgun sequencing’ achieves species-level accuracy similar to that offered by regular ones, making the WGS in a more economical way (Hillmann et al. 2018).

Different from specific targeted variable regions of mark gene amplification, full-length 16S rRNA gene by the third generation sequencing platforms has the potential of accurate microbiome classification at species or strain resolution (Johnson et al. 2019). To couple with such advantages by long-read sequencing platform data, algorithms or tools for sequence denoising, clustering and annotation should also be updated accordantly. Thus, the rapid development of microbiome profiling methods provides the fundamental of a broader view to the “microbiome data universe”.

Microbiome Data Integration Massive microbiomes have been produced by previous works such as Human Microbiome Project (Proctor et al. 2019), Earth Microbiome Project (Thompson et al. 2017) and American Gut Project (McDonald et al. 2018a). Usually, sequences are deposited are shared in online repositories, e.g. NCBI-SRA (Kodama et al. 2012), MG-RAST (Meyer et al. 2008), EBI Metagenomics (Harrison et al. 2019), JGI-IMG/M (Chen et al. 2019), MPD (Zhang et al. 2018) and so on. Such big data provides the "materials" for study on the global-wide microbial diversity and distribution, while also brings new problems in data integration and re-analysis. In these repositories, most specimens are sorted by their original studies, and metadata among studies are not normalized for feature selection and comparison, leading to the difficulty for revealing microbial patterns under a specific conditions or phenotypes.

Several works re-organized the microbiome data with unified metadata format (Yilmaz et al. 2011; Buttigieg et al. 2016) and re-processed the DNA sequences by standard operating procedures (SOPs) (Ten Hoopen et al. 2017), enabling the utilization and reusability of valuable microbiome big-data for further meta-analysis. GMrepo (Wu et al. 2020) database curated human gut metagenomes with constant profiling procedures and detailed metadata of host variables. Qiita (Gonzalez et al. 2018; McDonald et al. 2019) enables meta-analysis across different studies, and retrieve microbiomes with specific features (e.g. metadata, taxon terms, and sequence fragments) by SQL-like queries. In addition, Gc-Meta (Shi et al. 2019) implemented a data management system integrated with bioinformatical tools and workflows for analyzing data in a standardized way.

Recently, a Microbiome Search Engine (MSE) (Su et al. 2018; Jing et al. 2021a) has been proposed for rapid match of microbiomes in a "community to communities" mode. By a dynamic index and whole-microbiome-level similarity scoring functions (Jing et al. 2019; Su et al. 2014), MSE enables the real-time accessibility of microbiomes with aimed structure from huge number of samples. In this way, with a newly sequenced microbiome, people can answer what existing samples in the repositories or databases have an overall similar community to it, thus predict the environmental conditions or health status.

Technical variation is a key barrier to integrate microbiome datasets from multiple sources and batches. These factors mainly include (but not limited to) DNA extraction, PCR primers for marker genes, sub-regions of the marker gene amplification, sequencing platforms and types of sequence reads (Costea et al. 2017; Xiao et al. 2022). Technical differences can outweigh the taxonomy diversity among sample groups (Hacquard et al. 2015; Lozupone et al. 2013), e.g. human microbiomes from hosts with different states, ages, locations and diets, interfering the cross-study comparison, even such variation can be reduced by computational methods (Jing et al. 2021b). Thus, amplicon-based studies still need unified experimental protocols. In contrast, shotgun WGS is less sensitive to technical variations (Wirbel et al. 2019; Voigt et al. 2015), which is an optimal option for integration and comparison of cross-study datasets.

8.5 High-Resolution Bioinformatics on Single-Cell Level

One of the most important features of a microbial community is the complex interspecies interactions. For some diseases, dynamics among status is always associated with a series of microbiome structure variations. Detection and description of such variation is the basis to understand the action mode of the microbiome. However, the widely-used microbiome sequencing approaches typically only profile the overall taxonomy composition (e.g. species) or functional structure (e.g. gene families). On the other hand, multi-omics combination of the meta-transcriptome, meta-proteome, or meta-metabolome faces tremendous challenges in rapid-response monitoring of microbiome status for their destructive nature, tedious operation, and high cost. Furthermore, the significant shortage in biomarkers of the microbiome also presents major technical hurdles in tracing microbiome functions.

Fluorescence-Activated Cell Sorting Single-cell approaches including functional sorting, sequencing, and cultivation potentially enable the discrimination and validation of the function of individual cells within a microbial community. Currently, most function-based cell-sorting methods are based on fluorescence-activated cell sorting (FACS) (Song et al. 2017). However, FACS always requires cell labeling with fluorescent probes that target to specific proteins, metabolites, or nucleic acids, thus requiring *a priori* knowledge about the biomarkers of the aimed functions.

Single-Cell Ramanome "Ramanome" is a label-free, single-cell-level functional imaging tools that proposed for the "instant photography" of microorganisms (Lin et al. 2016). It is a collection of single-cell Raman spectra (SCRS) captured from individual cells within an isogenic population or consortium (Lin et al. 2016). Each SCRS consists of over 1500 Raman bands that correspond to the chemical bonds from the metabolites in a single cell, thus can be used to depict the profile of metabolites in the cell. Since the metabolite profile is sensitive to the genetic background, physiological state, and environmental changes of a cell, each SCRS can be considered as a digital photo of ~1500 pixels. As an optical approach, the ramanome is non-destructive to the cell and does not require external labeling or biomarkers. Thus, a ramanome works as a single-cell-resolution metabolome that can be obtained and analyzed with high speed and low cost.

