Chapter 1

Introduction

Observable human characteristics or traits are called phenotypes. Variation in phenotype emerges from the interplay of genetics, environment, and pure chance. The contributions of each can vary immensely from phenotype to phenotype. Traits for which genetic variation explains a non-zero percentage of phenotypic variation are heritable. Virtually all phenotypic traits are heritable to some degree, and twin studies provide upper bounds on this heritability by partitioning phenotypic variation into genetic and environmental components [1].

Genetic variation presents a unique opportunity to probe the causal molecular mechanisms underlying phenotypes. Information encoded in the genome has phenotypic consequences only after flowing through multiple molecular layers. This guiding principle is the central dogma of molecular biology, whereby the flow is directed from DNA to RNA to protein via transcription and translation. Barring somatic mutation, an individual's genome is fixed at conception, providing a causally upstream anchor that can be measured with relatively little error. A mainstay of the field of human genetics is uncovering the specific genetic variants that contribute to heritability of phenotypes through statistical association of variants and phenotypes. Although not immune to population-level biases like stratification [2] and collider bias [3], genetic association has intrinsic resistance to reverse causality, an issue that permeates observational studies of the causes of human phenotypes.

1.1 Genetic association studies for complex traits

1.1.1 Structure and variation of the human genome

The human genome is over three billion bp (base pairs) in length, containing 20 000–25 000 protein-coding genes that span 1–3 % of its length, with the remaining sequence being non-coding [4, 5]. Each diploid cell contains two copies of the genome, organised into 46 chromosomes comprised of 23 maternal-parental pairs: 22 pairs of homologous autosomes and one pair of sex chromosomes. Variation in the genome between individuals in a population exists in the form of single nucleotide polymorphisms (SNPs), short indels, and structural variants. For common population variants with minor allele frequency (MAF) >1 %, the vast majority (>99.9 %) are SNPs and short indels [5]. On average, a pair of genomes differs by one SNP per 1000–2000 bp [6]. Each version of a variant is called an allele; each individual has a maternal and parental allele at each variant.

The large number of variants in a population are inherited on a smaller number of haplotypes: contiguous stretches of the genome passed through generations via meiotic segregation. The fundamental sources of genetic diversity, mutation and meiotic recombination, generate new alleles and break apart haplotypes into shorter ones over evolutionary time. Variants that are physically close on a chromosome are less likely to flank a recombination event, hence are more likely to cosegregate from parent to offspring on the same haplotype (genetic linkage). Genetic linkage is one source of linkage disequilibrium (LD): the non-random association of alleles at two variants, differing from expectation based on their population frequencies and the law of independent assortment. LD can be quantified by r^2, the squared correlation coefficient between alleles in a specific population [7]. Recombination events are not distributed uniformly throughout the genome. The genome is a mosaic of haplotype blocks delimited by recombination hotspots, characterised by strong LD within blocks, and little LD between blocks [8, 9] (Fig. 1.1). The structure of correlated haplotypes reflects a population's unique evolutionary history, and can be used to trace the demography of populations back through time [10].

1.1.2 Lessons from the past fifteen years

Genetic variants can affect heritable traits by impacting the function or regulation of target genes. How genetic variation contributes to a particular trait defines its genetic architecture: the number of genes affecting the trait; and the frequencies, effect sizes, and interactions of trait-associated alleles [12, 13]. The number of genes defines a spectrum of traits from monogenic (where inheritance follows simple Mendelian patterns) to polygenic (where inheritance is complex). Proposed architectures differ strikingly among complex traits, even for traits with phenotypic similarities like type 1 diabetes (T1D) and type 2 diabetes (T2D) [12]. Consistently, however, the number of genes and genetic variants affecting a complex trait is large (ranging from dozens to many thousands), the average effect size of trait-associated variants is small, and the contribution of environment is substantial [14–16].

Since the 1980s, linkage analysis has been used to map the chromosomal positions and regions (loci) affecting traits by tracing the cosegregation of markers (variants with known positions) with the trait in family pedigrees [17–19]. Linkage analysis was complemented by early genetic association studies, which largely focused on variants in or near candidate genes selected on the basis of prior biological knowledge [20]. These methods saw much success for Mendelian traits, but application to most complex traits proved challenging. Small average effect sizes meant penetrance was too low to reliably observe cosegregation in pedigrees [19]. Early candidate gene studies were severely underpowered to detect such small effects [21].

