CHAPTER 7

Discussion and future perspectives
This thesis has contributed to a better understanding of the genetic determinants of complex human traits (e.g., infectious diseases) using computational biology approaches of integrating multiple molecular phenotypic measurements. Computational biology approaches exploit quantitative measurements, computational models, and high-throughput screening approaches to decipher the relationships among molecular components within a biological system, such as cells and tissues [1].

Thanks to improvements in next-generation sequencing (NGS) technologies, we can now profile various molecular traits in patients and healthy individuals cost-effectively at a large scale [2]. These advances have brought biology and medical researches into the era of multiple omics (multi-omics), including genomics, epigenomics, transcriptomics, proteomics, metabolomics, and microbiomics. Of note, the integration of multi-omics data bridges the gap between genotypes and phenotypes [3]. For instance, by combining multi-omics data from one individual who was followed for seven years, researchers revealed molecular component and pathway dynamics between healthy and disease conditions [4], highlighting the value of longitudinal multi-omics integration.

Omics studies started with genomics, in which the whole genome is examined to understand the structure, function, and evolution of the genome [5, 6, 7]. By combining genomics with other omics, such as transcriptomics and epigenomics, we can estimate allelic imbalances in order to study the regulatory networks involving human phenotypes [8]. This led me to study genetic regulatory relationships in health and diseases by identifying the effect of allelic imbalance, including allele-specific expression (ASE) and allele-specific open chromatin (ASoC), in Chapter 2, Chapter 3, and Chapter 6.

Advances in omics studies also offer opportunities to identify and study phenotype- or disease-associated genetic variants using methods like genome-wide association studies (GWAS). Over the past several decades, GWAS has blossomed and generated a large set of genetic associations with many phenotypes and diseases [9]. However, it is not yet possible to interpret most of these findings due to our limited biological knowledge, for instance, to understand the regulatory function of non-coding variants identified by GWAS [10, 11]. Therefore, dissecting the molecular functions of GWAS variants is important in the coming years of the “post-GWAS” era. My work in this direction is presented in Chapter 4 and Chapter 5.

Additionally, integrating multi-omics data empowers researchers to answer one of the key questions in biology studies: How does information flow in complex biological systems? The integration is biologically informative because it represents the flow of biological signals underlying the phenotypes of interest or disease conditions [1]. Moreover, single-cell sequencing methods deepen our biological knowledge by capturing the broad spectrum of diversity of cell characteristics and functions [12, 13]. Based on findings in Chapter 6, I will discuss how (single-cell) multi-omics integration is helpful in extending our understanding of human infectious diseases continuously.

In the current thesis, I explored genetic determinants of human molecular traits - allelic imbalance and host responses to pathogenic viruses - using computational biology approaches. In this final chapter, I discussed our conclusions and future directions from the perspective of computational biology, based on results from the current thesis and other recent studies elsewhere using multi-omics data.

Discussion

ASE and genetic variants

As diploid organisms, humans derive genetic and epigenetic characteristics from their parents, which introduces allelic differences between the two sets of haplotypes. These differences can result in distinct behaviors between alleles, for instance, ASE. One of the well-known mechanisms underlying ASE is epigenetic regulation, where allelic openness of the chromatin plays crucial roles, such as genomic imprinting [14] and X-chromosome inactivation [15]. Of note, researchers have revealed that genetic loci (e.g., disease-associated variants identified by GWAS) usually affect allelic states of chromatins [16, 17, 18] and disrupt the sequence of transcription factor binding sites (TFBSs) [19], which consequently modulates the phenotypes through the formed regulatory networks.