8.6 Machine Learning Brings the Opportunity of Disease Screening and Prediction in Precise Medicine

The volume of biological data has been intensively raised with the development of high-throughput sequencing. Usually, biological features are surveyed into either taxonomy units (e.g. species or OTUs) (Edgar 2013; Edgar 2010), or metabolic functions (e.g. gene families or pathways) (Franzosa et al. 2018; Truong et al. 2015). Then machine learning (ML) algorithms can reveal biological patterns of these features under different statuses, promoting data-driven disease detection and screening (Namkung 2020; Topçuoğlu et al. 2020; Cammarota et al. 2020). One of the most prevalent machine learning techniques, supervised classification, plays important role in the pattern recognition of human diseases (Duvallet et al. 2017a; Vangay et al. 2019) such as gingivitis (Huang et al. 2014; Huang et al. 2021), cancer (Poore et al. 2020; Wirbel et al. 2019), diabetes (Bajaj et al. 2012), inflammatory bowel disease (IBD) (Halfvarson et al. 2017; Gevers et al. 2014), etc. By training classifiers and models using taxonomical or functional profiles from patients and their healthy control, ML classifiers like support vector machines (SVM) (Cortes and Vapnik 1995) and random forest (RF) (Breiman 2001) can identify the status of new samples.

Basically, the data-based disease detection is a classification problem using biological profiles (Su et al. 2020a). With profiles $X = \{x_1, \ldots, x_n\}$ for n samples (here x_i is the profile of a sample that can be represented by the richness of features like species, genes, etc.) and their corresponding status meta-data (label) $Y = \{y_1, \ldots, y_n\}$, ML classifier solves a function $Y = f(X)$ that maps the profiles to their meta-data, thus identify the status of a new sample based on its profile. Here label y_i in meta-data Y is a discrete variable that represents a status (e.g. a specific disease). We review the commonly used three types of ML classifiers for health status detection, mainly including individual classifiers, ensemble classifiers and neural networks.

Individual Classifiers Logistic regression (LR) is a linear model based on a logistic function to model a binary variable (Kleinbaum et al. 2002). LR calculates the probability for an event, e.g. a microbiome sample is healthy or not. For its high interpretability and efficiency, LR is widely used in biological feature based disease recognition (Topçuoğlu et al. 2020; Song et al. 2020), although the accuracy is not as good as other methods. Support vector machine (SVM) captures non-linear associations between biological profiles and their phenotypes to obtain the maximum margin of samples that belong to different groups (Cortes and Vapnik 1995), thus exhibits better performance rather than LR. Another approach is k-nearest neighbors (k-NN), which directly labels an unknown sample by its nearest neighbors (Peterson 2009). The key problem of k-NN is how to suitably evaluate the neighborship among samples (Comin et al. 2020). For example, relations among microbiomes are always measured by distance metrices such as JCCARD, Bray-Curtis and Jensen-Shannon Divergence (Ricotta and Podani 2017), or phylogeny-based distances like UniFrac (McDonald et al. 2018b) or Meta-Storms (Jing et al. 2019) algorithms. Recently, A search-based approach uses microbiome search engine (MSE) (Jing et al. 2021a) to discriminate unhealthy microbiomes by a novelty score, and further detect their disease types via phylogeny-based distances, which outperforms regular k-NN in speed, robustness and sensitivity (Su et al. 2020b).

Ensemble Classifiers To further enhance the precision for status classification and disease detection, ensemble classifier are developed by integrating individual ML classifiers (Zhou 2009; Polikar 2012). Random forest (RF) builds multiple decision trees by randomly selecting samples and features in training data, then combines the predicted results of new samples by voting (Namkung 2020; Topçuoğlu et al. 2020;

Breiman 2001; LaPierre et al. 2019). Another approach, gradient boosting decision tree (GBDT), can assign a weight to each single sample, constructs the tree-like models in a stage-wise way (Friedman 2001; Friedman 2002), and iteratively updates parameters for minimum estimation errors (Ruder 2016). The ensemble classifiers of RF and GBDT not only exhibit advantages in precision than individual classifiers, but can also quantify and sort the importance of each biological features in status classification (Duvallet et al. 2017b; Pasolli et al. 2016).

Neural Networks For classification, feature extraction from input data is crucial for sensitivity and accuracy (Pouyanfar et al. 2018). Different from regular ML approaches, neural network based deep learning automatically performs feature selection and trains deep networks in an end-to-end mode (Glasmachers 2017), which also reduces the high dimensionality and sparsity by the data complexity and diversity. Neural networks (such as deep neural networks (DNNs) (Deng et al. 2016), recurrent neural networks (RNNs) (Mou et al. 2017), convolutional neural networks (CNNs) (Gu et al. 2018), etc.) have been transplanted from image processing to bioinformatical research with additional efforts in data adaption. For example, in computer vision, CNNs generate new variables using convolution operations on spatial neighboring pixels. However, such neighborship among biological features such as microbes or metabolic functions are not clear. To tackle this problem, Sharma et al. (Sharma et al. 2020) developed a novel CNN method by collapsing taxa to phylum-level clusters for human microbiome-based classification. Lo et al. (Lo and Marculescu 2019) also mapped biological features into a negative binomial distribution and solved over-fitting problem by data augmentation in CNNs.

References

A, S, et al. Aberrant RNA splicing in cancer; expression changes and driver mutations of splicing factor genes. Oncogene. 2016;35(19)

A, W.E, et al. Iterative rank-order normalization of gene expression microarray data. BMC Bioinformatics. 2013b;14:1.

Aanes H, et al. Normalization of RNA-sequencing data from samples with varying mRNA levels. PLoS One. 2014;9(2):e89158.

Abyzov A, Gerstein M. AGE: defining breakpoints of genomic structural variants at single-nucleotide resolution, through optimal alignments with gap excision. Bioinformatics. 2011;27 (5):595–603.

Abyzov A, et al. CNVnator: an approach to discover, genotype, and characterize typical and atypical CNVs from family and population genome sequencing. Genome Res. 2011;21(6): 974–84.

Amir A, et al. Deblur rapidly resolves single-nucleotide community sequence patterns. mSystems. 2017;2:2.

André FN, C.T. A. Pre-mRNA splicing and human disease. Genes Dev. 2003;17(4)

Armour CR, et al. A metagenomic meta-analysis reveals functional signatures of health and disease in the human gut microbiome. mSystems. 2019;4:4.