The past fifteen years have seen the rise of genome-wide association studies (GWASs) that systematically test common variants selected in a hypothesis-free manner across the whole genome (Fig. 1.2). Using large sample sizes to overcome small effects and the large multiple testing burden, thousands of associations have been discovered for complex traits and diseases, many robustly replicated across populations [19, 22]. A number of take-home messages have emerged. Most genetic variance is additive; the contributions of dominance and epistatic interaction are small [13]. Variants with effects on multiple phenotypes (pleiotropy) are widespread [19]. Even traits

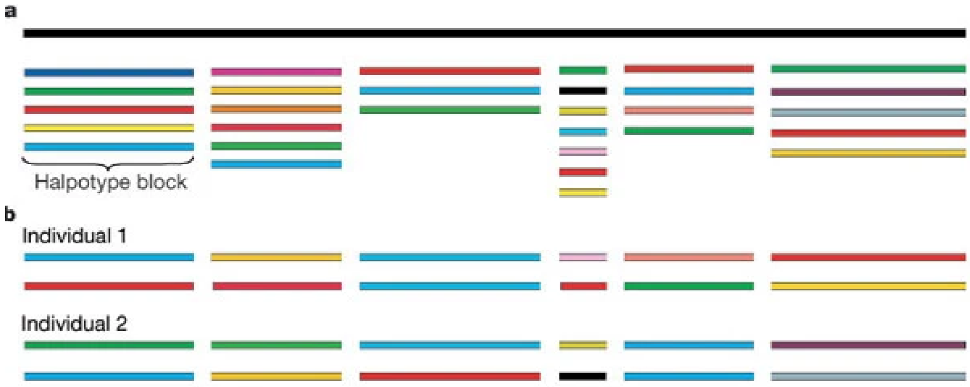


Figure 1.1: The mosaic structure of human genetic variation. Large parts of the genome can be divided into haplotype blocks between 5–200 kbp in length, with strong intra-block LD. For each block, three to seven common haplotypes (indicated by different colors) represent the majority of variation found in humans. An individual carries two haplotypes per block, one inherited from each parent. The exact structure and diversity of haplotype blocks varies between populations. Information on the haplotypes, their locations in the genome, and their frequencies in different populations form a “haplotype map” of the genome. Figure reprinted by permission from Springer Nature: Springer Nature, Nature, Pääbo [11], © 2003.

that are molecular rather than whole-organism phenotypes can be remarkably polygenic, with hundreds to thousands of associated loci [23]. GWAS sample sizes in the millions are increasingly commonplace, and the discovery of new associations with ever smaller effects as sample sizes increase shows no sign of plateauing [24, 25].

1.1.3 From complex trait to locus

GWASs rely on the tendency of common variants on the same haplotype to be in strong LD. As the number of haplotypes is relatively few, it is possible to select a subset of tag variants such that all other known common variants are within a certain LD threshold of that subset. In practice, there is enough redundancy that the number of variants measured on a modern genotyping array (in the order of 10^5–10^6) is sufficient to tag almost all common variants [27, 28]. Associations with unmeasured variants are indirectly detected through correlation with a tag variant. Furthermore, as unrelated individuals still share short ancestral haplotypes, study samples can be assigned haplotypes from a panel of haplotypes derived from reference samples by matching on directly genotyped variants. Genotypes at untyped variants can then be assigned from those haplotypes. This process—genotype imputation—allows ascertainment of many more variants than are directly genotyped [29] and helps to recover rarer variants that are poorly-tagged [22]. Modern imputation panels enable cost-effective GWASs testing tens of millions of variants as rare as 0.01–0.1 % in diverse populations [30].

Testing large numbers of variants incurs a massive multiple testing burden, but acknowledging the correlation between variants due to LD and restricting tests to common variants, there are only the equivalent of $\sim 10^6$ independent tests in the European genome, regardless of the number of tests actually performed [31]. The field has thus converged on a fixed discovery threshold of $0.05/10^6 = 5 \times 10^{-8}$ for genome-wide significance in European populations [32], akin to controlling the family-wise error rate (FWER) to below $\alpha = 0.05$ using the Bonferroni

