CHAPTER

Multi-omics approaches in immunological research

Frontiers in Immunology 12 (2021)
https://doi.org/10.3389/fimmu.2021.668045
Multi-omics approaches in immunological research

Xiaojing Chu 1,2,3†, Bowen Zhang 2,3†, Valerie Koeken 2,3,4, Manoj Kumar Gupta 2,3, Yang Li 1,2,3,4*

1 Department of Genetics, University of Groningen, University Medical Center Groningen, 9700 RB Groningen, the Netherlands
2 Department of Computational Biology for Individualised Medicine, Centre for Individualised Infection Medicine, CiiM, a joint venture between the Hannover Medical School and the Helmholtz Centre for Infection Research, Hannover, Germany
3 TWINCORE, Centre for Experimental and Clinical Infection Research, a joint venture between the Hannover Medical School and the Helmholtz Centre for Infection Research, Hannover, Germany
4 Department of Internal Medicine and Radboud Center for Infectious Diseases, Radboud University Medical Center, 6525 HP Nijmegen, the Netherlands

†These authors have contributed equally to this work.

* Correspondence:
Xiaojing Chu
Tsuki1128@outlook.com
Yang Li
Yang.Li@helmholtz-hzi.de

Abstract

The immune system plays a vital role in health and disease and is regulated through a complex interactive network of many different immune cells and mediators. To understand the complexity of the immune system, we propose applying a multi-omics approach in immunological research. This review provides a complete overview of current methodological approaches for the different omics data layers relevant for immunological research, including genetics, epigenetics, transcriptomics, proteomics, metabolomics, and celloomics. Thereafter, we describe the various methods for data analysis, as well as how to integrate different layers of omics data. Finally, we discuss the possible applications of multi-omics studies and the opportunities they provide for understanding the complex regulatory networks and immune variation in various immune-related diseases.

Keywords
Multi-omics; systems immunology; integrative analysis; immune-related diseases; immune variation.

Introduction

Infections cause millions of deaths each year, with the current COVID-19 pandemic underlining the devastating effects of these communicable diseases. At the same time, the incidences of immune-related diseases such as atherosclerosis (1) and autoimmune diseases such as type 1 diabetes mellitus (2) have been increasing. All these diseases are related to or mediated by the immune system. Thus, the immune system plays a vital role in health and disease, and it is our defense mechanism against harmful substances, infectious diseases and cancer. Within a properly functioning immune system, immune responses should be kept in a certain range, as both hypo-activation and hyper-activation lead to disorders of the immune system.

Understanding how the immune system works and what causes the immune system disorders may help us to efficiently fight immune-related diseases.

However, getting a comprehensive understanding of the immune system is a challenging task. First of all, the immune response is mediated through a complex interactive network of many different immune cells and molecules such as cytokines, immunoglobulins, and metabolites. At the same time, this network is highly variable depending on the exact threat posed by the wide variety of pathogens and other substances to which it can respond. Making things even more complex, the immune response to a specific stimulus or infection is highly variable between individuals, leading to population heterogeneity. This heterogeneity is exemplified by differences in severity of symptoms in patients suffering from the same infectious disease (3), variability in vaccine efficacy (4), and variations in responses to the same medical treatment (5). Many factors contribute to the immune network and the inter-individual variation of immune responses, highlighting both the promise and the challenge of multi-omics studies.

Until now, omics data have been used in many immunological studies to identify the determinants of immune variation and the molecular basis of the immune process in different population groups. Properly designed omics studies should make use of appropriate measurements as well as reasonable analytic approaches, which depend on the specific research question. Taking omics studies on COVID-19 as an example, a genome-wide association study revealed eight genetic regions to be associated with critical illness in COVID-19. By integrating both genome and transcriptome data, the authors prioritized one gene, IFNAR2, that might play a causal role in COVID-19 (6).

Another study focusing on transcriptome data of immune cells from the lung and blood identified several pro-inflammatory immune pathways related to the pathogenesis of COVID-19 (7). A proteomics and metabolomics study investigated the changes in COVID-19 patient sera and identified molecular changes implicating dysregulation in macrophage pathways, complement activation, and platelet degranulation, as well as suppression of metabolic pathways (8). A celloomics and single-cell transcriptome study also revealed dysregulation of the monocyte compartment, as well as two neutrophils clusters, that is specific to severe COVID-19 patients (9). Moreover, a study integrating single-cell transcriptome, celloomics, epigenome, and proteome comprehensively characterized complex dynamic changes in immune cells. Their results disclose an elevation of IFN-activated megakaryocytes and erythroid cells, hypomethylation around immune-signaling genes, and co-expression modules...
associated with clinical outcome (10). Additionally, a study on fecal fungal microbiota of COVID-19 patients showed enrichment of *Candida albicans* and a highly heterogeneous mycobiome configuration during hospitalization (11). All these studies make use of omics data from different angles to provide insights into the molecular pathology of COVID-19, which can eventually lead to improved therapeutic strategies.

Figure 1. Overview of omics data

In this review, we present an introduction to the multi-omics approaches used to investigate immune function and variation. The review is split into three parts. In the first section, we describe in brief the different layers of omics data relevant for immunological research, including genome, epigenome, transcriptome, proteome, metabolome, digestive system microbiota, and cellomics (12) (also called cytomics (13)) (Figure 1), and the methodological approaches commonly used to measure these different types of omics data. We also discuss important considerations and recommendations for appropriate study design. In the second section, we discuss how to analyze and integrate multiple omics platforms, including system genetic approaches to identify genetic factors, integration among multiple genetic profiles, and integration and association with other omics data layers. We demonstrate how recent studies applied a multi-omics approach to immune system research and discuss the interpretation of results from different approaches and their importance in immunological studies. In the third section, we discuss the immunological subjects that need specific attention and that may see progress in the next few years. For detailed information on computational algorithms and models in multi-omics integration (14), imputation of missing omics data (15), and the strengths and limitations of system approaches in infectious diseases research (16), we refer readers to other recent reviews.

