Chapter 5

Conclusion and outlook

5 Conclusion and outlook

5.1 The compromise between high-throughput and in-depth functional insights

With recent technological advances and the falling cost of whole-genome sequencing, the main challenge we now face is not the generation of genetic data but in the interpretation of biological mechanisms linking GWAS loci to function. We must consider what are the best approaches in understanding the different steps involved in the pathway from sequence variation to organismal phenotypes such as disease susceptibility. Addressing these challenges will aid the translation of GWAS findings to the clinic and capitalise on the power of genetics to predict and identify drug targets.

In this thesis, I have used a combination of approaches and phenotypic traits in an attempt to understand the cellular and functional consequences of genetic variation. These methods fall into two broad categories. First, large-scale annotation efforts such as those from the ENCODE and BLUEPRINT consortia provide broad functional insight into multiple loci across the genome. These data can be used to either annotate variants with epigenomic functionality or directly linking molecular phenotypes to variants in formal QTL association tests. Second, detailed bespoke investigations such as recall-by-genotype studies or targeted genome-editing provide in-depth insight but are generally lower-throughput and focus on a small number of loci. An important question as we endeavour to move from GWAS to function, is should our focus be on the application of GWAS in larger sample sizes (perhaps in millions of individuals) using existing traits or on increasing the phenotype complexity by continuing recent efforts to apply genetic approaches to functional and cellular traits?

In Chapter 1, I provided further demonstration of the power of molecular QTLs in providing workable mechanistic hypotheses for disease- and complex trait-associated variants as well as suggesting a relevant and specific experimental cellular model. I also provided further support for how the identification of genes dysregulated by variants can highlight potential therapeutic targets for disease treatment and management. The use of such phenotypes to provide functional insight through colocalisation and enrichment methods, as employed here, is a vast improvement on early methods of gene target identification, which relied on proximity to the sentinel variant with no indication of causal cell types involved. Indeed, QTL discovery is now being extended to a wider range of cell types and cellular contexts and will vastly improve our ability to search for the molecular consequences of significant loci. However, it is important to note that large-scale annotation efforts in multiple cell types and

large cohorts require substantial financial, logistical and analytical investment, as I experienced through working as part of the BLUEPRINT and UK Biobank consortia (Chen et al., 2016a, Astle et al., 2016). In addition, for certain cell types, it is challenging to access human samples. In some cases, important cell populations are present in low numbers and therefore provide technical challenges in both isolation and the application of genomic approaches. Some advances have already been made in applying ChIP-seq to small populations, such as haematopoietic progenitors (Lara-Astiaso et al., 2014). Many protocols for the differentiation of induced pluripotent stem cells (iPSCs) to different cell populations are already available and advances in this area will also facilitate the study of a wider range of cell types, particularly when coupled to genome-engineering approaches such as CRISPR. Extensively characterised iPSC cell lines are available to research groups through The Human Induced Pluripotent Stem Cells Initiative (HipSci, http://www.hipsci.org).

Even with the already vast amount of epigenome and QTL data available to the wider community, I have shown that to reach full mechanistic understanding still requires detailed, painstaking and often manual integration and annotation of important loci. Creating unified databases to summarise and visualise all available genetic, functional data and the multiple associations for each locus would greatly facilitate these efforts. Certainly, platforms such as HaploReg (Ward and Kellis, 2012), the Open Targets Platform (Koscielny et al., 2017) and DrugBank (Law et al., 2014) already improve the efficiency of this process, as I showed throughout my thesis. However, a unified browser of all available data from multiple cell types would greatly improve the formulation of functional hypotheses of individual genetic loci.

In Chapters 2 and 3 of this thesis, I moved beyond molecular processes, first performing a GWAS on novel neutrophil function traits and second performing an in-depth functional investigation into neutrophil count and surface receptor expression, which in both cases provided functional insight into the biology of the important immune cell type, neutrophils. Recall-by-genotype studies leverage the power of previous GWAS, in this case of nearly 174,000 individuals, to design follow up experiments that delve deeper into the functional processes.