Asshauer KP, et al. Tax4Fun: predicting functional profiles from metagenomic 16S rRNA data. Bioinformatics. 2015;31(17):2882–4.

Bajaj JS, et al. Linkage of gut microbiome with cognition in hepatic encephalopathy. Am J Physiol Gastrointest Liver Physiol. 2012;302(1):G168-75.

Bankevich A, et al. SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. J Comput Biol. 2012;19(5):455–77.

Baralle FE, Jimena G. Alternative splicing as a regulator of development and tissue identity. Nat Rev Mol Cell Biol. 2017;18:7.

Batzoglou S, et al. ARACHNE: a whole-genome shotgun assembler. Genome Res. 2002;12(1): 177–89.

Bisanz JE, et al. Meta-analysis reveals reproducible gut microbiome alterations in response to a high-fat diet. Cell Host Microbe. 2019;26(2):265–72. e4

Blaser MJ, et al. Toward a predictive understanding of Earth's microbiomes to address 21st century challenges. MBio. 2016;7:3.

Bolyen E, et al. Reproducible, interactive, scalable and extensible microbiome data science using QIIME 2 (vol 37, pg 852, 2019). Nat Biotechnol. 2019;37(9):1091.

Brandler WM, et al. Frequency and complexity of de novo structural mutation in autism. Am J Hum Genet. 2016;98(4):667–79.

Breiman L. Random forests. Mach Learn. 2001;45(1):5–32.

Buttigieg PL, et al. The environment ontology in 2016: bridging domains with increased scope, semantic density, and interoperation. J Biomed Semantics. 2016;7(1):57.

Callahan BJ, McMurdie PJ, Holmes SP. Exact sequence variants should replace operational taxonomic units in marker-gene data analysis. ISME J. 2017;11(12):2639–43.

Callahan BJ, et al. DADA2: high-resolution sample inference from Illumina amplicon data. Nat Methods. 2016;13(7):581–3.

Cammarota G, et al. Gut microbiome, big data and machine learning to promote precision medicine for cancer. Nat Rev Gastroenterol Hepatol. 2020;

Caporaso JG, et al. QIIME allows analysis of high-throughput community sequencing data. Nat Methods. 2010;7(5):335–6.

Carl P, et al. Global rank-invariant set normalization (GRSN) to reduce systematic distortions in microarray data. BMC Bioinformatics. 2008;9:1.

Carvalho CM, Lupski JR. Mechanisms underlying structural variant formation in genomic disorders. Nat Rev Genet. 2016;17(4):224–38.

Chen IA, et al. IMG/M v.5.0: an integrated data management and comparative analysis system for microbial genomes and microbiomes. Nucleic Acids Res. 2019;47(D1):D666–77.

Chen W, et al. Mapping translocation breakpoints by next-generation sequencing. Genome Res. 2008;18(7):1143–9.

Chen Y, et al. Parallel-meta suite: interactive and rapid microbiome data analysis on multiple platforms. iMeta. 2022;1(1):e1.

Climente-González H, et al. The functional impact of alternative splicing in cancer. Cell Rep. 2017;20:9.

Cole T, et al. Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and cufflinks. Nat Protoc. 2012;7:3.

Comin M, et al. Comparison of microbiome samples: methods and computational challenges. Brief Bioinform. 2020;

Cortes C, Vapnik V. Support-vector networks. Mach Learn. 1995;20(3):273–97.

Costea PI, et al. Towards standards for human fecal sample processing in metagenomic studies. Nat Biotechnol. 2017;35(11):1069–76.

D, R.M., M.D. J, and S.G. K, edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics Oxford, England, 2010. 26(1).

Deng Y, et al. A hierarchical fused fuzzy deep neural network for data classification. IEEE Trans Fuzzy Syst. 2016;25(4):1006–12.

Di W, et al. The use of miRNA microarrays for the analysis of cancer samples with global miRNA decrease, vol. 19. New York, N.Y: RNA; 2013. p. 7.

Douglas GM, et al. PICRUSt2 for prediction of metagenome functions. Nat Biotechnol. 2020;38 (6):685–8.

Duvallet C, et al. Meta-analysis of gut microbiome studies identifies disease-specific and shared responses. Nat Commun. 2017a;8(1):1784.

Duvallet C, et al. Meta-analysis of gut microbiome studies identifies disease-specific and shared responses. Nat Commun. 2017b;8(1):1–10.

Edgar RC. Search and clustering orders of magnitude faster than BLAST. Bioinformatics. 2010;26 (19):2460–1.

Edgar RC. UPARSE: highly accurate OTU sequences from microbial amplicon reads. Nat Methods. 2013;10(10):996–8.

Edgar RC. UNOISE2: improved error-correction for Illumina 16S and ITS amplicon sequencing. bioRxiv. 2016:081257.

Edgar RC. Accuracy of taxonomy prediction for 16S rRNA and fungal ITS sequences. PeerJ. 2018;6:e4652.

Eglė J, Arvydas K. Alternative splicing and hypoxia puzzle in Alzheimer's and Parkinson's diseases. Genes. 2021;12:8.

Elena B, et al. rnaSPAdes: a de novo transcriptome assembler and its application to RNA-Seq data. GigaScience. 2019;8:9.

English AC, et al. Assessing structural variation in a personal genome—towards a human reference diploid genome. BMC Genomics. 2015;16(1):1–15.

Fatih O, M.P. M. RNA sequencing: advances, challenges and opportunities. Nat Rev Genet. 2011;12:2.

Ferlaino M, et al. An integrative approach to predicting the functional effects of small indels in non-coding regions of the human genome. BMC Bioinformatics. 2017;18(1):1–8.

Feuk L, Carson AR, Scherer SW. Structural variation in the human genome. Nat Rev Genet. 2006;7 (2):85–97.

Forslund K, et al. Disentangling type 2 diabetes and metformin treatment signatures in the human gut microbiota. Nature. 2015;528(7581):262–6.