Apart from the canonical mechanisms mentioned above, we hypothesize that the ASE of a single nucleotide variant (SNV) can be explained by its genomic context, i.e., the genomic features around the SNV. To test this hypothesis, in Chapter 2, we built a gradient boosting tree model to predict the ASE of exonic SNVs using the genomic annotations of the SNV. The model performs fairly well (testing ROC AUC 0.8), suggesting the ASE of a SNV is related to its genomic annotations. For instance, we observed that low GerpN values [20], which have high feature importance influencing the model outcomes, drove the model outcomes for ASE. This is expected because GerpN values close to 0 represent high heterology at the locus.
However, our model has two main limitations. First of all, we built the model exclusively using variants from exonic regions. Consequently, it has no capacity to evaluate regulation from non-coding variants in intergenic and intronic sequences or the ASE effects of a transcript. Secondly, the model was trained using only whole blood samples, in which only a subset of genes was under active expression. However, ASE is tissue-specific and cell-type-specific [21], thus, including tissue-specific and single-cell sequencing results will further improve the performance of ASE predictors. The Genotype-Tissue Expression (GTEx) project [22] offers us opportunities to estimate the transcriptional ASE and identify non-coding variants that regulate the allelic imbalance.

To overcome the mentioned limitations above, we next developed a new method, ASDeep, in Chapter 3. ASDeep is a deep learning (DL) tool that predicts the ASE for the canonical transcript of a gene using the DNA sequence that potentially regulates the expression of the studied gene. Recent studies have revealed the ability of DL approaches to discover the functions of non-coding DNA sequences [23, 24, 25, 26], and this encouraged us to construct DL models to predict ASE using the DNA sequence.

However, the model does not perform as expected due to many limitations. In concrete, first of all, only 3500 samples were available which resulted in a high risk of over-fitting. Next, these samples were grouped into three classes (ASE prone to the paternal allele, non-ASE, and ASE prone to the maternal alleles) but the number of samples from each class was highly unbalanced. Lastly, due to the limitation of imputation algorithms for genotypes, the DNA sequences used for the model training resembled each other, which hampered the model to discriminate the identified classes.

In addition, both of our models considered only DNA sequences per se, neglecting other factors that affect gene expressions, such as DNA methylation and histone modifications in the context of epigenetics [27].

Host (epi)genetic determinants of infectious diseases

The human genomic architecture determines the distinct immune responses to microbes among individuals [28, 29]. Specifically, lethal infections caused by pathogenic viruses usually trigger strong immune responses during disease progressions. Therefore, studying the genetic determinants of human infectious diseases will provide us insights into the underlying mechanisms of microbe-human interactions and ultimately guide prognosis and treatment strategies. In addition, human immune responses to viruses are also under active epigenetic regulation, resulting from microbe-human interactions [30]. These facts led us to investigate the genetic and epigenetic factors that affect the manifestation of infections by pathogenic viruses, including HIV-1 and coronavirus SARS-CoV-2.

One of the obstacles to curing people living with HIV-1 (PLHIV) is the latent HIV-1 reservoir, i.e., immune cells in the body that are infected with HIV but not actively producing new HIV [31]. The combination of antiretroviral treatment (cART) substantially increases the life quality of PLHIV by suppressing HIV-1 viral loads in the human body. Recent studies have also revealed viral and host characteristics that correlate to the HIV-1 reservoir size and long-term dynamics in PLHIV [32, 33, 34, 35, 36, 37]. In Chapter 4, we identified PTDSS2, IRF7, RNH1, and DEAF1 as potential modifying factors of HIV-1 reservoirs by associating host genetic factors with HIV-1 reservoir traits, represented by cell-associated (CA) HIV-1 DNA, CA HIV-1 RNA, and their ratio (RNA:DNA).

Compared to previous studies [36, 37], we focused on the CA HIV-1 DNA and RNA isolated from CD4+ T cells, a choice that avoids the influence of cell proportion. Of note, our studies suggested that IRF7 plays a potential role to regulate HIV-1 latency and recent studies revealed that IRF7 mediated the HIV suppression resulting from the knockout of CNOT (the RNA deadenylase complex) [38]. In addition, there is evidence for IRF7 being involved in viral latency initialization and reactivation from a study comparing chronic gammaherpesvirus infections in If7-deficient and wild-type mice [39]. However, our study had several limitations: 1) the sample size of the discovery cohort was relatively small (200); 2) limited covariates were included in the association analysis even though many viral and host characteristics are associated with the HIV-1 reservoir traits; 3) our results were limited to Caucasian people; and 4) our findings would be further replicated using an independent cohort, which is currently unavailable.