**Measurements of omics data**

We can identify potential immunological mediators and study immune phenotypes using a wide range of omics comprising of various molecular and cellular phenotypes, including genome, epigenome, transcriptome, proteome, metabolome, digestive system microbiota and cellular phenotypes such as cell composition (Table 1). A single omics data layer characterizes a specific biological process from one aspect, for example the transcriptome, but this can only provide insights on genes at a transcriptional level. To achieve a holistic picture of the immune system, a systematic collection of multi-omics data is often required. The tissue (or source) to be measured is another important aspect to be considered. For example, the genome is usually regarded as a stable feature for each individual and is collected from an easily accessible tissue, usually blood. However, in some specific contexts, somatic mutations acquired after birth have to be considered and measured in specific tissues (17). In contrast, many other types of omics, such as transcriptome, proteome and metabolome, vary between cell types and tissues. Therefore, it is important to consider tissue in your experimental design and to aim to get as close to the relevant tissue as possible.

<table>
<thead>
<tr>
<th>Table 1. Typical approaches in omics measurements</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>sequencing-based</strong></td>
</tr>
<tr>
<td>genomics</td>
</tr>
<tr>
<td>whole-genome-seq, whole-exome-seq</td>
</tr>
<tr>
<td>epigenetics</td>
</tr>
<tr>
<td>ATAC-seq, whole-genome bisulfite-seq, RRBS-seq, DNAse-seq, FAIRE-seq, ChIP-seq, etc.</td>
</tr>
<tr>
<td>3D chromosome</td>
</tr>
<tr>
<td>H4, etc.</td>
</tr>
<tr>
<td>gene expression</td>
</tr>
<tr>
<td>RNA-seq, scRNA-seq, SLAM-seq</td>
</tr>
<tr>
<td>protein level</td>
</tr>
<tr>
<td>-</td>
</tr>
<tr>
<td>metabolites</td>
</tr>
<tr>
<td>-</td>
</tr>
<tr>
<td>microbiome</td>
</tr>
<tr>
<td>16s, RNA-seq, metagenomics, etc.</td>
</tr>
<tr>
<td>cellomics</td>
</tr>
<tr>
<td>single cell sequencing approaches</td>
</tr>
</tbody>
</table>

Given the complexity of the immune system, there is no gold standard for what to collect in multi-omics studies. The necessary data depends on the research questions and subjects. Understanding the different layers of omics data is helpful for setting up an appropriate study design. Therefore, in the following section, we introduce features and categories of different omics and describe important considerations when generating these data.

**Genome variation measurement**

Genotyping detects diversity in the genome. It describes small variations, such as single-nucleotide polymorphisms (SNPs) and insertion/deletions (InDels), as well as large-scale mutations such as insertions, deletions and amplifications. Genetic diversity can lead to variation in individual immune function (18).

To date, many techniques can be used to detect genotypes, including DNA sequencing, DNA microarrays (also known as genotyping chips) and PCR-based methods. These approaches can be categorized based on their measurement scales (high-throughput vs. low-throughput methods) or based on whether they include unknown variants (discovery vs. screening methods). Classical sequencing-based approaches detect genetic variants in a nearly unbiased manner on the genome (whole-genome sequencing) or within the exome regions (whole-exome sequencing),
including known or novel SNPs as well as structural mutation such as short insertions, deletions, and copy number variations.

Considering the cost and effectiveness of genotyping and cohort sizes, most population-based association studies choose genotyping screening methods such as DNA microarrays. These methods measure thousands to millions of known SNPs in well-studied organisms, such as humans and mice. The polymorphisms targeted depend on the chip designs. For example, the Immunochip contains 196,524 polymorphisms (718 InDels and 195,806 SNPs) on most reported loci involved in autoimmune and inflammatory diseases (19), whereas other custom genotyping chips such as the Metabochip (20) or the cardiovascular disease chip (21) contain loci designed for specific research areas. The number of variants that can be detected using genotyping chips has increased over the years, but even the high-density 5 million SNPs chip (Illumina OMNIV) covers only a small fraction of the 3.3 billion bases in the human genome.

To improve the power to discover genetic associations in the regions poorly covered by DNA microarrays, genotype imputation approaches are often used to expand coverage. For example, a commonly used genetic imputation server (https://imputationserver.sph.umich.edu/index.html#) takes the ~60,000 publicly available human haplotypes, covering ~40,000,000 SNPs, as a reference to impute millions of missing SNPs based on the measured genotypes and linkage disequilibrium (LD) structures (22).

Before association analysis, genotype data should pass a standard quality control (QC) at both individual and SNP level. Individuals with discordant sex information or outlying missing genotype or heterozygosity rate should be excluded (23). Duplicates and relatives could be identified by calculating identity-by-descent, and a multi-dimensional scaling plot merging with reference data such as the 1000 Genomes project (24) could help with the identification of individuals of divergent ancestry. SNPs that failed in genotyping and/or imputation and SNPs with low frequency and/or that deviate from the Hardy-Weinberg equilibrium are commonly removed before association analysis, especially in array-based studies, because these signals usually relate to poor genotyping quality. However, some low frequency SNPs may also contribute to rare diseases or phenotypes. With the increase in genotyping quality, many recent studies have focused on the function of rare alleles (minor allele frequency [MAF] < 0.01) (25-28).

Epigenome and 3D chromosome measurement

Epigenetics describes the study of chromatin traits (either in DNA or histones) that do not involve changes in the nucleotide sequence. Epigenetic measurements are mainly characterized by changes in histone modification (methylations and acetylations), DNA methylation, chromatin modification, chromatin accessibility, and chromosome structure.

DNA methylation is the process of adding methyl groups to DNA molecules, almost exclusively in CpG dinucleotides, with the cytosines on both strands being methylated.