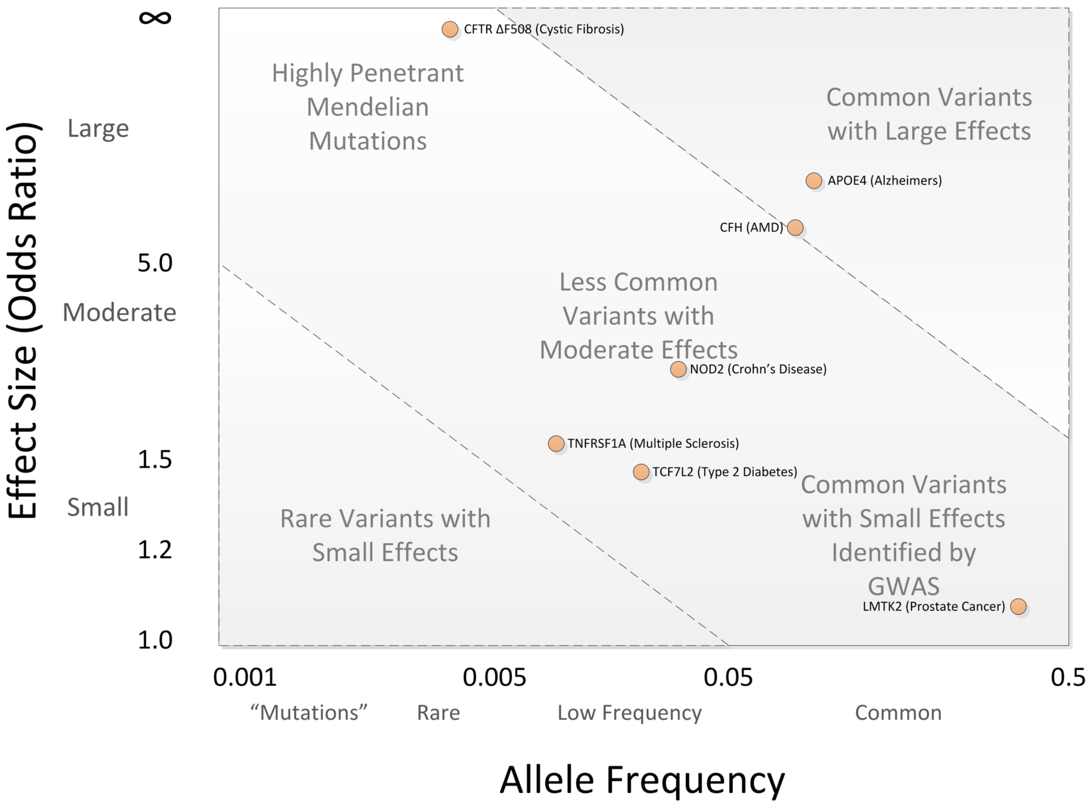


Figure 1.2: Effect size and allele frequency of trait-associated genetic variants. Different classes of genetic effects require different and complementary methods. Linkage analysis is suited to detecting Mendelian variants with large effects. GWAS is suited to detecting common variants with small effects. There are few common variants with large effects due to selection pressure. Rare variants with small effects are hard to distinguish from noise without very large samples. They are also poorly tagged by genotyping arrays and difficult to impute. Studies focusing on rare variants often employ whole-exome sequencing (WES) or whole-genome sequencing (WGS) to directly type variants. Figure reproduced from Bush *et al.* [26] under the CC BY 4.0 license (`creativecommons.org/licenses/by/4.0/legalcode`).

correction*.

1.1.4 From locus to causal variant

By design, a significant trait-associated variant from a GWAS needs not be a variant that causally affects the trait and may only tag a causal variant. The resolution of the associated locus depends on the local LD structure. Fine-mapping is the process of determining which of the many correlated variants in an associated locus are most likely to be causal, assuming the causal variants are observed either by direct genotyping or confident imputation. Due to incomplete power, the causal variants in a locus are not necessarily the ones with the strongest associations [34]. Bayesian fine-mapping methods take a variable selection approach, assigning each variant a posterior probability of causality. A credible set of variants likely to contain the causal variant in the locus with some probability can then be determined [34, 35]. The ability to separate causal and tag variants depends on factors including LD, sample size, and the effect size and number of causal variants [22, 34].

1.1.5 From causal variant to target gene

Most causal variants for Mendelian traits are coding variants (nonsense, missense, or frameshift) that impact protein sequence [36]. In contrast, over 90 % of GWAS loci fall in non-coding regions [37], and often too far from the nearest coding region to be in LD [38]. Even if the causal variants in a locus are fine-mapped, one of the greatest challenges following a GWAS is prioritising the target genes through which those variants affect the trait. A reasonable heuristic is to assign the gene with the nearest transcription start site (TSS) or body as the target, particularly for metabolite traits [39]. For improved accuracy across a variety of complex traits, integrative methods for gene prioritisation combine variant-to-gene distance with other metrics and data types drawn from numerous external sources [39–41].

1.2 Gene expression as an intermediate molecular phenotype

1.2.1 Regulation of gene expression

Gene regulation data are indispensable for gene prioritisation. Rather than directly impacting the coding sequence of a gene, many non-coding GWAS loci are hypothesised to affect traits by affecting the regulation of target gene expression [37, 42]. Unlike genotype, expression is dynamic across time and space. Diverse expression programs are responsible for the myriad of cell and tissue types generated during development, and enable adaptation in response to environmental stimuli.

Expression is the product of eukaryotic transcription, a multi-step process involving interactions between DNA, RNA, and hundreds of proteins [43]. Transcription of the pre-messenger RNA (mRNA) is initiated when RNA polymerase and transcription factors (TFs) form part

[*] The Bonferroni correction makes no assumptions about the dependence structure of the p-values, controlling the FWER (probability of at least one type I error) exactly under any structure. It is conservative (i.e. controls the FWER at a stricter level than the chosen threshold α) even for independent tests. In fact, it is always conservative unless the p-values have strong negative correlations [33].

of a protein complex around the promoter region and TSS of a gene. TFs can also bind to more distant *cis*-regulatory elements such as enhancers and repressors. These distant regulatory elements interact with the promoter region via DNA looping. Transcription can only happen in regions of open chromatin, where the packing of DNA-histone complexes (nucleosomes) is loose enough that the DNA is physically accessible to the transcriptional machinery. Chromatin accessibility is partially determined by histone modifications such as methylation, acetylation, phosphorylation, and ubiquitination [44]. The DNA itself can also be modified; methylation at CpG sites in promoters tends to repress transcription [45].