Franzosa EA, et al. Species-level functional profiling of metagenomes and metatranscriptomes. Nat Methods. 2018;15(11):962–8.

Friedman JH. Greedy function approximation: a gradient boosting machine. Ann Stat. 2001:1189–232.

Friedman JH. Stochastic gradient boosting. Comput Stat Data Anal. 2002;38(4):367–78.

Gevers D, et al. The treatment-naive microbiome in new-onset Crohn's disease. Cell Host Microbe. 2014;15(3):382–92.

Giuseppe B, et al. Alternative splicing in Alzheimer's disease. Aging Clin Exp Res. 2019;33(4)

Glasmachers T. Limits of end-to-end learning. In: Min-Ling Z, Yung-Kyun N, editors. Proceedings of the ninth Asian conference on machine learning; 2017., PMLR: Proceedings of Machine Learning Research. p. 17–32.

Gonzalez A, et al. Qiita: rapid, web-enabled microbiome meta-analysis. Nat Methods. 2018;15(10): 796–8.

Gonzalez-Garay ML. The road from next-generation sequencing to personalized medicine. Pers Med. 2014;11(5):523–44.

Gu J, et al. Recent advances in convolutional neural networks. Pattern Recogn. 2018;77:354–77.

Gu W, et al. SVLR: genome structural variant detection using long-read sequencing data. J Comput Biol. 2021;

H, S.M. et al., Oases: robust de novo RNA-seq assembly across the dynamic range of expression levels. Bioinformatics Oxford, England, 2012. 28(8).

Hacquard S, et al. Microbiota and host nutrition across plant and animal kingdoms. Cell Host Microbe. 2015;17(5):603–16.

Halfvarson J, et al. Dynamics of the human gut microbiome in inflammatory bowel disease. Nat Microbiol. 2017;2:17004.

Harrison PW, et al. The European nucleotide archive in 2018. Nucleic Acids Res. 2019;47(D1): D84–8.

Hedges DJ, et al. Evidence of novel fine-scale structural variation at autism spectrum disorder candidate loci. Mol Autism. 2012;3(1):1–11.

Heller D, Vingron M. SVIM: structural variant identification using mapped long reads. Bioinformatics. 2019;35(17):2907–15.

Hillmann B, et al. Evaluating the information content of shallow shotgun metagenomics. Msystems. 2018;3:6.

Hood L, Rowen L. The human genome project: big science transforms biology and medicine. Genome Med. 2013;5(9):1–8.

Huang S, et al. Predictive modeling of gingivitis severity and susceptibility via oral microbiota. ISME J. 2014;8(9):1768–80.

Huang S, et al. Longitudinal multi-omics and microbiome meta-analysis identify an asymptomatic gingival state that links gingivitis, *Periodontitis, and Aging*. mBio. 2021;12:2.

Huang X, Madan A. CAP3: a DNA sequence assembly program. Genome Res. 1999;9(9):868–77.

Huiling X, et al. Using generalized procrustes analysis (GPA) for normalization of cDNA microarray data. BMC Bioinformatics. 2008;9:1.

Huson DH, Reinert K, Myers EW. The greedy path-merging algorithm for contig scaffolding. J ACM (JACM). 2002;49(5):603–15.

J, H.T, K.K. A. baySeq: empirical Bayesian methods for identifying differential expression in sequence count data. BMC Bioinformatics. 2010;11:1.

Jiang H, Zhong F, and Zhu B. *Filling scaffolds with gene repetitions: maximizing the number of adjacencies*. in *Annual Symposium on Combinatorial Pattern Matching*. 2011. Springer.

Jiang H, et al. Scaffold filling under the breakpoint distance. in RECOMB International Workshop on Comparative Genomics. Springer. 2010.

Jiang H, et al. Scaffold filling under the breakpoint and related distances. IEEE/ACM Trans Comput Biol Bioinform. 2012;9(4):1220–9.

Jin Z, et al. *MultiTrans: an algorithm for path extraction through mixed integer linear programming for transcriptome assembly*. IEEE/ACM transactions on computational biology and bioinformatics, 2021. **PP**.

Jing G, et al. Parallel-META 3: comprehensive taxonomical and functional analysis platform for efficient comparison of microbial communities. Sci Rep. 2017;7:40371.

Jing G, et al. Dynamic meta-storms enables comprehensive taxonomic and phylogenetic comparison of shotgun metagenomes at the species level. Bioinformatics. 2019;

Jing G, et al. Microbiome search engine 2: a platform for taxonomic and functional search of global microbiomes on the whole-microbiome level. mSystems. 2021a;6:1.

Jing G, et al. Meta-apo improves accuracy of 16S-amplicon-based prediction of microbiome function. BMC Genomics. 2021b;22(1):9.

Johnson JS, et al. Evaluation of 16S rRNA gene sequencing for species and strain-level microbiome analysis. Nat Commun. 2019;10(1):5029.

Jones MB, et al. Library preparation methodology can influence genomic and functional predictions in human microbiome research. Proc Natl Acad Sci U S A. 2015;112(45):14024–9.

Juntao L, et al. TransComb: genome-guided transcriptome assembly via combing junctions in splicing graphs. Genome Biol. 2016a;17:1.

Juntao L, et al. BinPacker: packing-based De novo transcriptome assembly from RNA-seq data. PLoS Comput Biol. 2016b;12:2.

Juntao L, et al. TransLiG: a de novo transcriptome assembler that uses line graph iteration. Genome Biol. 2019;20:1.

Kelemen O, et al. Function of alternative splicing. Gene. 2013;514:1.

Kleftogiannis D, et al. Identification of single nucleotide variants using position-specific error estimation in deep sequencing data. BMC Med Genet. 2019;12(1):1–12.

Kleinbaum DG, et al. Logistic regression. Springer; 2002.

Knight R, et al. Best practices for analysing microbiomes. Nat Rev Microbiol. 2018;16(7):410–22.

Kodama Y, et al. The sequence read archive: explosive growth of sequencing data. Nucleic Acids Res. 2012;40(Database issue):D54–6.