My implementation of functional phenotypes here was not the first example of the utilisation of these traits in genetics (Orru et al., 2013, Roederer et al., 2015, Steri et al., 2017, Li et al., 2016b, Astle et al., 2016, Ahola-Olli et al., 2017). Indeed, for some cell types (particularly PBMCs and monocytes/macrophages) and assays (cytokine production), these measurements are as tractable as the genomic approaches used in molecular QTL studies. However, in Chapter 3, my work demonstrated that for some phenotypes and cell types,

measurement of function can be technically complex and subject to many sources of co-variation. Specifically, I outlined the difficulties in working with neutrophil function, which may represent a particularly challenging cell to work with compared to other haematopoietic cell lineages and I discussed possible reasons for this in Chapter 3. However, I gleaned mechanistic insight from small cohorts (10s-100s) using other functional measurements on these cells, namely flow cytometry of the surface expression of receptor proteins. In studying a specific neutrophil count associated locus, the important neutrophil receptor, GCSFR, I identified that while there was no evidence of association of the same common variant with the standard molecular phenotypes (gene expression, histone modifications), the locus was significantly associated with surface receptor expression, allele-specific expression and the expression level of certain *CSF3R* transcripts. Clearly, additional insight can be obtained by combining multiple layers of molecular, cellular and functional information.

Combining the lessons learnt from this thesis and from other similar studies, it is clear that many trait-associated variants affect not just one functional phenotype but many processes from the epigenome, to gene expression, post-transcriptional processes, protein levels, cell function, cell abundances and beyond. In reality, we cannot restrict our efforts to just one type of approach. Collating multiple phenotypes from unified large populations will enable phenome-wide association studies that identify multiple phenotypes affected by a single locus. Thus, molecular and functional networks could be constructed and provide insight into the complexity of interconnections in biological processes. Expanding sample sizes of these cohorts will increase power and enable detection of *trans* effects, which from our currently limited knowledge of these affects, seem to be even more subject to changes in cellular contexts (Delaneau et al., 2017). Providing dense, if not complete, genetic maps using high quality imputation or whole-genome sequencing combined with rich phenotype data will potentially lead to a full description of the genetic architecture of complex traits and perhaps eventually, prediction of an individual's risk based on their genetics. Such initiatives are seen with the recent biobank studies from, for example, the UK Biobank (Collins, 2012), INTERVAL (Moore et al., 2014) and the Precision Medicine Initiative in the United States (Sankar and Parker, 2017).

Below, I discuss two key challenges that we face in understanding functional genetic variation; that is the ever-increasing complexity of the regulatory epigenome and dissecting causal variants and causal relationships between different traits. To end, I highlight particular ongoing efforts to address these challenges and highlight a particular area in human disease that could also reveal interesting interactions with human disease; the microbiome.

5.2 The ever-increasing complexity of the epigenome and the regulatory code

Despite the vast amount we have learnt from genome annotation methods and large-scale consortia, we are still unable to read the regulatory code. We cannot yet predict epigenomic function directly from sequence alone. What we do know is that the regulatory genome is more complex than originally predicted. For example, the genomic regulatory function is highly cell type and context- specific as well as controlled in the three-dimensional space in addition to the early two-dimensional models of regulation of protein binding to a linear DNA sequence. Transcription factors and other regulators bind DNA with multiple cofactors forming large functional clusters, as I set out in Chapter 1. Complex interactions between cofactors located at varying genomic distances could perhaps underlie the observations of distal SNPs affecting TF binding even when they are not located within or nearby to the binding site (Wang et al., 2017b). Context specificity and the complexity of multiple layers of regulation suggests that in order to answer these questions we would require genome-wide binding profiles of *all* possible TFs and cofactors in *all* possible cell types and contexts and connect local effects to *all* interactions in the 3D space. While this seems like a daunting task, observed high functional correlation in genomic domains such as TADs, sub-TADs or variable chromatin modules may reduce the dimensionality of the regulatory genome thereby allowing us to study only the key “seed” factors that explain the majority of variability in the genomic region (Grubert et al., 2015, Waszak et al., 2015). Indeed, in this thesis, I demonstrated that different variant target genes between myeloid and hepatic cells at the CAD *SORT1* locus seemed to be due to the binding of lineage-specific pioneer factors, PU.1 and FOXA1 respectively. It seems remarkable that a small number of factors may be able to control cell-type specific transcription, but the challenge is establishing which layer of functionality underpins the causality at a particular locus and whether causal TF binding, as has been suggested, is a general genome-wide phenomenon or applies to specific cases (Wei et al., 2017).