This process usually acts in promoter regions to repress gene transcription and abnormal hypermethylation, which results in transcriptional silencing, and is often associated with immune diseases or used as a biomarker (29). Genome-wide techniques, such as whole-genome bisulfite sequencing (30), reduced representation bisulfite sequencing (31), and other non-targeted DNA methylation profiles, provide an opportunity to discover novel biomarkers. Other techniques, such as bisulfite-amplicon sequencing (32) and methylation arrays (33), detect the methylation status of CpG dinucleotides. Similar to genotyping arrays, the targeted regions from methylation arrays are based on the chip design. For the study of the human immune system, some well-established arrays can provide comprehensive coverage. For example, the MethylationEPIC BeadChip covers more than 850,000 methylation sites, making it ideal for an epigenome association study within big cohorts (34).

As the essential proteins that pack and order the DNA into structural units, histones play a role in gene regulation (35). Histone modification describes the post-translational modifications of histones, including methylation, acetylation, and other processes. Histone methylation often occurs as arginine (R), lysine (K), or histidine (H) residues of histone H3 or H4 being monomethylated (me1), dimethylated (me2), or trimethylated (me3). Array-based and sequencing-based approaches, such as ChIP-chip and ChIP-seq (36), are used to identify specific histone modifications that bind to DNA regions or domains.

Chromatin modification and accessibility is another important aspect of epigenetic changes. One of the most widely-used techniques to capture chromatin accessibility is called Assay for Transposable-Accessible Chromatin using sequencing (ATAC-seq). A standard “bulk” ATAC-seq measurement detects genome-wide open chromatin within a pooled sample or tissue, whereas, in order to capture cellular heterogeneity, single-cell ATAC-seq measures chromatin accessibility in thousands of individual cells and can generate genome-wide profiles from 10k to 100k cells per experiment (37). Alternative techniques are also used to investigate chromatin phenomena, including DNase-seq and FAIRE-seq, which measure open chromatin in regulatory regions, MNase, which identifies well-positioned nucleosomes, and ChIP-seq, which is used to detect binding sites of specific transcription factors (38).

Most epigenetic measurements also come with technical errors and biases. Biological replicates and technical replicates can help to characterize variability between samples and sequencing runs. Putting replicates of different conditions in the same batch is also important to avoid batch effects that could confound treatment effects. Large projects, such as the Encyclopedia of DNA Elements (ENCODE), have provided standard pipelines for processing many types of epigenetic data, such as ChIP-seq and ATAC-seq. However, this is not applicable in all cases. Applying appropriate QC strategies and software that accounts for bias effects according to the experiment design is essential to obtain robust results. To increase the coverage of epigenetic measurements, several methods, such as ChromImpute (39), Melissa (40), Avocado (41), and SCALE (42), provide imputation approaches for different epigenetic markers. However, existing imputation approaches have some limitations (15) and are not as widely applied as genotype imputation methods.
3D chromosome structure describes how chromosomes are folded, packaged, and organized into functional compartments and how different compartments are interconnected. Orthogonal ligation-based approaches include DNA-FISH, which can help with nuclear architecture visualization, and chromosome conformation capture (3C) techniques. One of the 3C techniques, Hi-C, is the most widely used approach to detect interactions between different genome regions (on gigabase scales) (38, 43). Single-cell adaptation of Hi-C methods are also used to investigate the interactions in individual cells (44).

Ligation-based approaches have the limitation of detecting DNA fragments connected with multiple genomic regions. To overcome this, orthogonal ligation-free methods including genome architecture mapping (45), split pool recognition of interactions by tag extension (46), and chromatin-interaction analysis via droplet-based and barcode-linked sequencing (47) were developed.

Transcriptome measurement

The transcriptome comprises all RNA molecules, including both coding and non-coding transcripts, in a single or population of cells. Traditional qPCR techniques can only quantify a limited number of genes at the same time. The most commonly used high-throughput techniques are RNA sequencing (RNA-seq) and microarray, and they can detect a large number of genes. Similar to genotyping methods, a sequencing-based approach (RNA-seq) can quantify the entire transcriptome, while microarray-based approaches (e.g., Affymetrix Genome U133 array and Illumina Whole-Genome Gene Expression BeadChips) are designed to target most known genes. In addition, typical RNA-seq can detect alternative splicing and rare isoforms, which microarray-based techniques cannot.

Depending on the aim of the study, specific levels of coverage are required for sequencing data. For instance, a bulk RNA-seq study for human differential expression profiling requires 10-25 million reads per sample, whereas alternative splicing or allele-specific expression analysis needs 50-100 million, and identifying novel transcripts requires >100 million reads per sample.

However, a “bulk” like measurement of the transcriptome cannot deal with cell heterogeneity and can be influenced by changes in cell composition. Single-cell RNA sequencing (scRNA-seq) was designed to uncover the transcriptome diversity in heterogeneous samples, characterizing the transcriptome at cell-level resolution. There are several scRNA-seq approaches, with plate-based (Smart-seq2) (48) and droplet-based (10x Genomics) the most commonly used. Usually, as few as 10,000 to 50,000 reads per cell are enough to detect cell types, and 500,000 reads can cover most of the genes (49).

To increase exonic coverage and the accuracy of gene quantification, polyA selection library preparation is commonly applied in scRNA-seq approaches such as 10x scRNA-seq (50). This will, however, miss important immune repertoire profiling, such as B-cell and T-cell receptors, which are mainly distinguished by their 5’ mRNA sequences. Thus, sequencing facilities, such as 10x genomics, provide full-length paired B-cell and T-cell repertoire sequencing, simultaneously, when examining cellular gene expression level. Combined with transcription measurement, this information can improve our understanding of clonal expansion and better characterize immune cell heterogeneity and functions (51).

SLAM-seq detects newly synthesized RNAs using a metabolic RNA-labeling approach. Compared to the other scRNA-seq techniques, this method can track transcriptome dynamics (52). For example, single-cell SLAM-seq has been applied to characterize the onset of infection with lytic cytomegalovirus in single mouse fibroblasts (53).

