Computational and Statistical methods to analyze multi-omic data

The advent of high-throughput technologies has enabled the simultaneous measurement of multiple types of omics data, such as genomics, transcriptomics, proteomics, and metabolomics. This multi-omic approach has revolutionized biological research by providing a comprehensive view of complex biological systems. However, the analysis of such data poses significant challenges due to its high dimensionality, heterogeneity, and interdependencies. In this context, computational and statistical methods play a vital role in extracting meaningful insights from multi-omic data. By incorporating this idea, we intend to address some key questions in this thesis, including (i) interactions and influences between different omics layers, (ii) identification of potential therapeutic targets for guiding drug development efforts, (iii) the role of cellular heterogeneity in disease programs, and (iv) mechanisms of cell communication within tissues or across the cellular system. While various methods are available, these questions still require appropriate answers. The traditional reductionist approaches fail to establish the bridge between biological significance and designed computational methods. This observation motivates us to design novel bioinformatics frameworks by using computational and statistical techniques to decipher unanswered questions regarding multi-omics data.

The first chapter of this thesis begins with a general introduction of the multi-omics data and provides background information on each omics and establishes the significance of the research. It involves an exploration of existing knowledge gaps, practical implications, and potential theoretical contributions. Additionally, the chapter delineates the scope of the studies, encompassing any limitations or constraints that may influence the findings' applicability. Furthermore, an overview of the research questions stated earlier is provided, elucidating how they are addressed, along with an outline of the methodology employed in each study. Finally, the chapter concludes by providing a summary of the thesis' overall structure, offering a brief preview of the subsequent chapters' content.

In the subsequent chapter, the main objective is to answer the interaction among the different omics data and their impact on each other. The existing approaches of regulatory networks play a pivotal role in unraveling the complex mechanisms underlying various biological processes and diseases. However, traditional methods fail to achieve a system-level understanding of gene regulation and association with other regulators. In this regard, we proposed a pipeline by incorporating network biology concepts to identify the regulators responsible for disease-specific gene dysregulation. This proposed method helps to decipher the relationships and also the key regulatory genes and pathways that play critical roles in disease development or involves in normal physiological processes.

Chapter 3 proposes models for the identification of potential therapeutic targets to guide drug development efforts. In the preceding chapter, transcription factors (TFs) are identified as promising drug targets due to their ability to modulate the expression of protein-coding genes and miRNAs at the transcriptional and post-transcriptional levels, respectively. However, targeting TFs for drug development is challenging due to their highly dynamic nature, which makes it difficult to identify stable drug-binding sites. To address this challenge, a novel concept called protein moonlighting is employed to determine suitable druggable sites on TFs. Additionally, the structure-function paradigm of proteins is utilized to gain valuable insights into their dynamic properties, which can guide the development of drugs and optimize their binding efficacy. To explore the dynamicity of TFs, an evolutionary study is conducted on the structure-function space of these proteins. Two entropy models, based on Shannon entropy and direct coupling analysis, are proposed to analyze the information content and coevolutionary patterns within the TFs. Moreover, centrality-based clustering techniques are employed to establish the structural network of the TFs. This approach helps identify interconnected regions within the network, known as the structural core. These structural cores represent specific protein conformations or dynamic properties that drugs can target to achieve optimal binding and therapeutic efficacy. By incorporating these models, this chapter provides a comprehensive framework for identifying potential drug targets within the highly dynamic landscape of TFs.

The studies discussed in the previous two chapters of this thesis are performed on bulk sequencing data. However, this sequencing technique does not capture the cellular heterogeneity within the population, potentially masking important subpopulations of cells with distinct gene expression profiles. In contrast, single-cell RNA-sequencing (scRNA-seq) enables the profiling of gene expression at the single-cell level, providing a comprehensive view of cellular heterogeneity within a tissue or sample. During COVID-19, it is imperative for researchers to contribute scientifically towards controlling the spread of the virus. The aim of our study is to investigate the impact of COVID-19 on specific organ cell types and elucidate the associated cell-specific pathway cascades. Through comprehensive case studies, it is found that some specific organs such as lungs, kidneys, liver, ileum, and bladder are affected due to COVID-19. Additionally, it was discovered that the virus exhibits a strong binding affinity with ACE2 and TMPRSS2 proteins. To gain insights into the organ-specific cell types and their differentially expressed biomarkers, clustering techniques were employed. Moreover, protein-protein interaction (PPI) analyses were performed to identify highly influential hubs within these networks, utilizing the K-means clustering algorithm. Furthermore, a pathway semantic similarity network was established to uncover the interactions between pathways responsible for regulating cell function. In another study, the maximal clique centrality (MCC) technique was incorporated to identify essential genes from local interaction networks. By identifying these essential genes, their critical roles in cellular processes and potential as therapeutic targets can be elucidated. Overall, the proposed

bioinformatics methods have the potential to contribute to the development of targeted therapeutic strategies and enhance our understanding of the disease.

Many diseases like cancer, neurodegenerative disorders, and immune-related conditions involve dysregulated cell communication and crosstalk. Simply identifying biomarkers associated with a disease may provide diagnostic or prognostic information, but it does not capture the underlying mechanisms driving the disease. In this regard, Chapter 5 proposed a new method for computing pathway activity scores of the cells to decipher the cell-to-cell communication in a particular tissue. The output of the proposed method provides information regarding the function of genes performed in biological processes. Finally, by incorporating pathway information in our proposed single-cell clustering method, we are able to align pathways across the three gynecological diseases and successfully identified disease-specific mechanisms.

As demonstrated in this thesis, the application of statistical and computational techniques has made significant contributions to addressing key problems in biology and holds great promise for future research. These approaches have proven to be valuable tools for analyzing complex biological data, uncovering hidden patterns, and gaining insights into fundamental biological processes.