Regulation of gene expression in macrophage immune response

Kaur Alasoo

Wellcome Trust Sanger Institute
Hughes Hall

University of Cambridge

This dissertation is submitted for the degree of Doctor of Philosophy
September 2016

Abstract

Regulation of gene expression in macrophage immune response

Kaur Alasoo

Gene expression quantitative trait loci (eQTL) mapping studies can provide mechanistic insights into the functions of disease-associated variants. However, many eQTLs are cell type and context specific. This is particularly relevant for immune cells, whose cellular function and behaviour can be substantially altered by external cues. Furthermore, understanding mechanisms behind eQTLs is hindered by the difficulty of identifying causal variants. We differentiated macrophages from induced pluripotent stem cells from 86 unrelated, healthy individuals derived as part of the Human Induced Pluripotent Stem Cells Initiative. We generated RNA-seq data from these cells in four experimental conditions: naïve, interferon-gamma (IFNɣ) treatment (18h), *Salmonella* infection (5h), and IFNγ treatment followed by *Salmonella* infection. We also measured chromatin accessibility with ATAC-seq in 31-42 individuals in the same four conditions. We detected gene expression QTLs (eQTLs) for 4326 genes, over 900 of which were condition-specific. We also detected a similar number of transcript ratio QTLs (trQTLs) that influenced mRNA processing and alternative splicing. Macrophage eQTLs and trQTLs were enriched for variants associated with Alzheimer's disease, multiple autoimmune disorders and lipid traits. We also detected chromatin accessibility QTLs (caQTLs) for 14,602 accessible regions, including hundreds of long-range interactions. Joint analysis of eQTLs with caQTLs allowed us to greatly reduce the set of credible causal variants, often pinpointing to a single most likely variant. We found that caQTLs were less condition-specific than eQTLs and ~50% of the stimulation-specific eQTLs manifested on the chromatin level already in the naive cells. These observations might help to explain the discrepancy between strong enrichment of diseases associations in regulatory elements but only modest overlap with current eQTL studies, suggesting that many regulatory elements are in a 'primed' state waiting for an appropriate environmental signal before regulating gene expression.

Declaration of Originality

This dissertation is the result of my own work and includes nothing which is the outcome of work done in collaboration except as declared in the beginning of each chapter. t is not substantially the same as any that I have submitted, or, is being concurrently submitted for a degree or diploma or other qualification at the University of Cambridge or any other University or similar institution. I further state that no substantial part of my dissertation has already been submitted, or, is being concurrently submitted for any such degree, diploma or other qualification at the University of Cambridge or any other University of similar institution. This dissertation does not exceed the word limit set by the Degree Committee for the Faculty of Biology.

Signature:

Date:

Kaur Alasoo
September 2016

Acknowledgements

First, I am grateful to many people who have inspired and nurtured my passion for science. I thank Hedi Peterson for drawing me into the field before I had even finished my first semester at the university. I thank Phaedra Agius for introducing me into machine learning and teaching me that it is completely normal in research not to know beforehand what works and what does not. I thank Prof. Jaak Vilo for supporting and guiding me during my undergraduate years, Harri Lähdesmäki for letting me play with new and exciting data, and Isabel Sá-Correia for giving me an independent experimental project with no prior experience. All of these experiences gave me the necessary skills and confidence to help me realize that research was what I wanted to do.

Many people have made the four years of my PhD studies such a rewarding and stimulating experience. Over the past four years, fair and constructive feedback from my supervisor Daniel Gaffney has greatly improved the clarity and focus of my research and writing. I also greatly value the independence that I had in shaping the projects that I was working on. I thank my co-supervisor Gordon Dougan for accepting me to his lab even though I had never seen neither human nor a bacterial cell before. Subhankar Mukhopadhyay was very patient in explaining both intricacies of immunology as well as basics of cell culture. Finally, most of the experimental work presented in this thesis would not have been possible without the dedication and hard work of Julia Rodrigues. Thank you, Julia, for not giving up even when cells were dying and differentiations failing. I will always remember the fruitful discussions with other members of the Gaffney on a wide range of experimental on computational topics.