To form a mature mRNA, the pre-mRNA is capped at the 5' end by a modified nucleotide and at the 3' end by a poly(A) tail. Exons are joined by spliceosomes that cut and rejoin the pre-mRNA at one or more pairs of splice sites, excising the intronic sequence between each pair. The choice of splice sites determines which of many alternatively-spliced transcripts is produced. Post-transcriptional regulation of mature mRNAs is also possible via RNA editing [46] and regulatory elements in the flanking 5' and 3' untranslated regions (UTRs) [47].

In line with the regulatory hypothesis, GWAS variants are heavily enriched in regulatory elements annotated by functional genomics projects (e.g. ENCODE [4]), including regions of open chromatin, histone binding sites, TF binding sites, enhancers, splice sites, and UTRs [48–52]. Furthermore, this enrichment is often observed in particular contexts (tissues, cell types, or cell states [22, 37, 42]). An example is the enrichment of fine-mapped SNPs associated with risk of immune-mediated inflammatory disease (IMID) in CD4$^+$ T cell enhancers, particularly in enhancers activated after stimulation [50]. These results put forth expression as an important intermediate that links non-coding GWAS variants to their associated traits, and helps nominate trait-relevant contexts and target genes.

1.2.2 Expression quantitative trait loci (eQTLs)

Expression is a complex molecular phenotype in itself, with a heritability of 15–30 % [53]. Genome-wide assays for expression, such as microarrays and RNA sequencing (RNA-seq), were among the earliest high-throughput technologies developed for quantifying molecular phenotypes. Genetic loci associated with quantified gene expression are called expression quantitative trait loci (eQTLs). Large-scale efforts such as the Genotype-Tissue Expression (GTEx) project [54] have pioneered the study of eQTLs and other molecular quantitative trait loci (molQTLs) over the past decade [55].

eQTL effect sizes are large relative to variants associated with whole-organism phenotypes, with the average eQTL explaining 5–18 % of additive genetic variance for its associated gene [53]. The eQTLs with the largest effects tend to be concentrated near the TSS of their target gene (*cis*-eQTLs), affecting TF binding sites and other local regulatory elements. eQTLs further away or on a different chromosome are called *trans*-eQTLs. The exact threshold separating *cis* from *trans* on the same chromosome is arbitrary; <1 Mbp and >5 Mbp are commonly used thresholds for *cis*- and *trans*-eQTLs respectively [56–58][*]. In general, eQTL effect size declines with distance

[*] Having a threshold is often a matter of practicality to reduce the number of variants tested. Assaying expression is still more costly than array genotyping, so eQTL mapping sample sizes are small compared to GWASs. Even though eQTL effects are relatively large, eQTL mapping genome-wide would be equivalent to performing GWASs on thousands of continuous phenotypes, incurring enormous computational and multiple testing burdens.

to the TSS, and *trans*-eQTLs have smaller effects compared to *cis*-eQTLs [55]. *Trans*-eQTLs often represent *cis*-eQTLs of regulatory molecules like TFs and RNA-binding proteins that may target many genes in *trans* as master regulators [57, 59]. Gathering large enough samples to detect *trans*-eQTLs remains a priority, as most expression heritability is driven by *trans* rather than *cis* effects, perhaps due to small but wide-reaching effects [60].

1.2.3 Context-dependent eQTLs

Like expression itself, the effects of eQTLs are highly context-dependent [55, 57]. When the effect size of an eQTL is not the same in all environments, but differs depending on the environment, the eQTL is said to interact with those environments. This can manifest as a statistical interaction in a regression model with a multiplicative genotype-environment term, where the effects of environment and genotype on expression are not additive at the chosen scale for measuring expression. A non-exhaustive list of environmental contexts that have been found to interact with eQTLs includes sex [61], age [61], ancestry [62–64], tissue [65, 66], purified cell type [62, 67–70], cell type composition in bulk samples [71–74], cell differentiation stage [75], disease status [68], and experimental stimulation (see Section 1.2.4). These contexts can be interdependent; for example, tissue-dependent effects may arise from a combination of cell type-dependence and varying cell composition between tissues.

A multitude of molecular mechanisms could facilitate genotype-environment interactions at eQTLs. Fu *et al.* [76] mapped eQTLs in blood and four non-blood tissues (Fig. 1.3), and proposed mechanisms that might explain discordant effects of an eQTL allele on a target gene between tissues, assuming the eQTL disrupts a regulatory factor's binding site. Different effect sizes of same or opposite signs could arise from tissue-dependent effects of the same factor, such as activating expression in one tissue and repressing it in another (e.g. due to cofactors, or from binding of different factors in different tissues at the same site). Effects specific to a tissue could arise from tissue-specific expression of a regulatory factor. A tissue-specific effect could also reflect tissue-specific target gene expression, as the eQTL effect will be zero in a tissue where the target is not expressed (e.g. due to chromatin inaccessibility). Tagging of different causal variants in the two tissues, potentially with differing tagging efficiency (i.e. LD), could also generate the above scenarios [76]. Furthermore, the complexity of human gene regulation means these mechanisms might be acting at epigenetic, pre-, co-, or post-transcriptional regulatory levels [53]. Detection of context-dependent effects merely exposes differences in regulatory architecture between contexts. Much like in GWASs, going from association to underlying mechanism requires considering data types beyond just genotype and expression.