Indeed, as we build our understanding of the regulatory genome, we find that the genome is even more complex and often challenges previously established functional paradigms. This, in turn, further complicates the interpretation of non-coding sequence variation. For example, I have described in this thesis how the histone modification, H3K4me1, is generally associated with poised or active enhancers and correlates with cell-type specific gene expression (Heintzman et al., 2009). Recently, however, a role for H3K4me1 bound at promoters was observed in inducible gene repression, rather than activation from a distal enhancer (Cheng et al., 2014). This repression seems to be mediated by the methyltransferase, MLL3/4 and appeared to restrict access to readers of the active promoter mark, H3K4me3 (Cheng et al., 2014). It was also recently shown that intragenic enhancers

can attenuate gene expression rather than activate it. By using CRISPR-cas9 knock-down, the authors showed that deleting an intragenic enhancer from the mouse ESC gene, *Meis1,* led to de-repression in the region and phenotypes consistent with ESC differentiation (Cinghu et al., 2017). Therefore, intragenic-enhancer repression appeared to have a physiological role. Interestingly, this effect was evident for genes that were not highly expressed but were expressed at medium-to-low levels, which may suggest that weaker intragenic enhancers could be repressive and stronger intragenic enhancers remain active (Cinghu et al., 2017).

Both of these examples show the importance of considering the full genomic context when interpreting variant function. They also demonstrate a need to increase our efforts to study all aspects of genomic regulation to help us ascribe function to important genetic variants.

5.3 Causal variants and causal relationships

Identifying causal variant(s) is an important step in fully understand the mechanism of action of genetic loci. In Chapter 2, I briefly discussed how the colocalisation method provides a posterior probability estimate of each variant being causal that can be used to fine-map potential causal variant(s) based on association evidence (Section 2.3.5.2) (Pickrell et al., 2016). Although not applied in this thesis, there are also other statistical methods available for fine-mapping causal variants (Spain and Barrett, 2015). In addition, I demonstrated how epigenomic information can facilitate fine-mapping, for example, if a particular variant disrupts a TF binding motif (CAD *SORT1,* Figure 2.10) or lies directly under a histone peak (AMD *TNFRSF10A*, Figure 2.17). However, for loci that contain many variants in LD that overlap the same epigenomic marks, we are still limited in detecting the causal variant. In this case, further experimental approaches could be employed to help dissect complex loci. For example, the combination of genome engineering and high-throughput production of induced human pluripotent stem cells (iPSCs) (Kilpinen et al., 2017) allows single nucleotide knock-out in a wide range of differentiated cell types. Coupling CRISPR-Cas9 approaches with tractable experimental read-outs, could help distinguish between closely correlated variants by experimentally comparing their effect sizes on an intermediate phenotype. In addition, using denser genetic information in GWAS also increases the chances of identifying the true causal variant. In Chapter 2, when I used the denser BLUEPRINT phase 2 cohort that included the predicted causal SNP, the associations at the AMD *TNFRSF10A* locus were more significant for all phenotypes of interest.

Establishing causality between different traits is also a complex challenge in human genetics. I described this concept in Chapter 1, where establishing whether a particular trait is a causal risk factor for a disease requires more than just identifying a high correlation between the two

factors. Establishing causality between intermediate (including molecular) phenotypes and disease risk provide useful readouts for monitoring disease progression, for potential therapeutic targets and also for experimental use in querying functional relationships at genetic loci. A clear therapeutic success started with the observation that variants within the HMG-CoA reductase gene, *HMGCR*, are associated with lipid levels. Now, *HMGCR* is targeted by cholesterol-lowering statins (Kathiresan et al., 2009). Cholesterol and LDL levels are known risk factors for coronary artery disease (Khera and Kathiresan, 2017).

In this thesis, I used a colocalisation method to provide a statistical assessment of regions of the genome that were associated with two different traits but this method does not provide a definitive demonstration of causality between traits. Mendelian randomization (MR) approaches have been successfully used to assess the causality between traits such as blood counts and complex diseases or LDL and CAD risk (Astle et al., 2016, Khera and Kathiresan, 2017, Holmes et al., 2017). MR application to molecular traits is complex. Often there is extensive QTL pleiotropy, where a variant affects multiple genes, which could all exert different effects on the outcome/disease. I also investigated examples where a genetic variant was associated with multiple molecular phenotypes such as TF binding, enhancer activity, and gene expression. In these cases, there may be multiple molecular routes through which a change in gene expression could affect an outcome. There also could be inter-relationships between the molecular function, which would essentially be described as reverse causation. In Figure 5.1, I highlight possible interactions and directionality between multiple epigenomic functions, partly based on my observations in Chapter 2. These complex molecular relationships can violate MR assumptions (Evans and Davey Smith, 2015).

