6 Conclusions

I have spent the past four years trying to understand how genetic differences between individuals lead to condition-specific differences in human macrophage gene expression. I have done this by first developing and validating a scalable cell culture model based on differentiating human induced pluripotent stem cells (iPSCs) into macrophages. I have subsequently used the model to study the genetics of gene expression and chromatin accessibility in macrophage response to IFNɣ stimulation and *Salmonella* infection.

6.1 Using iPSC-derived cells to map QTLs for molecular traits

Large iPSC generation initiatives such as the HipSci project (Kilpinen et al., 2016) provide both genetically and phenotypically well characterised cell lines from healthy individuals as well as from individuals with rare diseases. With the development of automated iPSC derivation and characterisation pipelines, the availability of these cell lines is likely to increase even further (Paull et al., 2015). Throughout the thesis, we have shown that it is feasible to use iPSC-derived macrophages to map QTLs for molecular traits such as gene expression and chromatin accessibility. Importantly, in Chapter 3 we have identified experimental factors (such as cell purity) that are responsible for a large amount of variability in the gene expression levels of iPSC-derived macrophages. These results can guide future QTL mapping experiments in iPSC-derived macrophages, but it is currently not clear how generalisable these observations are to other cell types and differentiation protocols.

Multiple studies have shown that a large fraction of eQTLs become visible only after specific environmental stimuli (Barreiro et al., 2012; Lee et al., 2014; Maranville et al., 2011) and even the duration of the stimulus can have a large effect (Fairfax et al., 2014). Furthermore, there can be a scores of relevant stimuli for any given cell type (Xue et al., 2014). Moreover, as we have shown in Chapters 4 and 5, applying two stimuli one after the other (e.g. IFNɣ + *Salmonella*) can reveal QTLs that are not visible with either of the stimuli alone. As a result, the logistics and the number of cells required for all relevant conditions can become prohibitively large for primary cells, especially if the cell type of interest is not easily accessible. IPSC-derived cells are free of these limitations because, in principle, large numbers of cells can be scalably produced from the same set of individuals over a long period of time.

A major limitation in expanding this approach to different cell types is the lack of reliable differentiation protocols for many of them. Secondly, even if the protocols exist, differentiated cells will always show some differences from their primary counterparts and the consequences of these differences are largely unknown. Furthermore, many differentiation protocols are highly complicated, contain multiple manual steps and require many different signalling molecules to be added at specific time points. Progress has been made towards automating iPSC differentiation, but only a small number of protocols have been successfully converted (Paull et al., 2015).

Even though there is no theoretical limit to the number of cells that can be produced from iPSC differentiations, working with large numbers of cell considerably increases the cost and complexity of the experiments. Therefore, to make it feasible to study tens of different stimuli at multiple time points, the experimental assays need to scaled down to small cell numbers. Fortunately, progress has been made over the years in reducing the numbers of cells required by RNA-seq (Picelli et al., 2014), ATAC-seq (Corces et al., 2016) and ChIP-seq experiments (Lara-Astiaso et al., 2014).

6.2 Alternative transcription QTLs

It is clear that since the DNA does not leave the nucleus, the effect of GWAS variants on cellular and organismal phenotypes must be somehow mediated by RNA. The fact that only a small fraction of GWAS associations overlap coding sequence (Maurano et al., 2012) has led to a surge in gene expression QTL (eQTL) mapping studies. Although current eQTL mapping studies have found thousands of independent genetic variants associated with mRNA levels of different genes, the number of GWAS hits that can readily be explained by eQTLs has remained relatively modest. One possible reason might be that the disease-causing eQTLs are active only in very specific cell types and conditions that have not yet been profiled by current eQTL studies.

Alternatively, GWAS variants might influence RNA level phenotypes other than the total gene expression level such as alternative transcript usage. We and others (Li et al., 2016a) have shown that eQTLs and transcript ratio QTLs (trQTLs) are predominantly independent from each other. A trQTL study in lymphoblastoid cell lines (LCLs) found that trQTL enrichment in GWAS

hits was comparable to eQTLs (Li et al., 2016a). Similarly, rare variants causing aberrant mRNA splicing have been linked to Mendelian disorders (Cummings et al., 2016).