The transcriptome reflects the dynamic changes in biological processes, making an appropriate sampling strategy crucial when collecting transcriptome data. In addition to QC, normalization is usually performed within a sample and between samples. When considering comparison analysis, it is also necessary to have biological replicates and check for batch effects using clustering-based approaches. There are many computational tools to handle batch effects. Of note, integration approaches (54), as included in the Seurat (55) and Harmony (56) packages, are commonly used in scRNA-seq analysis and detect consistent cell-type signals from different batches or measurements. However, when the batch difference is confounded by other group information, it is difficult to filter out the batch effects. This makes it important to consider batch effects in the experiment design. Moreover, when considering sampling tissues for immune responses, circulating leukocytes are often measured for systemic inflammatory responses, while inflamed tissues are measured for local inflammatory responses. In order to expand the capacity, deconvolution approaches have been applied to bulk RNA-seq data to characterize cell type-compositions (57, 58), while expression recovery methods have been applied to scRNA-seq data to reduce dropout noise (59, 60). Like imputation approaches in genome and epigenome studies, one should be aware and careful with the potential false signals in these recovery or deconvolution approaches.

Proteome and metabolome measurement

Proteins are the major transcriptional products and functional units in the immune system. Immune molecules like immunoglobulins and cytokines are usually detected and/or quantified by immunoassays such as immunofluorescent staining, enzyme-linked immunosorbent assay, enzyme multiplied immunoassay technique, or mass spectrometry (MS)-based approaches.

In addition to independent measurements, proteins can be also measured together with RNA transcripts. CITE-seq provides a method to identify surface proteins along with RNA-seq. This approach is often used for cell labeling in scRNA-seq (61). Cells in different research sets (e.g. under different treatments, from different tissues) can be labeled with different antibodies as hashtags then sequenced together as one pool. This process has two advantages: it lowers costs and excludes potential batch effects. In addition, as we also know that some immune cell types have specific cell markers, this approach can also be used to identify cell types. For example, detection of CD3e, CD4, and CD8a proteins on the cell surface could help to distinguish CD4 T cells from...
transcripts, proteins, or metabolites from microbiota, respectively (75). Including metatranscriptomics, metaproteomics, and metabolomics that target identify taxonomies and abundances. There are also other omics approaches (human) reads, microbiome reads are aligned to the known microbiome genomes to include 16s rRNA sequencing and metagenomics sequencing. After excluding host disease pathology (72–74). The commonly used approaches to study microbiota to influence host immune functionality, and to be involved in immune-mediated human digestive system. This community has been reported to vary among individuals, Microbiota refer to all the micro-organisms in a certain environment, for example the number of potential applications in immune-related studies (63).

The study of metabolic processes that regulate immune cell responses, which is referred to as immunometabolism, has become an exciting area in translational research and is paving the way for novel therapies in immune-related diseases. The intermediate or end products of cellular metabolism are metabolites, which include, but are not limited to, lipids, fatty acids, amino acids, bile acids, and cholesterol. Considering the regulatory effects of metabolites on the immune response (12, 64, 65), the metabolome has become an important subject of study in immunological research.

Approaches to study the metabolome can be classified into targeted and non-targeted techniques. Nuclear magnetic resonance (NMR) spectroscopy is one of the most commonly used techniques, detecting specific nuclei in the target molecule (66). Compared to NMR, MS-based approaches are more high-throughput and quantify metabolites in a non-targeted way, which detect the mass-to-charge ratio (67). However, MS-based approaches have a limitation in annotating metabolites, which is the major drawback of this method compared to NMR. Metabolite data could be acquired from different sources of samples. Among them, circulating metabolites are the most commonly measured. There are also many studies about fecal and urine metabolites.

Similar to transcriptome analysis, proper normalization (usually a log transformation) is required in processing both proteome and metabolome data. Secondly, biological replicates and batch effects have to be taken into consideration as well. In terms of sampling tissues, in addition to blood cells and inflamed tissues, proteome and metabolome can be also measured in urine, which is thought to be a rich source but underestimated in recent studies (68–70). In addition, fecal metabolites are usually studied together with microbiota, which affects immune homeostasis and susceptibility of the host to immune-mediated diseases. Of note, a recent study reported a reference map for serum metabolites (71), which can serve as a guide to control for irrelevant confounders in serum metabolite studies.

Digestive system microbiota measurement

Microbiota refer to all the micro-organisms in a certain environment, for example the human digestive system. This community has been reported to vary among individuals, to influence host immune functionality, and to be involved in immune-mediated disease pathology (72–74). The commonly used approaches to study microbiota include 16s rRNA sequencing and metagenomics sequencing. After excluding host (human) reads, microbiome reads are aligned to the known microbiome genomes to identify taxonomies and abundances. There are also other omics approaches including metatranscriptomics, metaproteomics, and metabonomics that target transcripts, proteins, or metabolites from microbiota, respectively (75).

Of note, studies on human microbiota usually have relatively low concordance compared to other omics data studies. A recent study reported a number of host variables that could confound human gut microbiota researches. To be exact, body mass index (BMI), sex, age, geographical location, alcohol consumption, bowel movement quality (BMQ), and diet should be balanced between cases and controls when comparing gut microbiota compositions (76). In the context of sample collection, most microbiota samples are currently acquired from stool, while urine and exhaled gas could be other important resources for microbiome detection (77, 78).

Cellomics measurements

Cellomics measurements often reveal the systemic responses at the level of cells and tissues, typically including cell composition, cellular localization, and trafficking analyses. Cell composition is measured as cell type abundance or proportion, which is commonly quantified by flow and mass cytometry (79) (FCM and CyTOF) or single-cell sequencing techniques. With the help of cell surface markers or cell-specific expression markers, both techniques can characterize hundreds of circulating cell subpopulations covering major immune cells involved in the innate and adaptive immune responses (i.e. neutrophils, monocytes, lymphocytes, and their subtypes). Additionally, high-content screening (HCS) is commonly used to track cellular changes, including their localization, trafficking, and morphologic phenotypes (80, 81).

Systems analysis on omics data

After data collection and pre-processing using appropriate strategies, the next challenge lies in linking different omics datasets and clinical phenotypes. For a specific trait or disease, a systems model can be built to specify the role and effect of different data layers. In this model, the qualitative or quantitative characteristics are linked by their relationships, which need to be estimated via comparison, association, and other systems approaches. These links can simply be a correlation but they can also reflect a regulatory or causal effect. In this section, we introduce general system approaches used with different kinds of omics data and provide representative examples of how they can be applied in immunological studies.