Individuals infected by SARS-CoV-2 display a wide range of clinical manifestations [40, 41, 42, 43] and these inter-individual variations could be due to the differences in (epi)genetic architecture among and across the populations [44, 45, 46, 47, 48]. This encouraged us to evaluate the (epi)genetic regulations in COVID-19 patients under different disease conditions.

In Chapter 6, our single-cell multi-omics integration results revealed transcriptomic and epigenomic (chromatin accessibility) profiles among COVID-19 conditions. The scRNA-seq results showed a number of cell proportion changes by the pair-wised comparisons among the COVID-19 disease conditions, which agrees with previous
studies [49, 50, 51]. In addition, based on our scATAC-seq results, we observed that C/EBPs, JUN/FOS, and SPI1 motifs were overrepresented in classical monocytes the patients. Additionally, our peak-to-gene linkage analysis suggested the long non-coding RNA, LUCAT1, was modulated by both SPI1 and C/EBPs transcription factors (TFs). These results led us to hypothesize a dysregulated cascade, in which increased C/EBP regulation enhanced LUCAT1 expression and further suppressed interferon responses to SARS-CoV-2 infection, leading to severe COVID-19.

We also identified ASoC SNPs and found that they were enriched in regulatory regions (i.e., promoters, enhancers and TFBSs). Further, our results revealed the potential role of CCR2 in COVID-19 in the context of genetics and epigenetics. CCR2 encodes the chemokine receptor for monocyte chemoattractant protein-1 (MCP-1/CCL2), which promotes the migration of monocytes to sites of inflammation [52]. The overrepresentation of MCP-1/CCL2 in bronchoalveolar fluid in severe COVID-19 patients suggests inflammation in lung and active recruitment of CCR2+ monocytes [53].

However, the mentioned study above has several limitations. First of all, only limited symptom categories were included in the study, namely severe, mild and convalescent individuals whereas, COVID-19 patients have a wide symptomatic spectrum ranging from asymptome to deadly severe [40, 41, 42, 43]. Next, we only included whole blood samples which could be less representative compared to primary lung tissues, such as epithelium [54]. In addition, limited by the sample size and sequencing depth, our study only validated a few COVID-19 GWAS loci, namely CCR2 and DPP9 [47]. Last but not least, our analyses exclusively focused on genetic, epigenetic, and transcriptomic aspects of the diseases at single-cell resolution, while proteomics, metabolomics, and metagenomics [55, 56, 57] analyses revealed prognosis and diagnosis markers. Therefore, we expect further evidence from other omics data to gain more insight into the disease mechanisms and to validate our results in the future.

**Cell-type-specific quantitative trait loci (QTL) of the cell-surface CCR5 abundance**

The CCR5 locus is a well-established association between host genetic architectures and HIV-1 infections, where CCR5d32 exists in HIV-1 controllers [58, 59]. This genetic locus has been consistently associated with primary resistance to HIV-1 diseases, which has boosted AIDS treatment approaches such as transplantation of CCR5 gene-edited CD4+ T cells in which CCR5 genes are made permanently dysfunctional [60, 61]. Apart from CCR5d32, recent studies have suggested that additional genetic loci affect the expression of CCR5 [62, 63]. In addition, beyond CD4+ T cells, other immune cells such as CD8+ T cells and dendritic cells are also potentially involved in the response to the HIV infection [64, 65, 66]. However, less is known about whether specific genetic loci differentially regulate cell-surface CCR5 abundance in subsets of immune cells (e.g., lymphocytes) from PLHIV. We estimated the cell-type-specific associations between genetic variants (SNPs) and CCR5 levels in lymphocyte subsets in Chapter 5, where we used the mean fluorescence intensity (MFI) of cell-surface CCR5 or CCR5+ cell proportion (CP) to represent CCR5 abundance. Our results revealed genetic loci associated with the cell-surface abundance of CCR5 in 11 lymphocyte subsets identified from PLHIV.