Alternative transcription can manifest in many different forms: alternative promoter usage, alternative splicing, alternative intron retention and alternative polyadenylation. In principle, if all possible alternative transcripts were annotated then all types of alternative transcription could be detected by quantifying transcript expression. There have been significant computational advances in recent years that have increased both the speed and accuracy of transcript expression quantification (Bray et al., 2016; Patro et al., 2016). However, as we have shown in Chapters 2 and 4, transcript annotations are still to a large degree incomplete. An alternative is to use approaches that rely less on reference transcript annotations and focus on reads mapping to exon-exon junctions instead. One such method is LeafCutter (Li et al., 2016b), but exactly because of its focus on junction reads it not able to detect changes to 5′ and 3′ untranslated regions or retained introns as we have shown in Chapter 4. On the other hand, using the reviseAnnotations tool developed in this thesis to split reference annotations into alternative 5′ and 3′ ends can be used to detect these events and approaches also exist to detect long 3′ UTRs *de novo* from RNA-seq data (Xia et al., 2014). An important area of future research will be to systematically analyse different types of alternative transcription events and characterise their genomic properties. Finally, combining better alternative transcription event annotations with RNA-seq data from hundreds of individuals will allow us to find trans-acting QTLs that regulate alternative transcription (Battle et al., 2014), thus providing new insights into the mechanisms of its regulation.

RNA transcripts consist of single long molecules. However, an important open question is how often different aspects of alternative transcription (i.e. alternative promoters, alternative exons, alternative 3′ UTRs) are regulated by shared mechanisms *versus* how often they are regulated by independent mechanisms. Preliminary results from Chapters 2 and 4 suggest that independent regulation might be the default mode of action. Future alternative transcription QTL mapping studies can answer this question by looking how often single QTLs are associated to single alternative transcription events as opposed to influencing multiple parts of the gene. Finally, direct long-read RNA sequencing has the potential to greatly improve reference transcript annotations (Garalde et al., 2016). However, if most alternative transcription events are regulated independently of the rest of the transcript then quantifying full transcript

expression for QTL mapping might actually reduce power, especially if the gene has multiple linked alternative transcription QTLs such as the IRF5 example highlighted in Chapter 4.

6.3 Information flow from DNA to protein

We and others have shown that there is considerable overlap between chromatin accessibility and gene expression QTLs. An early study in LCLs estimated that as many as 55% per cent of the eQTLs were also chromatin accessibility QTLs (caQTLs) but only 16% of the caQTLs were also estimated to regulate gene expression (Degner et al., 2012). In Chapter 5 we showed that in ~50% cases the caQTL underlying a condition-specific eQTL was already present in the naive state. Thus, a fraction of the discrepancy highlighted by (Degner et al., 2012) could be explained by 'primed' caQTLs that are waiting for the right environmental signal to start regulating gene expression. This observation illustrates an important concept where the propagation of regulatory effects from one level to the next (chromatin to RNA) can be regulated by changes in the environment that presumably influence the activity of trans-acting factors.

The situation is less clear for splicing and transcript ratio QTLs where we know less about what proportion are regulated at the chromatin level. While most variants disrupting canonical splice acceptor and donor sites and polyadenylation sites are unlikely to have any effect on the chromatin level, QTLs that influence alternative promoter usage could behave more like traditional eQTLs. Furthermore, there is evidence that DNA binding proteins such as CTCF can regulate splicing by influencing the pausing of RNA polymerase II (Shukla et al., 2011). Thus, this could be an interesting area of future research.

However, the functional unit for protein coding genes is the protein and not the mRNA molecule. Thus, it is important to know how genetic effects propagate from mRNA to protein level. Two of the largest joint protein QTL (pQTL) and eQTL mapping studies to date have been performed in human LCLs (Battle et al., 2015) and mouse liver (Chick et al., 2016). However, neither of these studies have looked at relationship between alternative transcription and protein expression level independent of the gene expression level. Since the role of 3′ and 5′ UTR sequences in regulating translation is well established (Wilkie et al., 2003), this could be an interesting area of future research. For example, re-analysing RNA-seq and proteomics data from (Chick et al., 2016) with splicing in mind might be a feasible starting point.