Genome-wide association analysis and quantitative trait locus mapping

Genome-wide association studies (GWAS) aim to scan the whole genome to find genetic determinants of certain traits. When considering a binary trait (e.g. case-control), they compare the allele frequency in two groups of individuals, for example a disease group and a healthy group. A chi-squared test is often applied to test for statistical significance. It is usually considered that there are ~1,000,000 independent loci in the human genome, so a p-value less than the Bonferroni corrected threshold of 0.05/1,000,000 (5 × 10^-8) is regarded to be genome-wide significant (82).

To date, GWAS have identified ~5000 genetic risk loci of immune-related diseases in ~400 studies (83). These findings have improved our understanding of the genetic factors that influence immune-mediated diseases, further pointing to the genetic basis of pathology as well as identifying treatment targets.
Generally, GWAS identify pathogenetic genetic factors contributing to phenotypes (diseases), although the variants identified will not cause disease directly, but rather affect intermediate molecules. Quantitative trait locus (QTL) analysis is a statistical method to discover the genetic basis of the mediated phenotypes, such as gene expression (eQTL) (84), splicing (sQTL) (85), metabolites (mQTL) (28), methylation (meQTL) (86, 87), and immune traits (88, 89).

After data normalization, a linear regression between each genetic variant and each quantitative trait is applied. Covariates are crucial aspects of the regression model of quantitative trait locus analysis. Based on the type of omics, different covariates should be included in the model to correct the detected phenotypes. In general, basic host features such as age and sex are considered, and the population structure also has to be taken into account, especially in large cohorts with samples from admixed ancestry (90, 91).

eQTLs are the associations between SNPs and expression of genes, which provide insights into the function of genetic variants. eQTLs have been shown to explain 10-50% of the heritability of a phenotype/disease (92), which means that gene expression variation is one of the major consequences of genetic variants. It is very useful for prioritizing pathogenic genes when there is an association between a gene’s expression and a pathogenic genetic variant. Based on their location, eQTLs are classified into cis-eQTL (eQTL within 250kb of the gene) and trans-eQTL (eQTL located outside 250kb of the gene), with trans-eQTLs being more tissue-specific than cis-eQTLs (85). Of note, tissue-specific eQTLs provide a way for prioritizing pathogenic tissues (93).

QTL analysis on the epigenome identifies associations between genetic variants and epigenetic modification. Most genome-wide significant disease-associated loci (~93%) are located in non-coding regions (94), particularly the regulatory elements identified by ENCODE (95) and Roadmap projects (96). These observations highlight the importance of the epigenome in the genetic regulation of diseases and immune functionality. Similar to eQTL analysis, this analysis could help us find the potential epigenetic mechanism responsible for the association between genetic variants and immune traits/diseases. For example, one study investigated genetic variants that affect the activity of cis-regulatory domains (aCRD-QTLs) or correlation structure within cis-regulatory domains (sCRD-QTLs) in 317 lymphoblastoid and 78 fibroblast cell lines and their consequence on gene expression (97). At the same time, genetic variants can also affect methylation (meQTL) by influencing the binding of DNA methyltransferase. Large meQTL studies in blood samples have shown significant enrichment in autoimmune diseases such as ulcerative colitis and Crohn’s disease (98).

pQTL mirrors the associations between genetic variants and protein level. About 40% of cis-protein quantitative trait loci (cis-pQTLs) are also eQTLs, as expected, indicating a sequential genetic regulation between gene expression level and protein levels. By applying pQTL analysis, we could identify the potential mechanism at the protein expression level, behind the association from genetic variants to immune-related phenotypes. As with cis-eQTLs, cis-pQTLs are also located around transcription start sites (TSS). Notably, pQTLs have shown significant enrichment in missense, 3UTR and splice regions (99). pQTLs could also help in prioritizing the causal proteins/genes of immune traits/diseases. For example, a pQTL of serum IL18R1 and IL1RL1 also associated with atopic dermatitis, and this association between genetic locus and protein level indicates the possible involvement of IL18R1 and IL1RL1 in atopic dermatitis pathology (99).

Metabolites that mediate the association between genetic variants to immune functionality and immune diseases could be discovered using a mQTL analysis. More than 140 genomic loci are associated with circulating metabolite features explaining a median 6.9% heritability (100). Overlap between mQTLs and immune trait QTLs suggest the role of metabolic processes in the genetic regulation of immune functionality. For instance, a mQTL study indicated that the mQTL loci ARHGEF3 (rs1354034) and LRRCA8 (rs13297295) also affect platelet function and neutrophil function, respectively (101).

Immune phenotypes such as circulating immune cell proportion and cytokine production capacity in response to stimulations are crucial parameters when characterizing immune activities. Understanding the genetic determinants of immune phenotypes can provide insights into immune function and immune-mediated diseases. The Human Functional Genomics Project has identified >20 genetic factors determining immune cell proportions and cytokine production upon stimulations, which provided a link between genetic control and inter-individual variation (89, 102).

Integration of multiple genetic association profiles

In the context of immunological research, multiple diseases and molecular and cellular profiles can be regulated by the same genetic factors, indicating an internal association between them. Integration with multiple genetic profiles can provide insights and build connections between associated phenotypes. Ideally, such genetic profiles can be directly built from GWAS and QTL analysis of different layers from the same individuals. Alternatively, they can be also collected from different population-based cohorts. A number of computational approaches have been developed to discover the links. In particular, approaches like colocalization (103), genetic correlation (104), and Mendelian randomization (MR) (105) take genetic variants as the instrumental variables to infer the association or causality when multiple traits are associated with the same locus.