Specifically, we found QTLs associated with both MFI and CP in CD4+ memory regulatory T (mTreg) cells, which suggests that the CCR5 abundance in mTregs is regulated by genetic factors. On the one hand, the mTregs derive from pathogen-specific Tregs that are activated and expand upon acute viral infections in vivo [67, 68]. In response to secondary infections and inflammations, mTregs supersede naive Tregs (nTregs) in migration to tissues and activation toward immunosuppression [67], suggesting that they can be the inducing target cell populations for vaccinations [68]. On the other hand, our results identified independent genetic loci associated with MFI and CP in mTregs. The leading SNP (rs11574435) for MFI is a proxy for the CCR5d32, while the leading SNP (rs60939770) for CP is a proxy for SNP rs1015164, which plays a role in protecting the CCR5 mRNA by affecting the expression of CCR5AS [62]. These results suggest different genetic architectures between cell-surface CCR5 intensities and proportions of CCR5+ proportions in mTregs. However, in-depth work is required to further dissect the underlying regulatory programs in HIV-1 infections.

However, several limitations hampered us to understand the hosts’ genetic factors impacting CCR5 surface expression in subpopulations of immune cells from PLHIV. First of all, only a limited number (200) of PLHIV was included in the association analysis, which lacks the power to detect loci of small effect size, expecting a larger cohort in the near future. Next, the other major limitation was that no other independent PLHIV cohort was included for the purpose of replication, which is expected to be solved by collaboration with other institutions and efforts to recruit more PLHIV participants in the coming years. Moreover, the use of a single monoclonal anti-CCR5 antibody might underestimate CCR5 that was not recognized by the antibody. In a study that evaluated the antiviral activity of different CCR5 antibodies, the monoclonal antibody 2D7, which we also used in
After obtaining GWAS summary statistics, the next step is to determine the causal variants with high confidence. The aim of GWAS is to identify the genetic determinants for human traits. However, the majority of identified associations are located in non-coding regions, which frequently contribute to the phenotypes by regulating gene expressions [70, 10]. To address this problem, allele-specific analysis, including ASoC, ASM (allele-specific methylation), ASTFB (allele-specific TF binding), and ASE, can be powerful tools to discover genetic variants and to underpin epigenetic markers depicting regulatory networks in disease scenarios. In diploid or polyploid organisms, the allelic analysis estimates differences between alleles at a genetic variant (e.g SNP). The difference includes chromatin accessibility, methylation state, TF binding affinity, mRNA level, and protein abundance. At the transcriptional level, genetic variants result in ASE by (i) modifying epigenetic markers, including histone modification and DNA methylation [16, 71], and (ii) affecting the allelic affinities for the transcription factor binding at the locus [8]. In Chapter 3 and Chapter 4, we studied the relationship between genetic variants and ASE that commonly exists in the human body. Our findings demonstrate that ASE is under active genetic and epigenetic regulations. They also provide chances to trace the causal genetic variants by modeling ASE and other allelic differences.

After obtaining GWAS summary statistics, the next step is to determine the causal variants for the studied traits. However, this process can be challenging because of the complex linkage disequilibrium (LD) patterns between causal variants and their proxies [72]. Therefore, many statistical methods have been developed to identify causal variants, the so-called fine-mapping processes. Fine-mapping can identify one or more causal variants for the traits or diseases and quantify the supporting evidence for the casualty. The resulting set of high-confident causal variants is further functionally analyzed to prioritize candidate genes based on publicly available genomic annotations. However, these genomic annotations may be less informative or incomplete, and follow-up for confirmation may require expensive and time-consuming laboratory-based functional analyses. Hence, further evidence, such as allelic analysis of multi-omics studies, is required to prioritize causal variants with high confidence.

Genetic variants identified by GWAS are significantly enriched in non-coding regions [70, 10]. These regions usually contain regulatory sequences (e.g. enhancers and suppressors) and potentially affect the studied traits or diseases by altering the expression of distal or proximal genes. For example, non-coding genetic variants can modulate gene expression by perturbing chromatin accessibility, DNA methylation, and TF binding affinity, which results in ASE [8]. Researchers can validate causal variants and dissect the underlying regulatory network by analyzing allelic differences in integrated multi-omics data. In Chapter 6, we identified ASoC SNPs using scATAC-seq results from COVID-19 patients. By combining with other publicly available omics and GWAS results, we prioritize CCR2 as a candidate gene in modulating the severity of COVID-19. Our findings offered an example in which allele-specific analysis validated GWAS associations in the context of human infectious diseases.