I have received great advice from many other people on the campus. Leopold Parts has been a great mentor by listening to my ideas, thinking about them hard and asking important questions. Conversation with Leo have made me to reconsider my experimental design more than once, ultimately allowing me to answer more important questions.

Last but not least, I am most grateful to my family. Aitäh isale nende õhtuste matemaatika ja füüsika olümpiaadiülesannete lahendamiste eest, mis olid küll väga põnevad, aga millega ma ise alati esimesel katsel hakkama ei saanud. Aitäh vennale, õele, vanaemadele, vanaisadele, tädile ja teistele suurtele ja väikestele sugulastele pidevalt meelde tuletamast, kui hea on kodus olla. Aitäh, Maret, et oled minu kõrval olnud ja mind toetanud kogu selle aja.

Table of Contents

1 Introduction

Virtually all cell types in the human body contain exactly the same DNA. In spite of this, human cells exhibit extraordinary functional, morphological and molecular diversity. This diversity is particularly evident in the human immune system: B-cells specialise in producing antibodies while macrophages in different tissues are able to phagocytose and kill invading bacteria, to just illustrate two of the many cell types. In addition to each cell type exhibiting specific phenotype and function, they must also be plastic enough to respond to various changes in their environment. This is particularly important for immune cells that must repel invading viruses and bacteria while minimising damage to the host. For example, tissue macrophages must produce inflammatory cytokines and reactive oxygen species only when they detect bacteria but intestinal macrophages have to limit these responses to avoid reacting to commensal bacteria with excessive inflammation (Krause et al., 2015). Underlying these cell type specific functional differences are unique gene expression profiles that are precisely regulated in response to changes in the environment.

Most human traits and complex diseases have a heritable component (Visscher et al., 2008) and genome-wide association studies (GWAS) have identified thousands of genetic loci associated with those traits. Since over 90% of these loci are in the non-coding regions of the genome and highly enriched for chromatin marks specific to gene regulatory elements (Maurano et al., 2012), an emerging consensus is that they likely influence disease risk by regulating gene expression levels in one or more cell types and conditions. This observation in turn has led to a surge in studies to identify genetic variants that are associated with gene expression levels. While gene expression quantitative trait loci (eQTL) mapping experiments have identified thousands of regulatory variants, they have, to date, explained only a small fraction of GWAS associations and have also highlighted that considerable proportion of eQTLs are cell type and context specific. Thus, to create a complete catalogue of gene regulatory variation in humans, we need to measure gene expression levels in larger numbers of individuals, cell types and conditions.

However, constructing a comprehensive catalogue of human regulatory variation has been limited by the relative inaccessibility of most cell types and the large number of environmental stimuli potentially relevant for each cell type (Xue et al., 2014). However, scalable cell culture

systems based on human induced pluripotent stem cells (iPSCs) have the potential to overcome these limitations and identify functional regulatory variants in many more cell types and cell states. In this thesis, I will establish an iPSC-derived macrophage model to study the genetics of context specific gene expression and apply it to understand how genetics shapes gene expression in human macrophages in response to interferon-gamma stimulation and *Salmonella* infection.

In this introductory chapter, I will give an overview of our current understanding of the principles and mechanisms that regulate cell type and context specific gene expression by focussing on key studies performed in macrophages and B-cells. I will describe how macrophages sense and respond to changes in their environment and introduce experimental and computational techniques that are widely used to measure gene expression and chromatin state. Next, I will introduce iPSC-derived macrophages as a scalable system to study context specific gene expression. Finally, I will give an overview of how genetic variation influences gene regulation and how these studies can be used to interpret disease associations.

1.1 Regulation of cell type and condition specific gene expression

One of the first examples of gene expression controlled by environmental signals is the *lac* operon in *Escherichia coli* that contains three genes required for lactose import and metabolism (Jacob and Monod, 1961). The *lac* operon has two regulatory mechanisms. First, in the absence of lactose, lactose repressor protein strongly binds to a short DNA sequence downstream of the promoter and prevents the transcription of the operon. The second control mechanism is the catabolite activator protein that, in the absence of glucose, binds to a specific 16 base pair (bp) sequence upstream of the lac promoter and assists RNA polymerase binding to the DNA. Thus, the expression of the lac operon is highest when lactose is present in the environment and there is no glucose. This seminal study highlighted how sequence specific factors regulated by external signals can regulate gene expression.