Koren S, et al. Canu: scalable and accurate long-read assembly via adaptive k-mer weighting and repeat separation. Genome Res. 2017;27(5):722–36.

Lander ES. Initial impact of the sequencing of the human genome. Nature. 2011;470(7333):187–97.

Langille MG, et al. Predictive functional profiling of microbial communities using 16S rRNA marker gene sequences. Nat Biotechnol. 2013;31(9):814–21.

LaPierre N, et al. MetaPheno: a critical evaluation of deep learning and machine learning in metagenome-based disease prediction. Methods. 2019;166:74–82.

Lasse M, Andreas SJ, Anders K. Bayesian transcriptome assembly. Genome Biol. 2014;15:10.

Li H. Minimap and miniasm: fast mapping and de novo assembly for noisy long sequences. Bioinformatics. 2015;32(14)

Li J, Tibshirani R. Finding consistent patterns: a nonparametric approach for identifying differential expression in RNA-Seq data. Stat Methods Med Res. 2013;22:5.

Lin T, et al. Label-free, rapid and quantitative phenotyping of stress response in E. coli via ramanome. Sci Rep. 2016;6:267.

Liu N, et al. An improved approximation algorithm for scaffold filling to maximize the common adjacencies. IEEE/ACM Trans Comput Biol Bioinform. 2013;10(4):905–13.

Lixin C, et al. CrossNorm: a novel normalization strategy for microarray data in cancers. Sci Rep. 2016a;6:1.

Lixin C, et al. ICN: a normalization method for gene expression data considering the over-expression of informative genes. Mol BioSyst. 2016b;12:10.

Lo C, Marculescu R. MetaNN: accurate classification of host phenotypes from metagenomic data using neural networks. Bmc Bioinformatics. 2019;20(12):314.

Lovén J, et al. Revisiting global gene expression analysis. Cell. 2012;151:3.

Lozupone CA, et al. Meta-analyses of studies of the human microbiota. Genome Res. 2013;23(10): 1704–14.

Lu J and Salzberg SL, *Ultrafast and accurate 16S microbial community analysis using Kraken 2.* bioRxiv, 2020: p. 2020.03.27.012047.

Luo R, et al. SOAPdenovo2: an empirically improved memory-efficient short-read de novo assembler. Gigascience. 2012;1(1):2047-217X-1-18.

M. M.A, et al. iReckon: simultaneous isoform discovery and abundance estimation from RNA-seq data. Genome Res. 2013a;23:3.

Ma J and Jiang H. *Notes on the 6/5-Approximation Algorithm for One-Sided Scaffold Filling*. in *International Workshop on Frontiers in Algorithmics*. Springer; 2016.

Ma J, et al. On the solution bound of two-sided scaffold filling. Theor Comput Sci. 2021;873:47–63.

Macintyre G, Ylstra B, Brenton JD. Sequencing structural variants in cancer for precision therapeutics. Trends Genet. 2016;32(9):530–42.

Mackeh R, et al. Single-nucleotide variations of the human nuclear hormone receptor genes in 60,000 individuals. J Endocr Soc. 2018;2(1):77–90.

McDonald D, et al. American gut: an open platform for citizen science microbiome research. mSystems. 2018a;3:3.

McDonald D, et al. Striped UniFrac: enabling microbiome analysis at unprecedented scale. Nat Methods. 2018b;15(11):847–8.

McDonald D, et al. Redbiom: a rapid sample discovery and feature characterization system. mSystems. 2019;4(4)

Meng Z, et al. Analysis of long noncoding RNAs highlights region-specific altered expression patterns and diagnostic roles in Alzheimer's disease. Brief Bioinform. 2019;20:2.

Meyer F, et al. The metagenomics RAST server - a public resource for the automatic phylogenetic and functional analysis of metagenomes. BMC Bioinformatics. 2008;9:386.

Mihaela P, et al. StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. Nat Biotechnol. 2015;33:3.

Mingfu S, Carl K. Accurate assembly of transcripts through phase-preserving graph decomposition. Nat Biotechnol. 2017;35:12.

Mitchell G, et al. Ab initio reconstruction of cell type-specific transcriptomes in mouse reveals the conserved multi-exonic structure of lincRNAs. Nat Biotechnol. 2010;28:5.

Mo C, M.J. L. Mechanisms of alternative splicing regulation: insights from molecular and genomics approaches. Nat Rev Mol Cell Biol. 2009;10:11.

Moritz A, et al. SplicingCompass: differential splicing detection using RNA-seq data, vol. 29. Oxford, England: Bioinformatics; 2013. p. 9.

Mou L, Ghamisi P, Zhu XX. Deep recurrent neural networks for hyperspectral image classification. IEEE Trans Geosci Remote Sens. 2017;55(7):3639–55.

Muoz A, et al. Scaffold filling contig fusion and gene order comparison. BMC Bioinformatics. 2010;11:304.

Nakagawa H, Fujita M. Whole genome sequencing analysis for cancer genomics and precision medicine. Cancer Sci. 2018;109(3):513–22.

Nalbantoglu U, et al. Large direct repeats flank genomic rearrangements between a new clinical isolate of Francisella tularensis subsp. tularensis A1 and Schu S4. PLoS One. 2010;5(2):e9007.

Namkung J. Machine learning methods for microbiome studies. J Microbiol. 2020;58(3):206–16.

Norris AL, et al. Nanopore sequencing detects structural variants in cancer. Cancer Biol Ther. 2016;17(3):246–53.

Ozery-Flato M, Shamir R. Sorting cancer karyotypes by elementary operations. J Comput Biol. 2009;16(10):1445–60.

Pasolli E, et al. Machine learning meta-analysis of large metagenomic datasets: tools and biological insights. PLoS Comput Biol. 2016;12(7):e1004977.

Peng L, et al. Integrative analysis with ChIP-seq advances the limits of transcript quantification from RNA-seq. Genome Res. 2016;26:8.