Colocalization analysis evaluates the association from each individual loci and helps to identify the phenotypes that share the same genetic regulation. Examples of colocalization analysis include a study that integrated genetics, epigenetics, and transcription to identify colocalization of molecular traits from CD14+ monocytes, CD16+ neutrophils, and naïve CD4+ T cells (106). Results from this analysis illustrated the molecular mechanisms at work within autoimmune disease-associated variants, including an alternative splicing signal around SP140 in T cells which might be involved in Crohn’s disease pathology.
Genetic correlation considers the full summary statistics to describe to what extent the genetic background is shared between two phenotypes. An example from a linkage disequilibrium (LD) regression-based genetic correlation approach showed a shared genetic basis of autoimmune diseases such as Celiac disease and type 1 diabetes, indicating a similar pathological mechanism between these two diseases (104).

MR is a statistical method working at the step from association to causality. If one trait (exposure) is causal for another trait (outcome), then the genetic factors contributing to the exposure should also contribute to the outcome. This will be reflected in the correlation between the effect sizes of the same genetic variant on exposure and outcome. There are many examples of immune-related studies that have applied MR, and these have led to the identification of causal relationships between IL-6 signaling and rheumatoid arthritis (107), IL-18 and inflammatory bowel disease (108), and eosinophilic indices and asthma (25).

**Comparison and association of the epigenome and 3D chromosome structures**

Systems analysis of epigenetic changes can investigate their influence on immune functionality or variation, and changes induced, as well as disease susceptibility and development (109, 110). In one example, the impact of cytokines was studied on the epigenome of insulin-releasing cells (β cells) from type 1 diabetes pancreases. By measuring ATAC-seq, Chip-seq, and RNA-seq, the authors identified that proinflammatory cytokines induced neo/primed epigenetic events in human β cells (111). Moreover, in the immune system, the effects of epigenetic changes lead to long-term alterations in metabolic and transcriptional pathways and further induce immune memory (112) or immunological diseases (113). Thus, epigenomics is another vital area for better understanding of the personalized immune system.

While genetics is stable, the epigenome is subject to dynamic changes that can be induced or affected by host and environmental factors such as smoking, drug usage, diet, aging, inflammation, disease, and exposure to pets. Considering that epigenetic changes affect gene transcription levels, the epigenome is a pivotal part to study when trying to understand immunological networks.

In a case-control study, differential accessible regions could be identified in ATAC-seq data and differential methylation positions/regions identified using bisulfite sequencing and methylation array. Instead of comparison analysis, association analysis could be applied to continuous phenotypes to identify associated regions. Based upon the position of these regions, researchers could further map them to the corresponding genes. More specifically, by checking which gene TSS regions are overlapped with the peaks/regions, these peaks/regions could be matched to genes, and pathway analysis then used to get more biological meanings. For example, in a multi-omics study on mixed-phenotype acute leukemia, researchers associated scATAC-seq with transcription responses from scRNA-seq and antibody captured from CITE-seq. Despite widespread epigenetic heterogeneity of chromatin accessibility within patients, they reported common malignant signatures across patients and thus revealed both distinct and shared molecular mechanisms of mixed-phenotype acute leukemia (114).

Another application of epigenetic analysis is to annotate the function of the identified regions based on the signals from epigenetic markers. One tool ChromHMM used a multivariate hidden Markov model applied to annotate regulatory elements (e.g. TSS, enhancers, promoters) with histone markers (e.g. H3K4me1, H3K4me3, H3K27me3, H3K9me3, H3K36me3) that bind to the chromosomes. One application of this method identified the chromatin states in mice and humans and reported the up-regulation of immune regulatory regions in Alzheimer’s disease (115).

Analyses of 3D chromosomes are generally similar. In a case-control study with Hi-C data, the compartment switches could be identified using a comparative analysis. Researchers could then further predict the interactions between those segments (116). By referring to public epigenetic databases or genome annotations, they could check the overlap between switched compartments or interactions and known epigenetic markers or elements. Based on this information, researchers could again associate the changes with other immune profiles or annotate the involved regulatory elements. For example, in a study of lineage commitment of early T cells with Hi-C data, the authors found wide compartment re-organizations across chromosomes from a transition between T cell double-negative-2 stage to double-negative-3 stage, and later double-negative-4 stage to double-positive stage. They annotated the changes with domain scores and, more interestingly, found that the changes in the domain scores between the two transitions are positively correlated, which suggests the re-organization at the former transition is actually reinforced at the later transition (117). Another example is a study on activated T cells that identified activation-sensitive interactions related to autoimmune diseases captured by Hi-C data (118).

**Comparison and association of transcriptome and proteome**

As the downstream products of genetic and epigenetic regulation, proteome and transcriptome changes directly reflect the influence of genetic and epigenetic variants. Comparison and association studies of transcriptome and proteome have allowed researchers to estimate functional units and validate hypotheses in immune regulation.

As for a case-control study, the first and direct analysis is identifying differentially expressed genes/proteins (DEGs/DEPs), followed by pathway analysis. If the corresponding phenotypes are continuous, the associated genes/proteins will be identified before pathway analysis. Examples include many transcriptome/proteome studies of COVID-19. Transcriptome measurement across samples from healthy, moderately affected patients and severely affected patients suggests an overall acute inflammatory response in COVID-19 patients wherein transcriptional responses of high cytotoxic effector T cells are associated with moderate patients and deranged interferon responses are associated with severe patients (119). Moreover, a urine proteome study identified 1986 urine proteins that show significant changes in COVID-19 patients compared to healthy controls (120).

Compared to bulk RNA-seq, the added information in scRNA-seq - cell composition - provides more analysis potential. In a case-control study, in addition to identifying DEGs and enriched pathways within each cell cluster/type, cell proportion could be compared between groups while novel cell subpopulations could be also identified in...
particular cases. For example, a scRNA-seq study on two COVID-19 cohorts reported identical dysfunctional neutrophil clusters in severe patients’ blood (9). When considering the T cell receptor/B cell receptor (TCR/BCR) analysis, it would be interesting to explore the clonotype expansion and diversity under different conditions (121, 122), immune development stages (51), or antigen specificities (123). Usually, a clonal expansion means an adaptive immune response targeting certain stimulation, since a specific receptor is the mediator of specific antigen recognition.