The basic principle of sequence specific transcription factors (TFs) binding to DNA and thereby activating or repressing gene expression is also conserved in eukaryotes and many of the sequence motifs have already been identified (Weirauch et al., 2014). However, an extra layer of complexity is that, in contrast to prokaryotes, eukaryotic DNA is located in the nucleus and

tightly packed around the nucleosomes. This adds two additional levels of regulation. First, since protein synthesis happens in the cytoplasm, the localisation of TFs can be regulated as well. For example, the NF-κB complex is normally sequestered to the cytoplasm and is only localised to the nucleus after the repressor proteins have been degraded (Verma et al., 1995). Secondly, because nucleosomes have much stronger affinity for DNA than single TFs do, a single instance of a TF motif is usually not sufficient for a TF to bind (Polach and Widom, 1996). Recent studies have highlighted the importance of collaborative interactions between TFs in competing with nucleosomes and establishing active regulatory elements (Deplancke et al., 2016; Heinz et al., 2010).

1.1.1 Principles of cell type specific TF binding

Since gene expression is regulated by TFs, to understand cell type specific gene expression we first need to understand the principles of cell type specific TF binding. Genome-wide profiling of TF binding has led to three key observations: (1) different factors in the same cell type often bind to the same locations (MacArthur et al., 2009), (2) the same factor in different cell types can often have different binding sites (Odom et al., 2004) and (3) the same biological processes (such as self-renewal) can be regulated by distinct set of regulatory elements in different cell types (Soucie et al., 2016). To illustrate possible mechanisms behind these observations, I will now focus on PU.1 - a key TF required for both B-cell and macrophage differentiation *in vivo*, that shares approximately half of its binding sites between the two cell types (Heinz et al., 2010).

(Heinz et al., 2010) sought to identify what underlies the cell-type specific binding pattern of PU.1. They found that macrophage specific PU.1 binding sites were co-enriched for AP-1 and C/EBPβ motifs, two additional factors that are required for macrophage development and function (Friedman, 2007). Conversely, B-cell specific PU.1 binding sites were enriched for motifs of E2A, EBF1 and OCT2 - three factors that are known to play important roles in B-cell development and function (Medina and Singh, 2005). Furthermore, they showed that knock-out of E2A leads to loss of PU.1 in B-cells at sites where the E2A motif is present and that can be rescued by inducible expression of E2A in knock-out cells. Similarly, PU.1 knock-out in macrophages led to reduced binding of C/EBPβ at loci where both of the binding sites were present. Together, this evidence indicates that cell type specific enhancers are established by collaborative binding of a small number of cell type specific pioneer TFs that are able to compete with the nucleosomes.

The second line of evidence to support this model of collaborative binding of cell type specific pioneer TFs comes from a follow-up study of macrophage enhancers in two genetically distinct inbred mouse strains (Heinz et al., 2013). They found that PU.1 motif mutations in one strain resulting in strain-specific loss of PU.1 binding were frequently associated with corresponding loss of C/EBPα binding. Conversely, they also found that mutations in the C/EBP motif leading to the loss of C/EBPα binding were similarly associated with the loss of PU.1 binding.

1.1.2 Signal dependent TFs bind to established enhancers

A second key observation is that although different cell types often respond to the same extracellular signal by activating the same signalling pathways and TFs, the binding sites that these TFs occupy are often cell type specific. One proposed mechanism that could explain this observation is that TFs activated by external signals may largely bind to enhancers that have been previously established by cell type specific pioneer TFs. Some of the evidence for this comes from an early study which found that 34% of the oxysterol-responsive nuclear receptor Liver X Receptor beta (LXRβ) binding sites colocalised with PU.1 binding sites in macrophages and LXRβ binding was reduced at these sites in PU.1 deficient cells (Heinz et al., 2010). On the other hand, PU.1 binding at these sites was not affected by LXRβ knock-out, indicating that LXRβ is not directly involved in establishing cell type specific enhancers.