Peng Y, et al. IDBA-UD: a de novo assembler for single-cell and metagenomic sequencing data with highly uneven depth. Bioinformatics. 2012;28(11):1420–8.

Peterson LE. K-nearest neighbor. Scholarpedia. 2009;4(2):1883.

Philipp D, et al. Accurate detection of differential RNA processing. Nucleic Acids Res. 2013;41:10.

Piazza A, Heyer W-D. Homologous recombination and the formation of complex genomic rearrangements. Trends Cell Biol. 2019;29(2):135–49.

Poirion O, et al. Using single nucleotide variations in single-cell RNA-seq to identify subpopulations and genotype-phenotype linkage. Nat Commun. 2018;9(1):1–13.

Polikar R. *Ensemble learning*, in *Ensemble machine learning*. 2012, Springer. p. 1–34.

Poore GD, et al. Microbiome analyses of blood and tissues suggest cancer diagnostic approach. Nature. 2020;579(7800):567–74.

Pop M. Genome assembly reborn: recent computational challenges. Brief Bioinform. 2009;10(4): 354–66.

Pouyanfar S, et al. A survey on deep learning: algorithms, techniques, and applications. ACM Computing Surveys (CSUR). 2018;51(5):1–36.

Proctor LM, et al. The integrative human microbiome project. Nature. 2019;569(7758):641–8.

Qi F, et al. *Improved probe selection for DNA arrays using nonparametric kernel density estimation.* Conference proceedings : ... Annual International Conference of the IEEE Engineering in Medicine and Biology Society. IEEE Engineering in Medicine and Biology Society. Annual Conference, 2005. **2006**.

Qiang et al. *Structural variation in amyloid-beta fibrils from Alzheimer's disease clinical subtypes.* Nature, 2017.

Qin J, et al. A human gut microbial gene catalogue established by metagenomic sequencing. Nature. 2010;464(7285):59–65.

Randal J. The human genome project. Lancet. 1991;334(8678):1535–6.

Rasko DA, et al. Bacillus anthracis comparative genome analysis in support of the Amerithrax investigation. Proc Natl Acad Sci. 2011;108(12):5027–32.

Ratan A, et al. Identification of indels in next-generation sequencing data. BMC Bioinformatics. 2015;16(1):1–8.

Ricotta C, Podani J. On some properties of the Bray-Curtis dissimilarity and their ecological meaning. Ecol Complex. 2017;31:201–5.

Rognes T, et al. VSEARCH: a versatile open source tool for metagenomics. PeerJ. 2016;4:e2584.

Ruan J, Li H. Fast and accurate long-read assembly with wtdbg2. Nat Methods. 2020;17(2):155–8.

Ruder S., *An overview of gradient descent optimization algorithms.* arXiv preprint arXiv:1609.04747, 2016.

Ruolin L, L.A. E, D.J. A. Comparisons of computational methods for differential alternative splicing detection using RNA-seq in plant systems. BMC Bioinformatics. 2014;15:1.

Sam K, et al. Transcriptome assembly from long-read RNA-seq alignments with StringTie2. Genome Biol. 2019;20:1.

Sanchis-Juan A, et al. Complex structural variants in Mendelian disorders: identification and breakpoint resolution using short-and long-read genome sequencing. Genome Med. 2018;10 (1):1–10.

Sankoff D, et al. Gene order comparisons for phylogenetic inference: evolution of the mitochondrial genome. Proc Natl Acad Sci. 1992;89(14):6575–9.

Schloss PD, et al. Introducing mothur: open-source, platform-independent, community-supported software for describing and comparing microbial communities. Appl Environ Microbiol. 2009;75(23):7537–41.

Scholz M, et al. Strain-level microbial epidemiology and population genomics from shotgun metagenomics. Nat Methods. 2016;13(5):435–8.

Sedlazeck FJ, et al. Accurate detection of complex structural variations using single-molecule sequencing. Nat Methods. 2018;15(6):461–8.

Segata N, et al. Metagenomic biomarker discovery and explanation. Genome Biol. 2011;12(6):R60.

Segata N, et al. Metagenomic microbial community profiling using unique clade-specific marker genes. Nat Methods. 2012;9(8):811–4.

Sharma D, Paterson AD, Xu W. TaxoNN: ensemble of neural networks on stratified microbiome data for disease prediction. Bioinformatics. 2020;

Shi W, et al. gcMeta: a global catalogue of metagenomics platform to support the archiving, standardization and analysis of microbiome data. Nucleic Acids Res. 2019;47(D1):D637–48.

Shihao, S., et al., *rMATS: robust and flexible detection of differential alternative splicing from replicate RNA-Seq data.* Proc Natl Acad Sci U S A, 2014. 11151.

Simon A, Alejandro R, Wolfgang H. Detecting differential usage of exons from RNA-seq data. Genome Res. 2012;22:10.

Sindi S, et al. A geometric approach for classification and comparison of structural variants. Bioinformatics. 2009;25(12):i222–30.

Song B, et al. MetaSee: an interactive and extendable visualization toolbox for metagenomic sample analysis and comparison. PLoS One. 2017:7, 11.

Song K, Wright F, Zhou Y-H. Systematic comparisons for composition profiles, taxonomic levels, and machine learning methods for microbiome-based disease prediction. Front Mol Biosci. 2020;7:423.

Sonia T, et al. NOIseq: a RNA-seq differential expression method robust for sequencing depth biases. EMBnetjournal. 2012;17(B)

Stefan C, et al. CIDANE: comprehensive isoform discovery and abundance estimation. Genome Biol. 2016;17:1.

Su X, et al. GPU-meta-storms: computing the structure similarities among massive amount of microbial community samples using GPU. Bioinformatics. 2014;30(7):1031–3.

Su X, et al. Identifying and predicting novelty in microbiome studies. MBio. 2018;9:6.

Su X, et al. *Method development for cross-study microbiome data mining: challenges and opportunities*. Comput Struct Biotechnol J. 2020a;

Su X, et al. Multiple-disease detection and classification across cohorts via microbiome search. Msystems. 2020b;5:2.