Another aspect that is completely unknown is if there is additional condition specificity on pQTL level beyond that observed at the mRNA level. For example, similarly to the constitutive caQTLs becoming eQTLs that we described in Chapter 4, it would be interesting to find constitutive eQTLs that become pQTLs after stimulation. If these eQTL-pQTL pairs do exist, a potential mechanism for them might come from the (Chick et al., 2016) study that identified an abundance of trans-acting pQTLs that were not present on the RNA level. They found that a large proportion of these QTLs could be explained by stoichiometric buffering whereby the expression level of a single protein in a larger complex influences the levels of other members of the same complex, probably because proteins bound in a complex are more stable than the unbound molecules. Thus a constitutive eQTL might become a pQTL in another condition when other members of the same complex are more highly expressed.

6.4 What are we going to do with all of the QTLs?

A major motivation for performing molecular QTL mapping studies is their potential to aid the interpretation of GWAS associations in order to identify causal genes and variants. However, even if a molecular QTL has been identified in the same region with a GWAS hit, it still remains challenging to distinguish a single shared causal variant driving both traits from two independent causal variants that are in high linkage disequilibrium. Although multiple statistical approaches have been developed to test colocalisation between associations (Giambartolomei et al., 2014; Zhu et al., 2016), they have limited power in regions with large number of variants, where it can be impossible to decide on the sharing of causal variants. The second challenge is pleiotropy, where the same causal variant influences too traits, but the traits themselves are not causally linked. For example, eQTLs can simultaneously regulate the expression of multiple gene at the same time. If the same causal variant is then associated with a complex trait then it might not possible to tell which gene mediates the GWAS associations based on statistical evidence alone.

Although deciding if a given molecular trait (such as gene expression) is causally linked to a complex disease is challenging based on a single association alone, we can be more confident if we see multiple associations pointing in the same direction. For example, multiple independent genetic associations with lower levels of low density lipoprotein (LDL) in blood are all linked to reducing cardiovascular disease risk (Ference et al., 2016). This association has also been confirmed in clinical trials, where the administration LDL-lowering drugs (such as

statins and PCSK9 inhibitors) has been shown to reduce cardiovascular disease risk. Thus, one paradoxical conclusion is that we need to discover even more QTLs to be able to take the full advantage of all the QTLs that we have found thus far.

However, even with larger studies we are unlikely to be able to characterise the function of all regulatory variants using QTL mapping approaches. This is especially true for rare variants and rare cell types that we do not know how to differentiate *in vitro*. Moreover, it is deeply unsatisfying if the only way we can predict the function of a non-coding genetic variant is to directly measure its activity experimentally. To achieve true understanding of the underlying biology, we need to be able to generalise from thousands of measured QTLs to new variants that have not been observed. Hence, in the long term, large QTL maps could provide us the necessary training data to build computational models that can predict the function of non-coding variants. In that respect, progress has recently been made to predict the effect of genetic variation on chromatin accessibility and transcription factor binding (Alipanahi et al., 2015; Kelley et al., 2016; Zhou and Troyanskaya, 2015). Progress has also been made building models to link enhancers to their target genes (Marbach et al., 2016; Whalen et al., 2016) and this is an area where large condition-specific eQTL maps can provide valuable training data.

6.5 From natural to engineered variation

In my thesis, I have used iPSC-derived cells to study the consequences of common natural genetic variation. However, another promising avenue of future research is studying the consequences of engineered genetic variation, especially because iPSCs can be readily genetically modified using the CRISPR technology. The first opportunity here is to use IPSCs to study the consequences of specific engineered mutations at several phenotypic levels and in many different cell types. The main advantage of iPSCs over primary cells is that iPSCs are self-renewing, meaning that it will be possible to construct clonal cell lines with specific engineered mutations in many different genetic backgrounds. These lines can then be shared and compared between different laboratories.

Another area where engineered genetic variation has a large potential are phenotypic screens. In this framework, a large library of mutant cells is first generated where each cell has a loss-of-function mutation in a single gene (or a regulatory element). The cells then go through either positive or negative selection, after which it is possible to determine which mutations had either

advantageous or deleterious effect on the phenotype. CRISPR screens have successfully identified genes required for cancer survival (Munoz et al., 2016) as well as genes important in innate immune response (Parnas et al., 2015). An advantage of iPSCs is that a wide range of phenotypes and cell types can be used for screening that are currently not available. This includes developmental processes; otherwise inaccessible cell types as well was artificial reporter constructs that can be introduced into the cells. Consequently, studying both natural and engineered genetic variation in iPSCs has a great potential to uncover the genetic architecture of a large variety of human traits.