In a follow up study, Heinz *et al* (Heinz et al., 2013) used two genetically distinct inbred mouse strains to study the strain specific binding of NF-κB after TLR4 activation. They found that 61% of NF-κB binding sites in the activated cells were already bound by either PU.1 and/or C/EBPα in the naive condition. Furthermore, most strain-specific NF-κB binding sites were bound by PU.1 or C/EBPα only in the strain that showed NF-κB binding. Finally, they were able to attribute 34% of strain-specific NF-κB binding events to mutations in AP-1, PU.1 or C/EBPα binding motifs and only 9% to mutations in NF-κB binding motifs. These observations suggest that the landscape of NF-κB binding sites after TLR4 activation are largely predetermined by enhancers occupied by PU.1, AP-1 or C/EBPα TFs in the naive state where no active NF-κB is present in the nucleus.

In summary, these studies highlight a hierarchy between cell type specific pioneer factors that establish enhancers in closed chromatin regions and TFs activated by external signals that

predominantly bind to pre-established enhancers. Similar results have also been described for TGFβ (Mullen et al., 2011), BMP and Wnt pathways (Trompouki et al., 2011).

1.1.3 Role of signal dependent TFs in establishing new enhancers

While most signal-dependent TF binding occurs at pre-established enhancers, Ostuni *et al* showed that up to 15% of the enhancers activated by LPS were undetected in the unstimulated cells (no PU.1 binding or H3K4me1 histone modification signal) (Ostuni et al., 2013). They referred to these elements as latent enhancers and they found that different stimuli each activated a distinct set of latent enhancers. To mechanistically study the latent enhancers they focussed on IFNɣ stimulation. They found that, although STAT1 was phosphorylated within 10 minutes after IFNɣ stimulation, latent enhancers were only established hours after stimulation, suggesting that nucleosomes might act as a barrier inhibiting TF binding. They observed that although many latent enhancers contained PU.1 binding motifs and displayed PU.1 binding after stimulation, there was no PU.1 binding in the naive state. Furthermore, they found that PU.1 motifs in the latent enhancers had considerably lower binding affinities than motifs in constitutive enhancers, indicating that PU.1 binding at these sites depended on stimulus-specific cofactors. Thus, while the hierarchical enhancer activation model is conceptually useful, signal dependent TFs can also facilitate the eviction of nucleosomes and the binding of cell type specific TFs. One apparent distinction between these different modes of regulation, as illustrated by the IFNɣ example, is that pre-existing enhancers can facilitate cellular responses on the order of minutes while remodelling nucleosomes can take hours.

1.1.4 Long range interactions between cell type specific and signal dependent TFs

The evidence presented so far has relied on two different types of experimental approaches. The first relied either on deleting or ectopically expressing specific TFs and looking at the effects of these changes on the binding profiles of other TFs. The second approach relied on subtler perturbations caused by segregating variants disrupting TF binding sites between different mouse strains. However, because both of these approaches resulted in changes to thousands of TF binding events, they were limited to looking at average genome-wide effects on overlapping regulatory elements and were not able to reliably identify if TF binding at any one specific locus affected TF binding at other regulatory elements further away. Detecting these

individual effects can be achieved by QTL mapping approaches or directly disrupting single TF binding sites by precise genome editing.

Evidence that cell type specific TFs can influence the binding of signal-induced TFs at neighbouring enhancers comes from an elegant study of an enhancer cluster upstream of the WAP gene in mouse mammary tissue (Shin et al., 2016). The enhancer cluster consists of three elements E1, E2 and E3 and the 1000-fold induction of the WAP gene during mouse pregnancy depends on all of them. The E1 enhancer has binding sites for three TFs: ELF1, NFIB and STAT5A. STAT5A binding can be observed at E1 during early pregnancy prior to transcriptional activation of the WAP gene. However, WAP transcription is induced only after STAT5A is also bound at the E2 and E3 enhancers. Intriguingly, the authors found that jointly disrupting ELF1, NFIB and STAT5A binding sites in the E1 enhancer not only abolishes the enhancer, but also prevents the E2 and E3 enhancers from being established later during pregnancy and, in turn, the gene from being transcriptionally activated. Thus, the E1 enhancer contains binding sites for tissue-specific TFs ELF1 and NFIB and acts as a 'seed' enhancer for the neighbouring E2 and E3 enhancers that only contain binding sites for STAT5A.