1.2.4 Response eQTLs (reQTLs)

A important class of context-dependent eQTLs are response expression quantitative trait loci (reQTLs), where the interacting environment is experimental stimulation, revealing regulatory effects not detectable in the baseline state [55, 77]. The vast majority of reQTL studies to date have been conducted on immune cells. This is not only due to the abundance of immune cells

Studies focused specifically on *trans*-eQTL mapping reduce the number of tests in other ways, such as testing only significant GWAS variants for eQTLs [58].

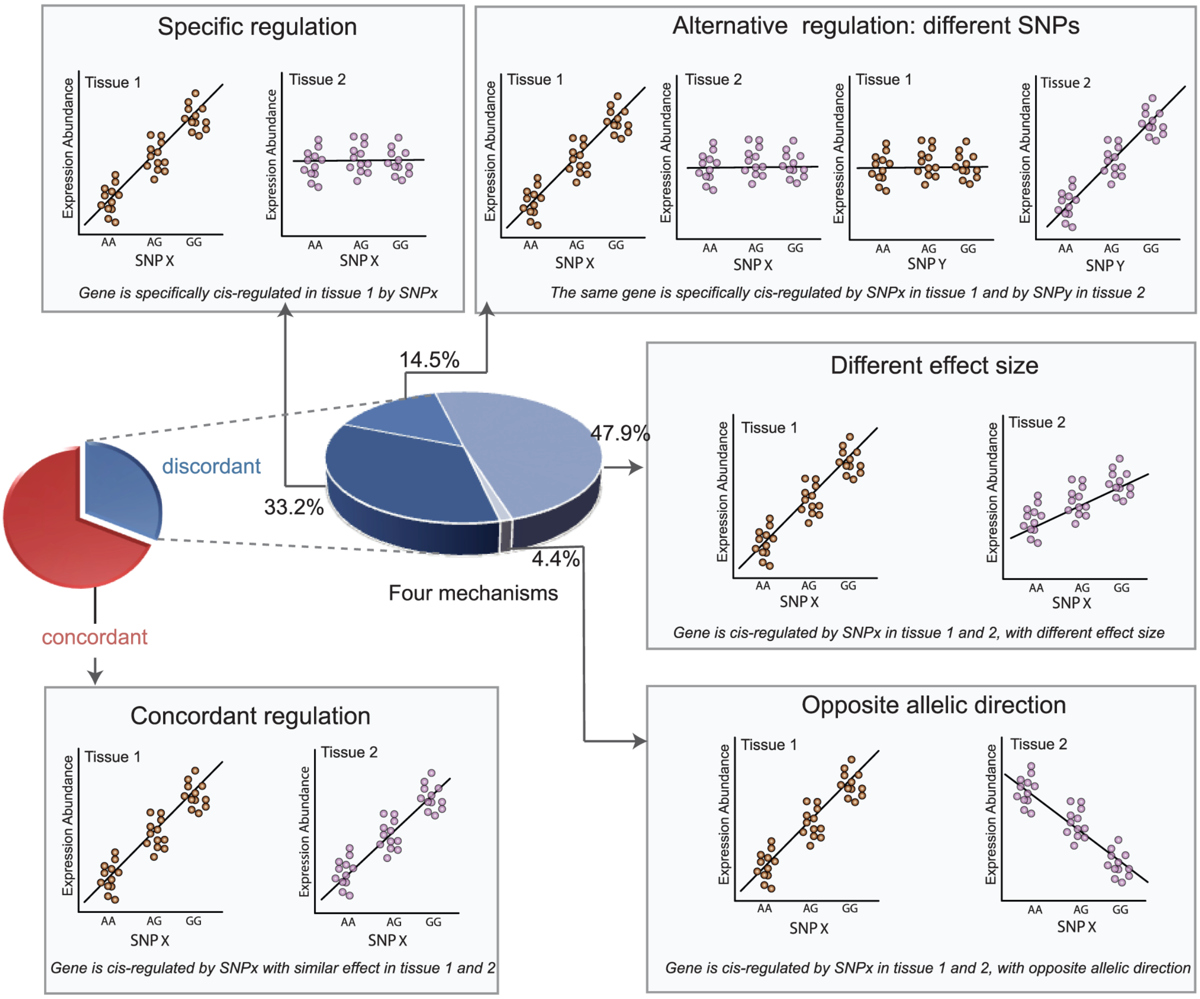

Specific regulation
Tissue 1
Tissue 2
Expression Abundance
AA AG GG
SNP X
Gene is specifically cis-regulated in tissue 1 by SNPx
Alternative regulation: different SNPs
Tissue 1
Tissue 2
SNP Y
The same gene is specifically cis-regulated by SNPx in tissue 1 and by SNPy in tissue 2
14.5%
47.9%
33.2%
4.4%
discordant
concordant
Four mechanisms
Different effect size
Gene is cis-regulated by SNPx in tissue 1 and 2, with different effect size
Concordant regulation
Gene is cis-regulated by SNPx with similar effect in tissue 1 and 2
Opposite allelic direction
Gene is cis-regulated by SNPx in tissue 1 and 2, with opposite allelic direction