Sudmant PH, et al. An integrated map of structural variation in 2,504 human genomes. Nature. 2015;526(7571):75–81.

Sunagawa S, et al. Metagenomic species profiling using universal phylogenetic marker genes. Nat Methods. 2013;10(12):1196.

Ten Hoopen P, et al. The metagenomic data life-cycle: standards and best practices. Gigascience. 2017;6(8):1–11.

Thompson LR, et al. A communal catalogue reveals Earth's multiscale microbial diversity. Nature. 2017;551(7681):457–63.

Ting Y, et al., *TransRef enables accurate transcriptome assembly by redefining accurate neo-splicing graphs.* Briefings in bioinformatics, 2021.

Topçuoğlu BD, et al. A framework for effective application of machine learning to microbiome-based classification problems. MBio. 2020;11:3.

Truong DT, et al. MetaPhlAn2 for enhanced metagenomic taxonomic profiling. Nat Methods. 2015;12(10):902–3.

Uritskiy GV, DiRuggiero J, Taylor J. MetaWRAP-a flexible pipeline for genome-resolved metagenomic data analysis. Microbiome. 2018;6(1):158.

Vangay P, Hillmann BM, Knights D. Microbiome learning repo (ML Repo): a public repository of microbiome regression and classification tasks. Gigascience. 2019;8:5.

Venter JC, et al. The sequence of the human genome. Science. 2001;291(5507):1304–51.

Vezzi F, Cattonaro F, Policriti A. E-RGA: enhanced reference guided assembly of complex genomes. EMBnet J. 2011;17(1):46–54.

Voigt AY, et al. Temporal and technical variability of human gut metagenomes. Genome Biol. 2015;16:73.

Wang W, et al. Identifying differentially spliced genes from two groups of RNA-seq samples. Gene. 2013;518:1.

Wenger AM, et al. Accurate circular consensus long-read sequencing improves variant detection and assembly of a human genome. Nat Biotechnol. 2019;37(10):1155–62.

Wen-Ping H, et al. Kernel density weighted loess normalization improves the performance of detection within asymmetrical data. BMC Bioinformatics. 2011;12:1.

Wirbel J, et al. Meta-analysis of fecal metagenomes reveals global microbial signatures that are specific for colorectal cancer. Nat Med. 2019;25(4):679.

Wood DE, Salzberg SL. Kraken: ultrafast metagenomic sequence classification using exact alignments. Genome Biol. 2014;15(3):R46.

Wu S, et al. GMrepo: a database of curated and consistently annotated human gut metagenomes. Nucleic Acids Res. 2020;48(D1):D545–53.

Xi W, C.M. J. SeqGSEA: a Bioconductor package for gene set enrichment analysis of RNA-Seq data integrating differential expression and splicing. Bioinformatics. (Oxford, England). 2014;30:12.

Xiao L, Zhang F, Zhao F. Large-scale microbiome data integration enables robust biomarker identification. Nat Comput Sci. 2022;2(5):307–16.

Yarza P, et al. Uniting the classification of cultured and uncultured bacteria and archaea using 16S rRNA gene sequences. Nat Rev Microbiol. 2014;12(9):635–45.

Ye SH, et al. Benchmarking metagenomics tools for taxonomic classification. Cell. 2019;178(4): 779–94.

Yilmaz P, et al. Minimum information about a marker gene sequence (MIMARKS) and minimum information about any (x) sequence (MIxS) specifications. Nat Biotechnol. 2011;29(5):415–20.

Yinlong, X., et al., SOAPdenovo-trans: de novo transcriptome assembly with short RNA-Seq reads. Bioinformatics (Oxford, England), 2014. 30(12).

Yu, P., et al., IDBA-tran: a more robust de novo de Bruijn graph assembler for transcriptomes with uneven expression levels. Bioinformatics Oxford, England, 2013. 29(13).

Zhang T, et al. MPD: a pathogen genome and metagenome database. Database (Oxford). 2018;2018

Zhou Q, Su X, Ning K. Assessment of quality control approaches for metagenomic data analysis. Sci Rep. 2014;4:6957.

Zhou Q, et al. RNA-QC-chain: comprehensive and fast quality control for RNA-Seq data. BMC Genomics. 2018;19(1):144.

Zhou Z-H. Ensemble learning. Encyclopedia of Biometrics. 2009;1:270–3.

Concluding Remarks

The multi-omics studies have already proven themselves as a powerful approach for the in-depth understanding of biology systems, from a data-driven perspective. Based on these multi-omics studies, hundreds to thousands of studies have been conducted for both basic researches and precision medicine applications, deepen our understanding of the disease regulations patterns as well as dynamic models at multiple levels. In recent years, this trend has become more obvious, with more multi-omics studies conducted and more mature data integration and mining techniques, a broad spectrum of applications have been realized, helping for fast translation of molecular understanding of diseases to clinics. No doubt, that with the increasing type and amount of multi-omics data, we could see more success stories based on the analyses of these data.

The multi-omics studies will continue to grow, at least in two directions: first, from multi-omics for the single organism or single species, to single-cell level omics studies, as well as to population and community level studies (Fig. 1a); second, the tight integration of multi-omics with data science as well as with clinical applications (Fig. 1b).