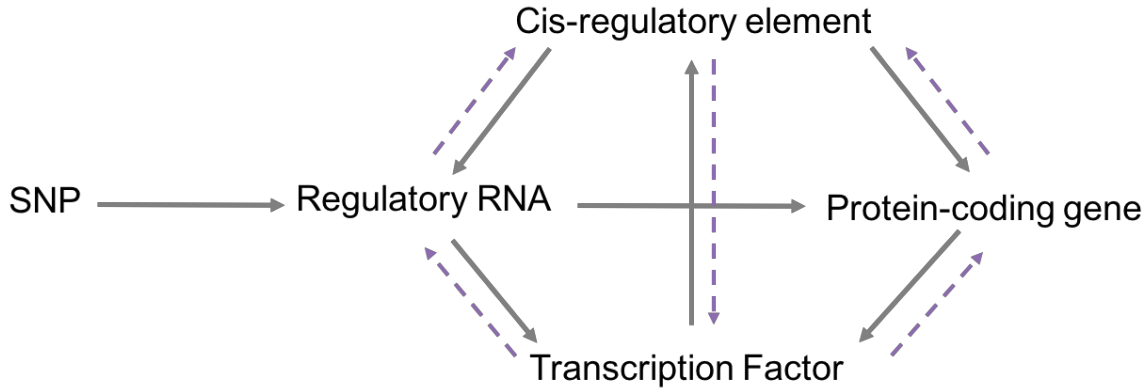

Figure 5.1: Complex molecular regulatory mechanisms
This schematic summarises possible complex regulatory mechanisms that could occur at one genetic locus. Block arrows represent the direction of regulatory effect and dashed arrows represent possible feedback mechanisms or relationships between functions. The generally accepted model is that pioneer factors first bind and then stimulate further chromatin modifications to form cis-regulatory elements, but for some TFs, open chromatin or certain functionality is required (hence the dashed arrow). There could also be feedback mechanisms between the level of gene expression and molecular regulators.

In light of this, I would argue that currently, integrating QTLs with disease GWAS SNPs is state of the art in forming molecular hypotheses, but currently demonstrating a causal relationship must come from downstream experimental testing. Many common variants, however, have small effect sizes on diseases or complex traits, making it difficult to experimentally demonstrate an effect on disease as a result of manipulation of gene function. Mouse models can highlight the consequences of extreme cases of gene knock-out or overexpression, which can identify relevant phenotypes. However, if a target has effects on many pathways, such as *TNFRSF10A*, this could be difficult to ascertain. Intermediate phenotypes that have been shown to be causal risk factors for a particular disease can be used as tractable experimental readouts to establish causal mechanisms. For coronary artery disease, our knowledge of causal intermediate risk factors is quite advanced due to extensive MR analysis and experimental evidence. In particular, LDL levels can be measured in mouse models and the effect of variant knock-down or overexpression on these levels can and have been evaluated (Musunuru et al., 2010). Indeed, cellular responses are also tractable experimental measures. I have already discussed the example of measuring the propensity of macrophages to form foam cells after exposure to oxidised LDL (Reschen et al., 2015).

However, amenable intermediate risk factors have not yet been established for all diseases, particularly for diseases that involve tissues that are difficult to access such as age-related macular degeneration, Alzheimer's disease or schizophrenia. If there is a causal role for peripheral blood cells in these diseases, it would be interesting the perform MR analysis using blood counts and these diseases to assess whether these are causal risk factors. Future work must involve efforts to identify tractable and causal disease intermediates. This, combined with experimental investigation would enable definitive assessment of the causality between peripheral immune function and for example AMD.

While a definitive causal relationship between immune factors and some of the diseases studied here has yet to be determined, therapeutics targeting dysregulated immunological processes that reduce disease severity or progression or manage severe symptoms. Therefore, there is a potential benefit to the management of these disorders through identifying pathways that increase disease severity for example.