Figure 1.3: Types of tissue-dependent *cis*-eQTL effects. The effect size of an eQTL SNP on expression can be concordant or discordant between tissues. Discordant effects represent genotype-tissue interactions and can be classified into four subtypes. Specific regulation refers to a gene with a significant eQTL in only one tissue. Alternative regulation refers to regulation of the same gene by independent SNPs in the two tissues—genes can have multiple eQTLs with tissue-specific effects. Different effect size refers to an eQTL having tissue-dependent effects with concordant sign but discordant magnitude. Opposite allelic direction is a tissue-dependent effect where the sign is discordant. Pie charts show the proportion of different types of effects in pairwise comparisons of blood and four non-blood tissues by Fu *et al.* [76]. Figure reproduced from Fu *et al.* [76] under the CC BY 4.0 license (creativecommons.org/licenses/by/4.0/legalcode).

easily accessible in peripheral blood, amenable to purification and stimulation, but because the immune system is specialised for mounting different responses to different pathogens and perturbations.

When stimulation is applied *in vitro*, variables such as cell type and abundance; and the nature, length, and intensity of stimulation can be precisely controlled. A seminal study by Barreiro *et al.* [78] mapped eQTLs in monocyte-derived dendritic cells (DCs) before and after 18 h of infection with *Mycobacterium tuberculosis*. reQTLs were detected for 198 genes: 102 specific to the uninfected state and 96 specific to the infected state. They observed a 1.4-fold enrichment of reQTLs among GWAS variants associated with susceptibility to pulmonary tuberculosis, but no enrichment of eQTLs shared between uninfected and infected DCs. From overlap of reQTLs and GWAS variants, three genes (*DUSP14*, *ATP6V0A2*, *RIPK2*) were prioritised as candidates affecting tuberculosis susceptibility. Since then, numerous *in vitro* reQTL studies have been conducted with a variety of stimulations (often cytokines, pathogens, or pathogen-associated molecular patterns (PAMPs)), applied to purified [70, 79–91] or mixed cell types [83, 92].

A complementary approach is reQTL mapping with *in vivo* stimulation. An isolated mixture of cells *in vitro* cannot hope to replicate the innumerable interactions involved in human immune response. *In vivo* designs suit whole-organism stimulations and response phenotypes, such as vaccination and vaccine-induced antibody response. Published *in vivo* reQTL studies are comparatively few. Idaghdour *et al.* [93] mapped whole blood eQTLs in 94 West African children admitted to hospital for malaria and 61 age-matched controls. reQTLs with a significant case-genotype interaction were detected for five genes: *PRUNE2*, *SLC39A8*, *C3AR1*, *PADI3*, and *UNC119B*. As *SLC39A8* is upregulated with T cell activation, a postulation was made that T cell activation is important to malaria infection response. In Franco *et al.* [94], whole blood eQTLs were mapped in 247 healthy adults given trivalent inactivated influenza vaccine (TIV). Twenty genes involved in membrane trafficking and antigen processing were prioritised to be important to vaccine response, on account of having post-vaccination reQTLs or differential expression, and an expression correlation with antibody response. Lareau *et al.* [95] focused on epistatic effects of SNP-SNP interactions on expression fold-change after smallpox vaccination in 183 individuals. Eleven significant interactions were found where the effect of two independent SNPs on expression was non-additive. Apoptosis-related genes (e.g. *TRAPPC4*, *ITK*) were enriched among target genes. Most recently, Davenport *et al.* [96] mapped whole blood eQTLs in 157 systemic lupus erythematosus (SLE) patients in a phase II clinical trial of an anti-IL-6 monoclonal antibody. Nine reQTLs with effect sizes magnified by drug exposure were found to disrupt the binding site of IRF4, highlighting it as a key regulatory factor downstream of IL-6. Overall, *in vivo* reQTL studies have delivered insight into the biology of a diverse set of whole-organism phenotypes. However, ethical requirements can limit sample size and choice of stimulation. Many environmental factors (e.g. diet, lifestyle, immune exposures) cannot be controlled, potentially leading to greater experimental noise compared to *in vitro* designs, and complicating interpretation of results.

1.2.5 Gene prioritisation using eQTLs

eQTLs are enormously valuable for target gene prioritisation after GWAS. They propose both target gene and mechanism of action, where the effect of variant on complex trait is mediated through expression. GWAS variants for many traits are indeed enriched for eQTLs [97], but care must be taken when conducting enrichment analyses to avoid false positives due to the abundance of eQTLs. At current sample sizes, 60–80 % of genes have at least one detectable eQTL [55, 58]. Assuming that a locus is associated with both a trait of interest and with expression of a particular gene, how can one separate the scenario where the same causal variants affect both trait and expression (pleiotropy) from coincidental overlap between distinct sets of causal variants that may possibly be in LD? Bayesian colocalisation methods address this problem by extending Bayesian fine-mapping methods to multiple phenotypes [98–100]. Using information from all measured variants in the locus, they estimate the posterior probability that the same causal variants are associated with both phenotypes, distinguishing pleiotropy from LD.