**Future perspectives**

**Allelic-specific analyses prioritize causal variants for GWAS**

The aim of GWAS is to identify the genetic determinants for human traits. However, the majority of identified associations are located in non-coding regions, which frequently contribute to the phenotypes by regulating gene expressions [70, 10]. To address this problem, allele-specific analysis, including ASoC, ASM (allele-specific methylation), ASTFB (allele-specific TF binding), and ASE, can be powerful tools to discover genetic variants and to underpin epigenetic markers depicting regulatory networks in disease scenarios. In diploid or polyploid organisms, the allelic analysis estimates differences between alleles at a genetic variant (e.g SNP). The difference includes chromatin accessibility, methylation state, TF binding affinity, mRNA level, and protein abundance. At the transcriptional level, genetic variants result in ASE by (i) modifying epigenetic markers, including histone modification and DNA methylation [16, 71], and (ii) affecting the allelic affinities for the transcription factor binding at the locus [8]. In Chapter 3 and Chapter 4, we studied the relationship between genetic variants and ASE that commonly exists in the human body. Our findings demonstrate that ASE is under active genetic and epigenetic regulations. They also provide chances to trace the causal genetic variants by modeling ASE and other allelic differences.

After obtaining GWAS summary statistics, the next step is to determine the causal variants for the studied traits. However, this process can be challenging because of the complex linkage disequilibrium (LD) patterns between causal variants and their proxies [72]. Therefore, many statistical methods have been developed to identify causal variants, the so-called fine-mapping processes. Fine-mapping can identify one or more causal variants for the traits or diseases and quantify the supporting evidence for the casualty. The resulting set of high-confident causal variants is further functionally analyzed to prioritize candidate genes based on publicly available genomic annotations. However, these genomic annotations may be less informative or incomplete, and follow-up for confirmation may require expensive and time-consuming laboratory-based functional analyses. Hence, further evidence, such as allelic analysis of multi-omics studies, is required to prioritize causal variants with high confidence.

Genetic variants identified by GWAS are significantly enriched in non-coding regions [70, 10]. These regions usually contain regulatory sequences (e.g. enhancers and suppressors) and potentially affect the studied traits or diseases by altering the expression of distal or proximal genes. For example, non-coding genetic variants can modulate gene expression by perturbing chromatin accessibility, DNA methylation, and TF binding affinity, which results in ASE [8]. Researchers can validate causal variants and dissect the underlying regulatory network by analyzing allelic differences in integrated multi-omics data. In Chapter 6, we identified ASoC SNPs using scATAC-seq results from COVID-19 patients. By combining with other publicly available omics and GWAS results, we prioritize CCR2 as a candidate gene in modulating the severity of COVID-19. Our findings offered an example in which allele-specific analysis validated GWAS associations in the context of human infectious diseases.

**Multi-omics integration to decipher complex information flow in biological systems**

Multi-omics integration supplies a tool to understand complex biological systems by comprehensively considering the information flow from different perspectives within the system. A biological system modulates its information flow depending on its architecture which can be profiled by multiple molecular traits generated by different omics approaches [3]. Hence, multi-omics data integration enables researchers to discover the driving factors for complex phenotypes like human infectious diseases because this integration combines several snapshots of the biological information flow (BIF) from different perspectives. Additionally, single-cell multi-omics provide another layer of information that exhibits cell-type-specific characteristics and functions [73].

According to the ways of depicting the information flow, integrative strategies can be grouped into two classes: multi-stage and multi-dimensional [74]. The multi-stage approaches are often performed into two steps: correlation between two types of omics data followed by validation of the identified correlation. For instance, researchers can underpin and validate causal genetic variants for a disease by combining multiple types of (publicly available) QTLs. Obviously, the multi-stage method hypothesizes a relatively straightforward approach in which the BIF propagates linearly or hierarchically, even though we know BIF could be a more complicated network. Consequently, the multi-stage strategy may overlook latent BIFs that play crucial roles in modulating phenotypes. In contrast, the multi-dimensional strategy hypothesizes that the BIF moves along the edge between
two nodes that present the omics profiles of the phenotype. Multi-dimensional approaches require more advanced statistical frameworks to fuse the information, capture the latent BIF, and abstract the key driving factors for the phenotypes or diseases. Frameworks like machine learning/deep learning (ML/DL) methods are often designed to simulate complex relationships among multiple layers of data and are, thus, promising for the integration of multi-omics data in a multi-dimensional context.