In summary, the DNA in eukaryotic cells is tightly wrapped around the nucleosomes and collaborative interactions between multiple TFs are often needed to evict nucleosomes and establish accessible chromatin. Overlapping sets of TFs are often expressed in multiple cell types (such as PU.1 in B-cells and macrophages) and cell type specific binding is achieved by regulating the expression level of individual TFs as well as the pool of available cofactors. Transcription factors activated by multiple signalling pathways (IFNɣ, TLR4, TGFβ, Wnt, etc.) predominantly bind to regulatory elements pre-established by cell type specific factors, although over prolonged periods of time they might also contribute to establishing new enhancers. The extent of this is likely to depend on the exact TFs being activated and their intrinsic ability to compete with nucleosomes (Romanoski et al., 2015). Finally, as the example of the WAP gene suggests, TF binding at one locus can also facilitate the binding of TFs at other regulatory elements multiple kilobases (kb) away. The mechanisms by which this happens have not yet been elucidated.

1.2 Macrophage biology in the context of immune response

Macrophages are key phagocytic cells associated with innate immunity, pathogen containment and modulation of the immune response (Murray and Wynn, 2011; Wynn et al., 2013). Macrophages have multiple receptors to recognise pathogen-associated molecular patterns such as toll-like receptors (TLRs), nod-like receptors (NLRs) and RIG-i like receptors (Mogensen, 2009). Macrophages also respond to regulatory signals produced by other cells such as interferon-gamma (IFNɣ), interferon-beta (IFNβ), interleukin-4 (IL-4), interleukin-10 (IL-10), tumour necrosis factor (TNF) and many others (Xue et al., 2014). In the following section I will give a more thorough overview of macrophage response to bacterial lipopolysaccharide, IFNɣ and *Salmonella* infection, because these three stimuli are the main focus of the rest of the thesis.

1.2.1 Signalling pathways activated by lipopolysaccharide and interferon-gamma

Lipopolysaccharides (LPS) are a component of the outer membrane of gram-negative bacteria. Macrophages recognise LPS via the TLR4 receptor on their cell surface (Medzhitov and Horng, 2009). Ligand binding to TLR4 leads to the activation of the Myd88 dependent pathway that culminates with the activation of NF-κB and AP-1 transcription factors that recognise specific sequence motifs in the nucleus (Takeuchi and Akira, 2010) (Figure 1.1). This pathway is also shared with other toll-like receptors such as TLR2, TLR3 and TLR9. In addition, TLR3/4 activation also leads to the activation of Myd88-independent pathway culminating with the activation of interferon response factors 3 and 7 (IRF3/7) transcription factors that recognise the canonical interferon-response element (ISRE) motif (Doyle et al., 2002).

One of the genes activated by IRF3/7 is IFNB1 that codes for IFNβ protein (Doyle et al., 2002). IFNβ is secreted by the cells where it is then recognised by interferon-alpha receptor (IFNAR). Activation of IFNAR predominantly leads to activation of the ISGF3 complex composing of STAT1, STAT2 and IRF9 that recognises the same ISRE motif (Ivashkiv and Donlin, 2014).

Interferon-gamma (IFNɣ) is an inflammatory cytokine produced by T-cells and natural killer (NK) cells (Schroder et al., 2004). IFNɣ binding to the IFNɣ receptor leads to the phosphorylation of STAT1 and formation of STAT1 homodimers that bind to the gamma-activated sequence (GAS) motif (Platanias, 2005). One of the immediate targets of STAT1 is IRF1 transcription factor that

is involved in the cooperative regulation of gene expression of many target genes (Ramsauer et al., 2007) including the master regulator of major histocompatibility complex (MHC) class II genes CIITA (Reith et al., 2005).