Given the effect of an eQTL can be starkly context-dependent, eQTL datasets from trait-relevant contexts are most useful for gene prioritisation. For instance, immune *in vitro* reQTLs are enriched more so than non-reQTLs among GWAS associations for immune-related phenotypes, such as susceptibility to infectious [78, 92] and immune-mediated diseases [85, 92]. Supplementing shared eQTL effects with cell type-specific eQTL effects finds many additional colocalisations with complex traits [74, 101]. The increasing number of context-dependent eQTL datasets available for large-scale colocalisation analyses means eQTLs can propose not just target gene and mechanism, but also the specific environments most relevant to a trait.

1.3 Phenotypes of immune response

1.3.1 An overview of the immune system

Immunology began as the study of host defense against infection [102]. It is now recognised that the immune system is also involved in pathogenesis of diverse conditions encompassing allergic diseases, autoimmune and immune-mediated diseases, and cancer. This subsection provides a basic overview of parts of the immune system relevant to this thesis.

The two major arms of the immune response are the innate and adaptive response. The innate response is rapid and non-specific, occurring in the first few minutes to days after the initial (primary) exposure to infection. This triggers the adaptive response, which takes days to weeks to develop, but delivers a powerful and specific response capable of eliminating pathogens that have evaded the innate response. The adaptive response can also create immunological memory lasting years to decades, where re-exposure to the same pathogen induces a faster and more powerful recall response[*]. Both arms distinguish self from non-self through complex interactions between many cell types via surface receptors and signalling molecules.

Immune cell types differentiate from common myeloid progenitor or common lymphoid progenitors, which themselves are descended from pluripotent hematopoietic stem cells (HSCs) in the bone marrow. By and large, the cells of the innate response are of the myeloid lineage,

[*] There is increasing evidence the innate immune system also has a form of immunological memory [103].

and the cells of the adaptive response are of the lymphoid lineage. Immune cells are also called leukocytes or white-blood cells as many types can be found in peripheral blood, but certain types are confined to tissues or parts of the lymphatic system.

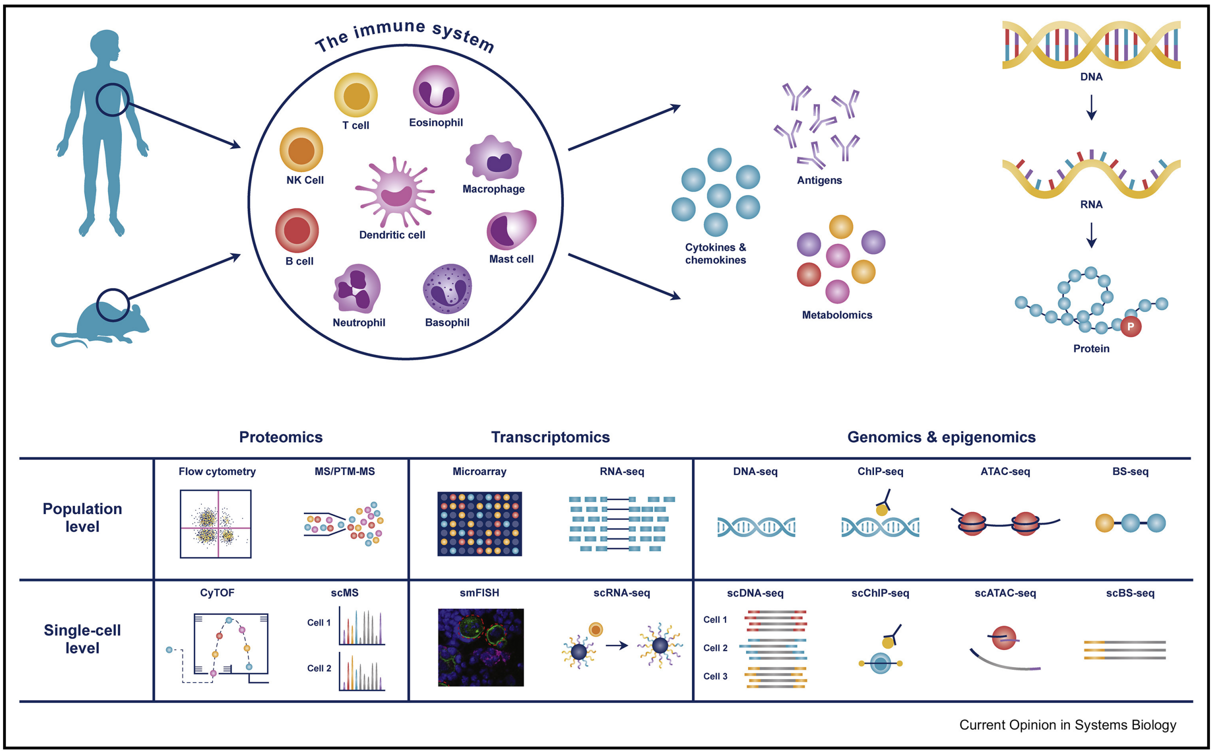
Innate response begins with the detection of pathogens by phagocytotic sensor cells—primarily neutrophils, tissue-resident macrophages, and DCs. These cells express pattern recognition receptors (PRRs) that recognise conserved PAMPs not present in host cells, then secrete small proteins (cytokines) that trigger the inflammatory response: a massive recruitment of multiple cell types from blood into infected tissues. Recruitment is partially mediated by a family of cytokines called chemokines, which chemically attract immune cells by creating a concentration gradient (chemotaxis). Recruited neutrophils clear pathogens by phagocytosis and secrete antimicrobial molecules by degranulation. Natural killer (NK) cells detect and kill virus-infected and tumour cells. Circulating monocytes migrate to the site of infection and differentiate into macrophages and DCs. Macrophages perform phagocytosis, modulate inflammation, and can also engage in antigen-presentation—but it is DCs that are considered to be the specialist antigen-presenting cell (APC) type. Antigen-presentation by DCs is a key link between the innate and adaptive responses.