From the aspect of multiple levels of omics data, we expect to see more types of multi-omics data, expanding the scope of multi-omics: from the single organism or single species, to single-cell level omics studies (Chappell et al. 2018), as well as to population and community level studies (Rahnavard et al. 2021). We have already seen rapid progress in this direction, largely due to the sequencing technical advances (Buermans and den Dunnen 2014). For example, single-cell level omics studies have already revealed the spatial-temporal patterns of how a biological system develop at single-cell level, especially during disease progression (Chappell et al. 2018). On the population and community level, population genetics have already revealed the differences between different ethnic groups (Huang et al. 2015), while recent microbiome studies have also examined the patterns that have never been discovered before on how microbes could assert such strong influences on human health (Sharma and Gilbert 2018). All of these expansions of multi-omics

K. Ning (ed.), *Methodologies of Multi-Omics Data Integration and Data Mining*, Translational Bioinformatics 19, https://doi.org/10.1007/978-981-19-8210-1

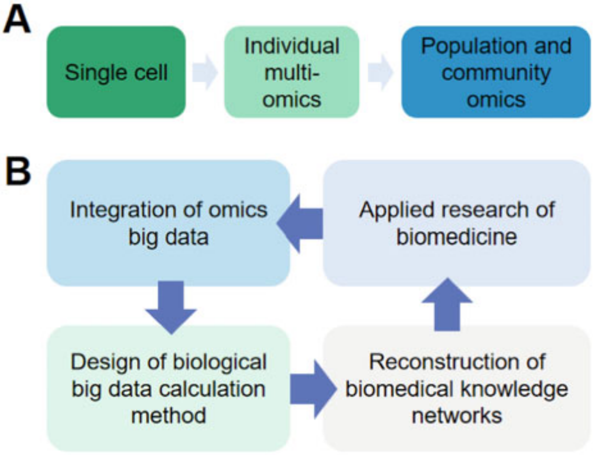
A
Single cell
Individual multi-omics
Population and community omics
B
Integration of omics big data
Applied research of biomedicine
Design of biological big data calculation method
Reconstruction of biomedical knowledge networks

Fig. 1 The two main directions of multi-omics research. (**a**) One direction is from single cell to individual species' multi-omics to population and community level omics. (**b**) Another direction is from integration of omics data to data analysis, to knowledge discovery, to applications

studies at different scales have pushed the multi-omics researches to be closer to a full picture of the dynamic regulation system in human.

From the aspect of the integration of multi-omics with data science as well as with clinical applications, there are very hard challenges still lying ahead. Firstly about "from bedside to bench", since more types of multi-omics data could be obtained, while how these multi-omics data at different levels correspond to each other remains illusive, so new sampling as well as sequencing techniques are needed for a solid multi-omics study. Secondly about data integration and data mining, new tools especially deep learning tools are urgently needed for mining of hidden patterns (Shi et al. 2019). Thirdly about "from bench to bedside", any analytical results, regardless of biomarker, dynamic pattern or functional gene, should be validated by wet-lab experiments, and the fast and reliable translation of multi-omics knowledge to clinical practices is critical (Plebani 2021), which might also need a shift of mindset: the combination of data-driven and clinical problem driven.

From the aspect of the biological computing, it is the trend to integrate advanced AI technology and cutting-edge biotechnology through long-term and large-scale, towards a new type of multi-omics detection and analysis, high-throughput experimental simulation, intelligent molecular discovery engine, and accelerating the development of new drugs and diagnostic products. There have been many studies in this area with good results, including biocomputing drives innovation and development of plant-based natural medicines (Thomford et al. 2018), simulating the biomolecules of interest in research (Giannakis et al. 2020), the design of biological therapeutics with silico methods (Roy et al. 2017). It would be desirable to connect laboratory instruments and computing systems to build a dry and wet closed-loop model, organically combine artificial intelligence models with wet experiments, and overcome the limitations of artificial intelligence models that do not have enough experimental data to modify and test parameters. For examples, at present, drug development of pharmaceutical companies is mainly based on human assumptions and existing experimental capabilities. In this way, the potential target space or pharmaceutical space that can be explored is greatly limited by the accumulation of existing research and development. With artificial intelligence methods, more complex data can be considered comprehensively, and higher-dimensional information can be observed. Based on this, it can take advantage of AI models and computing resources, combine self-produced experimental data and medical and pharmaceutical expertise to discover new drug targets. And a broad spectrum of applications could follow this scheme to be equipped with AI for more knowledge discovery.

Collectively, it is now a multi-omics age, and everyone should keep pace with it. It is based on multi-omics data analysis and interpretation that many of current clinical applications are possible, and there is no doubt that more clinical applications would be heavily dependent on multi-omics studies. In this book, we have described the needs for multi-omics, the basic database and tools for omics data analysis, as well as representative applications based on muti-omics studies. We hope the readers could gain basic idea about multi-omics, and could use the information provided in this book to help for designing and conducting their own project. And we believe that many pieces of information provided here could also help those who are working in area of systems biology to gain more idea to improve their studies.

Good luck!

References

Buermans HPJ, den Dunnen JT. Next generation sequencing technology: advances and applications. Biochim Biophys Acta (BBA) - Mol Basis Dis. 2014;1842(10):1932–41.

Chappell L, Russell AJC, Voet T. Single-cell (multi)omics technologies. Annu Rev Genomics Hum Genet. 2018;19(1):15–41.

Giannakis K, et al. Particular biomolecular processes as computing paradigms. Cham: Springer International Publishing; 2020.

Huang T, Shu Y, Cai Y-D. Genetic differences among ethnic groups. BMC Genomics. 2015;16(1): 1093.

Plebani M. Chapter 3—"Omics" translation: a challenge for laboratory medicine. In: Wehling M, editor. Principles of translational science in medicine (third edition). Boston: Academic; 2021. p. 21–32.

Rahnavard A, et al. Omics community detection using multi-resolution clustering. Bioinformatics. 2021;37(20):3588–94.

Roy A, et al. In silico methods for design of biological therapeutics. Methods. 2017;131:33–65.

Sharma A, Gilbert JA. Microbial exposure and human health. Curr Opin Microbiol. 2018;44:79–87.

Shi Q, et al. Deep learning for mining protein data. Brief Bioinform. 2019;22(1):194–218.

Thomford NE, et al. Natural products for drug discovery in the 21st century: innovations for novel drug discovery. Int J Mol Sci. 2018;19(6):1578.

K. Ning (ed.), *Methodologies of Multi-Omics Data Integration and Data Mining*,
Translational Bioinformatics 19, https://doi.org/10.1007/978-981-19-8210-1