Genetic and epigenetic regulation, which are massive but methodical, result in the cellular heterogeneity of a biological system (e.g., the human body) [75, 76]. Genetically, every cell in the human body carries an identical copy of heritable information from the same zygote, at least as long as no somatic mutations happen during the body’s growth and development. However, after division and differentiation, these cells can display diverse characteristics and distinct functions, which is due to the time- and space-specific regulation. These divergences or heterogeneities bring about inter-cell-type variability in response to pathogenic microbes like viruses [77, 78]. A subset of crucial cells could play essential roles in defending the human body [78]. Therefore, studying the regulation programs in a specific cell type in the scenario of infectious diseases can demonstrate how these cells handle external stimulations and invasions. This knowledge may ultimately improve medical practices by providing such as personalized medicine and vaccination methods.

Single-cell methods zoom into the cell-type-specific functions and characteristics. The integration of single-cell multi-omics data is becoming popular in routine biology and medical studies as it bridges various biological features at individual-cell resolution. Single-cell multi-omics methods break the barriers of cell heterogeneity by functioning at a higher resolution than traditional single-cell methods (e.g., cell imaging and flow cytometry). Chapter 6 shows how our single-cell multi-omics results reveal the epigenetic and genetic regulatory mechanisms underlying the severity of COVID-19 diseases. These findings prove the value of applying single-cell multi-omics integration in infectious diseases studies and these approaches are also applicable in other biological and medical studies.

ML/DL promotes computational biology

The successful interpretation of integrative multi-omics data depends on attentive and close cooperation by experts or specialists from different disciplines, which is considered to be the bottleneck of computational biology studies in the multi-omics era. Many statistical methods were developed to combine multi-omics data efficiently and consistently in order to break the bottleneck. ML/DL can mine information from the data of multiple sources by building a model that acquires experience and knowledge from the data during the mining. Subsequently, after its performance meets predefined criteria, the model makes predictions or decisions for forthcoming data. The application of ML/DL methods is spreading widely in biology and medical research thanks to improvements in computational and NGS technologies [79]. These recent ML/DL applications can be roughly grouped into (i) analyzing nucleotide/protein sequence [80, 81], (ii) predicting the properties, functions, and interactions of proteins and RNAs [82], (iii) processing biomedical images [83], and (iv) constructing causal networks for diseases [84]. These results suggest that a well-tuned ML/DL model can perform as professionally as experts and specialists from a specific field.

However, due to the complexity and diversity of parameters, ML/DL models are deemed black boxes because of their lack of interpretability compared to simple models like linear/logistic regression, and this is one of the main barriers to the use of ML/DL approaches in medical and clinical applications [85, 86]. In fact, ML/DL models can be interpreted and explained by dissecting the model, e.g., visualizing the information flow and customizing the model based on prior knowledge [87, 84, 88]. These feature-attribution methods provide convenient and robust tools to recognize the data patterns that drive the outcome [87]. For instance, researchers can identify genetic variants that affect the binding affinity in TF motifs based on the feature-attribution methods [89]. The other way to achieve high interpretability for ML/DL models is to use custom model architecture using prior knowledge, which heavily depends on expertises. For instance, in a recent study, a biologically informed model (BIM) was constructed according to the hierarchical structure of biological pathways [84]. In the model, each neuron node in every hidden layer is biologically meaningful, for instance, a gene or biological processing. The well-informed design of the model enables researchers to visualize the BIF in the network and prioritize neuron nodes based on their ability to drive the prediction outcomes. These examples open up the field of interpretable ML/DL models in biological and medical studies, which is expected to benefit disease treatment and prevention by dissecting the biological systems (i.e., the human body).

Concluding remarks

In the current thesis, I explored the genetic determinants of human phenotypes from allelic expression and infectious diseases. These studies inspire and encourage me to
interpret disease associated genetic variants, especially non-coding variants through application of ML/DL integration of multi-omics data in my future research.

Reference