The main forces of the adaptive response comprise B and T lymphocytes. Naive lymphocytes express antigen receptors that recognise parts of specific antigens called epitopes. When they encounter this antigen, they activate, proliferate (clonal expansion), then differentiate into effector cells. To initiate adaptive response, CD4$^+$ (helper) T cells recognise peptide fragments from the antigen presented in a complex with major histocompatibility complex (MHC) class II on the surface of APCs. CD4$^+$ T cells then differentiate into several subsets; these activate and regulate other immune cell types such as macrophages, CD8$^+$ T cells and B cells. Activated CD8$^+$ (cytotoxic) T cells recognise antigens presented by MHC class I on infected cells and directly kill the cell. Activated B cells differentiate into plasma cells that secrete large quantities of antibodies, the soluble form of the B cell receptor (BCR). Antibody-mediated immunity is also called humoral immunity, whereas T cell and innate immune responses comprise cell-mediated immunity. A small subset of activated B and T cells can become memory cells, responsible for long-term immunological memory.

1.3.2 High-throughput immunology

To understand the immune system and its intricate interactions, “systems immunology” studies take a holistic rather than reductionist experimental approach [104–106]. The basic principle is the same: experimentally perturb the immune system and observe its response. Drugs and vaccines can be used as safe and synchronised perturbations—one of the largest subfields of systems immunology is systems vaccinology, which I review in Section 2.1.4. A range of high-throughput technologies are applied to measure response at many layers of the immune system (Fig. 1.4). Longitudinal designs are common, aiming to sample timepoints corresponding to baseline, innate, and adaptive immunity. The complexity of the immune response presents a major challenge, with the richness of sampling required often restricting the sample sizes of systems immunology studies due to cost and logistics.

There are three major themes to systems immunology. Initial studies of immune response to



Current Opinion in Systems Biology

Figure 1.4: High-throughput technologies for systems immunology. Profiling can be done in humans and model organisms, at multiple levels of the immune system, and at bulk or single-cell resolutions. An additional dimension not shown here is profiling at multiple timepoints before and after perturbation. Figure reprinted from Yu *et al.* [107], © 2019, with permission from Elsevier.

a particular perturbation are often descriptive, aiming to find correlations between components of the immune response. Predictive studies then evaluate the ability to use measurements of relevant components to predict individual responses to the perturbation. Feature sets that are molecular phenotypes (e.g. gene expression) with validated predictive accuracy are known as molecular signatures. Causal inference is the third and most difficult goal. Fortunately, the heritability of immune parameters (e.g. cell counts, surface marker expression, serum protein levels) is substantial (20–40 % [108–111]), with greater heritability for innate than adaptive immune parameters [110]. Much akin to GWAS and quantitative trait locus (QTL) studies, to identify causal links in the immune system, one can leverage genetic variants as naturally-occurring perturbations [105, 112]. Controlled variation can also be systematically generated by RNA interference or genome editing [113]. Obtaining causal understanding is essential for clinical translation, to determine the interventions that can be made to promote effective response to pathogens and vaccines, and impede pathways that lead to immune dysregulation in disease.

1.4 Thesis outline

This thesis examines longitudinal response to *in vivo* immune perturbations by vaccines and drugs. Chapter 2 is a descriptive differential gene expression (DGE) study of transcriptomic and antibody responses to pandemic influenza vaccine (Pandemrix) in the Human Immune Response Dynamics (HIRD) cohort of healthy adults. Chapter 3 integrates HIRD genotype data to map the regulation of expression response to Pandemrix using an *in vivo* reQTL study design. In

Chapter 4, I mirror the design of the previous two chapters, exploring clinical response to biologic anti-tumour necrosis factor (TNF) therapy for Crohn's disease (CD) patients in the Personalised Anti-TNF Therapy in Crohn's Disease (PANTS) cohort. Finally, Chapter 5 presents an overview of shared themes and limitations, and provides recommendations for future analyses and study designs for immune response phenotypes.