Since transcriptome/proteome data is responding rapidly to environmental changes, performing transcriptome/proteome analysis in a time-series could allow us to associate the dynamics with infection or stimulation in order to comprehensively understand the host immune responses. A nice example is demonstrated in a study of influenza vaccination efficiency where the authors measured the hemagglutination-inhibition antibody titers and transcriptional responses at baseline and multiple time points post-vaccination. By comparing the profiles between day 28 and day 180, the authors identified individual categories such as temporary and persistent responders and illustrated the underlying molecular mechanism (124). Of note, a dynamic study can be also achieved by a time-series design or by applying a trajectory analysis such as pseudotime analysis (125, 126) or RNA velocity analysis (127) in scRNA-seq analyses. In a recent study on SARS-CoV-2 infection, researchers longitudinally measured samples at several time points after symptoms and applied pseudo-time trajectory inference on scRNA-seq data of epithelial cells from the upper respiratory tract. Based on the trajectory, the authors predicted a new, alternative differentiation pathway that is dependent on the interferon response and marked by interferon-stimulated genes such as IFIT1, IFIT2, and CXCL10 (128).

Co-expression analysis among transcriptome or proteome can provide information about gene co-regulation and interactions. These co-expression relationships are inferred by different association methods, such as a weighted gene co-expression network analysis (129) applied to the transcriptome to identify consistent expression patterns among genes. The associations identified by gene expression could be applied to predict gene co-regulatory networks and to prioritize genes involved in the same pathways (130). At protein level, parts of these co-expression relationships could be further explained by protein-protein interactions, which are also collected by several protein-protein interaction databases, including innateDB (131), which focuses particularly on immune interactions. In application, similar to gene co-expression networks, protein-protein interaction relationships could help with functional/pathway enrichment analysis (132).

In recent single-cell experiments, co-expression relationships are further applied to predict the cell-cell interactions. By detecting the correlation between known ligand and receptor genes among different cell sub-clusters, researchers could infer potential communications between cell populations (133). This analysis fits well with immune network analysis. For example, by detecting ligand and receptor gene signals, a recent study identified cross-talk between CD8+ T cells and epithelial cells altered in the colon of ulcerative colitis patients compared to healthy controls (134). Additional methods, such as NicheNet (135), also use knowledge of gene regulatory networks or protein-protein interaction networks in public databases and the literature to build a model to further predict the activated targets of the cell-cell interactions by correlating the ligand’s expression level with its potential downstream gene or protein level interactions. In an example study of the cell-cell interactions underlying the tissue-specific imprinting of macrophages, the authors deciphered the interaction signals driving monocyte recruitment, engraftment, and acquisition of Kupffer cell-associated transcription factors and identified the contributions of different cells to the Kupffer cell niche (136).

**Comparison and association on metabolome/microbiome**

Metabolome or microbiome are additional factors that reflect, or affect, a person's state of health (137, 138). Similar to the transcriptome or proteome, comparison and association analysis could be applied to metabolome and microbiome data. However, the metabolome can be relatively harder linked to genes than other omics, which leads to different strategies of interpretation. Taking KEGG (139) and HMDB (140) as references, the online tool MetaboAnalyser performs metabolic pathway enrichment and network analysis on the identified metabolites (141). An example serum study on COVID-19 detected accumulation of 11 steroid hormones and suppression of amino acid metabolism in patients (8).

As to the gut microbiome, diversity analysis could be applied to taxonomy data. There are different strategies available for functional profiling of gut microbiome data. For example, the HUMAnN pipeline takes metagenomic or metatranscriptomic sequencing data as input to identify gene families and abundances (142). Gene families could be further matched to broader functional categories, such as MetaCyc metabolic pathways, and GO categories for functional interpretation. For example, a study that associated gut microbiome features to cytokine production capacity identified two microbial metabolic pathways, palmitoleic acid metabolism and tryptophan degradation to tryptophol, that showed associations with TNFα and IFNγ production (143).

As in transcriptome and proteome analyses, timeseries studies could provide valuable information in metabolome and microbiome data. For example, in a study of the metabolic functions of gut microbes from patients with inflammatory bowel disease, fecal samples were collected at baseline and 2, 6, and 14/30 weeks after induction of therapy to collect metabolic and microbiota profiles. The observed association in dynamics of metabolites and diversity shifts of microbiota reveals the heterogeneity of the disease, and the authors used it to build an in silico model that might be used to identify patients likely to achieve clinical remission from the therapy (144).

**Integration of epigenome, transcriptome, proteome, metabolome, microbiota, and cellomics**

In addition to associations between omics data and genetics, simple association analysis between two different kinds of non-genetic omics data could be applied to the data measured in the same cohort with a large sample size to find underlying co-regulation (Table 2). For instance, eQTM (associations between methylation and gene expression) provide a resource to integrate methylation and gene expression.
High methylation can block the binding of transcription factors on promoters and enhancers. In line with expectations, most eQTMs showed negative correlations between methylation and gene expression and negatively correlated eQTMs are enriched in active TSS regions (145). In another example, a study carefully characterized the changes in the gut microbiota of patients suffering inflammatory bowel diseases and the interplay between microbiome composition and gut metabolites (146).

Table 2. System analysis between omics

<table>
<thead>
<tr>
<th>Genomics</th>
<th>epigenomics</th>
<th>gene expression</th>
<th>protein level</th>
<th>metabolomics</th>
<th>microbiome</th>
<th>celiomics</th>
</tr>
</thead>
<tbody>
<tr>
<td>GWA</td>
<td>eQTL, GWAS</td>
<td>eQTL, mQTL, pQTL</td>
<td>eQTL, cQTL, mQTL</td>
<td>eQTL, mQTL, pQTL</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

For more complex multi-omics integration, advanced techniques like building a multivariable regression model could take features from different omics to evaluate the cumulative effects/prediction power on a certain phenotype. One example study integrated genomic, metagenomic, metabolomic, immune cell composition, hormone levels, and platelet activation profiles with cytokine response profiles in a population-based cohort. Results from multivariable linear regression and machine learning approaches such as elastic net then showed the accumulative contribution and prediction power of genetic and non-genetic factors on cytokine response (147).