5.4 Ongoing efforts and future goals in functional genomics

The next few years promise to be an exciting era for human genetics both in studies of vastly increased sample sizes and in new efforts to understand cellular function. In one of the largest single genetic cohorts and most comprehensive resource, genetic data and detailed phenotypes of 500,000 individuals has been released by UK Biobank project. The increased power of this dataset will allow identification of many more rare variants and will transform our knowledge of the allelic architecture of complex traits (Vasquez et al., 2016).

There are also exciting ongoing efforts in improving our understanding of human cellular biology. The Human Cell Atlas is an international collaboration aiming to use cutting-edge single-cell approaches to classify all human cells (Rozenblatt-Rosen et al., 2017). Although we have learnt much from genomic approaches applied to bulk tissue samples, the results provide an average picture across all possible cellular sub-types. Accurate molecular blueprints on a single cell basis for every type of cell in the human body will provide unprecedented insight into cellular interactions, different cellular states, transitions involved in differentiation, and potentially uncover previously unknown cell subtypes. Integration of these data with GWAS variants will allow us to gain a more detailed understanding of how genetic variation affects cellular phenotypes ultimately influencing disease risk (Rozenblatt-Rosen et al., 2017).

Another important factor in modulating immune responses that was not explored in this thesis is the gut microbiome. Changes in the composition of these bacteria and their

taxonomy have been implicated in multiple human diseases such as inflammatory bowel disease, multiple sclerosis, rheumatoid arthritis and asthma (Hall et al., 2017, Berer et al., 2017). A role for host genetic variation in bacterial composition alternations was revealed through using microbiome analysis such as 16S rRNA or metagenomics sequencing as the phenotype in GWAS (Hall et al., 2017). Identified loci had a clear role in disease, for example, 48 IBD risk genes were also associated with altered gut microbiome composition (Hall et al., 2017). The interaction between the host and microbiome also affects the production of host cytokines. Up to 9.7% of the variation in cytokine production was explained by the gut microbiome (Netea et al., 2016). Clearly, the influence of the microbiome in immune responses and genetic disease is an important area to consider. Large cohorts including the Framingham 4000 cohort (Mahmood et al., 2014) and TEDDY (N = 10,000) (Group, 2007) will perform microbiome GWAS and enable further exploration of the host-microbiome inter-relationship. As ever, there is the challenge of assessing causality between these factors and disease outcome.

I discussed how experimental approaches using causal risk factors as tractable readouts and MR approaches are currently employed to assess causality between different phenotypes. In future, longitudinal studies where susceptible individuals are tracked prior to disease onset would be invaluable in assessing the causality between immune function, microbiome composition or molecular changes with disease. This requires the establishment of detailed population-based biobanks that collate a wide-range of rich phenotype data including molecular, cellular and functional measurements as well as lifestyle factors (Leading Edge Voices, 2017). The INTERVAL study is such an example where the design is analogous to a randomised clinical trial. Here the aim is to link genetic determinants to the propensity of individuals to develop anaemia after repeat blood donations (Moore et al., 2014). Rich phenotypic data will be collected at several time points.

For complex diseases with later-in-life onset, such as age-related macular degeneration or Alzheimer's disease, realising the potential for causality assessments may take several generations of data collection and analysis. However, currently these resources will allow us to build our knowledge of pleiotropy, heritability and genetic architecture of many traits. Biobank resources may also help us understand why some individuals who carry the risk variants do not develop disease (Leading Edge Voices, 2017). Importantly, these biobanks also provide the opportunity of engaging the public in scientific endeavours as we move to collating data from larger and larger populations. Such large and potentially dynamic data resources bring with them consent and ethical challenges, which will need to be addressed (Caulfield and Murdoch, 2017).

In summary, although there remain challenges in interpreting the function of GWAS associations, a decade after the initial GWA studies, we have seen many examples of the power of genetics to uncover novel biological paradigms and potentially improve the success of therapeutic candidates in clinical trials. Indeed, GSK and Regeneron recently committed to sequence the first 50,000 UK Biobank samples, providing a denser variant set than the original array genotyping (GlaxoSmithKline plc., 2017). AstraZeneca announced efforts to build an integrated genomic database consisting of two million genomes as well as clinical trial and electronic health records data (AstraZeneca, 2016). It has even been estimated that by 2025, between 100 million and two billion human genomes would have been sequenced (Stephens et al., 2015). This wealth of genetic data may well enable us to completely resolve the genetic heritability and allelic architecture of complex traits and in doing so transforming our knowledge of basic science and the genetic risk of complex disease.