On the other hand, if the sample size is insufficient for association analysis, it might be applicable to check the intersections between the findings from different omics. For example, we could easily compare the regions identified in ATAC-seq, methylation array and Hi-C data. In addition, by matching a DAR to genes, and intersecting with DEGs, we could further check whether an epigenetic change has the potential to regulate gene expression.

Discussion and perspectives

In this review, we have discussed multi-omics applications in immunological studies, from measurements and analysis to comparison or association of several typical layers (Figure 2). For system studies – in particular for newly discovered infectious diseases or rare diseases about which there is little prior knowledge – the choice of which data layers to collect and the selection of measuring approaches, target or non-target technique, or bulk or single-cell level can be as important as the analysis models and algorithms. Here, we discuss a few points that need specific attention in study design and interpretation and subjects the may see progress in the next few years.

There are some commonly used strategies for interpreting genetic associations. As the starting point of the central dogma of molecular biology (Figure 2), genetics has so far received a lot of attention and was associated with many types of data or phenotypes. In the interpretation of genome-wide associated loci, the genes around them have also been regarded as the necessary and most essential compartments. The strategy to properly link loci with affected genes so far has been focused on the position and associations between gene expression and genetic variants (i.e. eQTLs). In addition, functional annotation of identified loci, such as whether the variants are located in the regulatory elements or affect protein structure, may provide additional clues for loci interpretation in particular cases. Nevertheless, there are existing debates about several aspects, for example whether the host genome could influence the gut microbiome. It will never be nitpicking to be very careful in interpreting your microbiome QTLs.

Figure 2. Central dogma and regulations of different omics layers.

Epigenetics could be used as a window to study environmental influence. In contrast to genetics, epigenetics often links the external factors to immune phenotypes. This is particularly true when considering external effects as a risk factor in immune diseases, for example the role of smoking in asthma, because epigenetic modifications such as methylation are usually related to environmental exposures. Considering the various kinds of epigenetic changes, multiple types of epigenetic data are commonly used in one study, and they often validate and complement each other. In two examples of this, an active TSS region was identified by low methylation as well as high DNA accessibility (148), while the enhancer involved in a neo chromatin interaction identified in Hi-C data was characterized as a neo opening region in ATAC data (149). Considering the functional relationships, epigenetic data is commonly integrated with gene expression measurements. As the direct consequence of epigenetic modification, alteration in corresponding gene expression could be the best validation of the importance of epigenetic studies.

ScRNA-seq is usually applied together with cellomics measurements. A cell composition discovered in scRNA-seq data could be validated with FCM-based approaches. FACS is also commonly used as a pre-filtering step to help with concentrating target cell types for scRNA-seq analysis. In particular, for the rare cell
types (e.g. Tregs in PBMCs), a pre-sorting process is necessary for concentrating on cells of interest.

The proteome and metabolome showing downstream immune functions requires more attention. As the downstream products of gene expression, protein or metabolite level measurements are not as popular as transcriptome measurements in current studies. This might be because gene expression analysis takes advantage of the efficiency of next-generation sequencing and well-established microarray chips. Thus, there appears to be much room for further studies on the proteome and metabolome in immune studies.

Proper measurement techniques and sampling tissues are crucial in an omics study. When considering the purpose of measurements, it is often appropriate to apply high-throughput and/or non-targeted approaches at the discovery stage, while single and/or targeted approaches are more commonly used for validation. Moreover, excepting the genome, all other omics have tissue-specificity, and data from the same tissue are more commonly associated. For example, associations between omics from blood samples could be easily interpreted, but it would be trickier and require more biological basis to associate blood features with gut features.

To fully elucidate the biological processes involved in the immune system, several aspects remain unknown in omics studies. First, due to sample inaccessibility, fewer studies have been performed on tissues other than blood. Taking meQTLs as an example, several large studies have been carried out on blood samples (98, 150, 151), but there are only very limited sample sizes and/or studies about meQTLs in other tissues (152). Second, considering the high dynamics, rapid response, and spatial specificity of the immune system, temporal and spatial studies can provide more insights into the dynamic process and spatial heterogeneity in immune activities and/or immune-related disease etiology. For example, the process by which immune cells are activated by interacting physically and chemically with synapses is highly dynamic and depends on the spatial position of immune cells, neurons, and glial cells. Despite its importance in immune functionality and immune-mediated diseases, our current knowledge is not sufficiently advanced, which calls for more comprehensive studies (153-155). Third, with respect to population-based studies, many more of them have been carried out in healthy individuals of European ancestry, and studies in under-represented populations and in patients require greater attention.

Considering the complexity of our immune system and patient heterogeneity in severity or treatment responses for many immune-related diseases, the generation of personalized medicine is one of the most significant goals we can achieve through multi-omics studies (156). Personalized medicine stratifies a heterogeneous group of patients based on certain characteristics and provides treatment based on this stratification. In the case of infectious diseases, one personalized medicine trial is now ongoing in which tuberculous meningitis patients are being stratified based on genotype prior to treatment (158).

In conclusion, we have systematically reviewed the measurements and analyses that can be applied in immunological studies, which provides insights for personalized medicine. Through the development of high-throughput techniques, e.g. scRNA-seq and mass cytometry, we now possess the tools to unravel the many complexities of the immune system in health and immune-related diseases, including infectious diseases, allergies, and auto-immune diseases. With unbiased measurements and effective integration, multi-omics studies can help us understand the immune system and could lead to the development of personalized medicine.

Conflict of Interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Author contribution

X.C. conceived and designed the review. X.C. and B.Z. drafted the manuscripts, supervised by Y.L. All authors contributed to revision and text editing and approved the submitted version.

Funding

X.C. was supported by the China Scholarship Council. Y.L. was supported by an ERC Starting Grant (948207) and the Radboud University Medical Centre Hypatia Grant (2018) for Scientific Research.

References


Section 1

Genetic, transcriptomic, metabolic, and environmental dominators of immune functionality in health and diseases