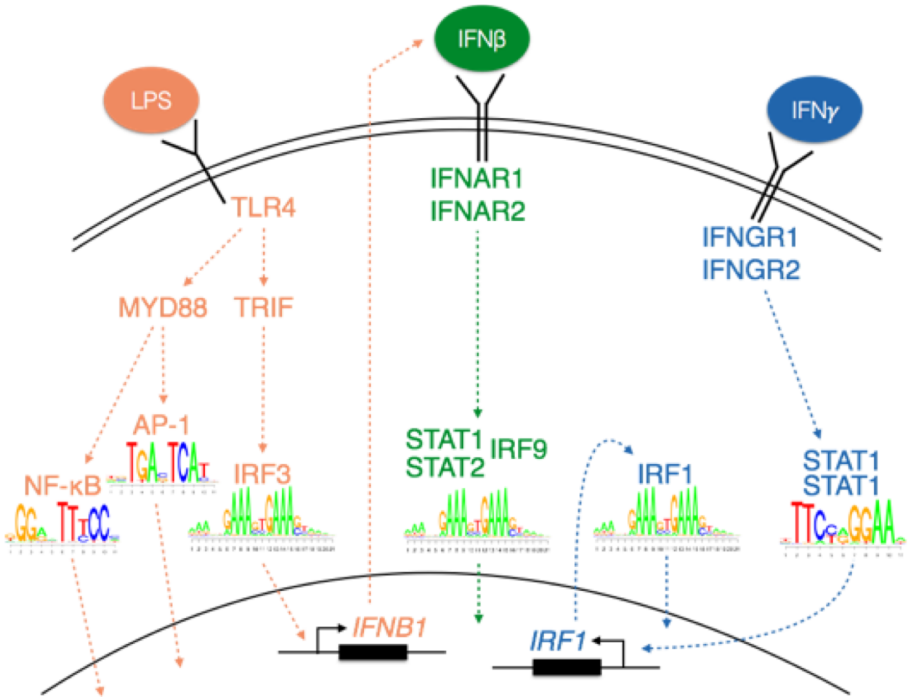

Figure 1.1: Main signalling pathways activated in macrophages after *Salmonella* infection and IFNɣ stimulation. Macrophages recognise LPS on the *Salmonella* cell wall via the TLR4 receptor (Medzhitov and Horng, 2009). Ligand binding to multiple TLRs such as TLR2, TLR3, TLR4 and TLR9 leads to downstream activation of NF-κB and AP-1 transcription factors (Takeuchi and Akira, 2010). However, TLR3/4 activation also leads to specific activation of the IRF3 transcription factor and downstream antiviral response genes (Doyle et al., 2002). IFNɣ, on the other hand, activates signal transducer and activator of transcription 1 (STAT1) and IRF1 TFs.

Thus, different environmental signals lead to the activation of distinct signalling pathways and downstream TFs that are responsible for specific changes in gene expression (Xue et al., 2014). Furthermore, simultaneous activation of multiple signalling pathways can have synergistic effects on gene expression, leading to activation of genes that are not activated by either of the stimuli alone (Qiao et al., 2013).

1.2.2 Macrophage response to *Salmonella* infection

Macrophages recognise many different components of *Salmonella* including LPS (TLR4), flagella (TLR5), fimbrae/pili, peptidoglycan (TLR1/2, NOD2), bacterial DNA (TLR9) and type III secretion systems (T3SS) (NLRC4) (de Jong et al., 2012). In addition, *Salmonella* can also directly modulate macrophage immune response by releasing effector molecules encoded via the type III secretion systems that can promote bacterial uptake and intracellular survival (Haraga et al., 2008).

Salmonella infection and LPS stimulation induce similar transcriptional response in mouse macrophages (Rosenberger et al., 2000), suggesting that LPS plays an important role in early response to bacterial infection (4 hours). Similarities between *Salmonella* and LPS response have also been observed in human macrophages where the core transcriptional response was conserved between many different species of bacteria and bacterial components (such as LPS) and this response was predominantly mediated by TLR4 and TLR2 signalling (Nau et al., 2002). This is not to say that differences in response between live bacterial infections and LPS stimulation do not exist. For example, *Mycobacterium tuberculosis* is able to actively suppress interleukin-12 (IL12) production (Nau et al., 2002). Rather, it suggests that in common experimental designs of bulk infections (resulting in only 20-30% of macrophages being infected) early response (the first few hours) is dominated by TLR signalling and other signalling mechanisms have either weaker effects or influence smaller proportion of cells. Single cell RNA-seq is a promising approach to address this question.

1.3 Tissue culture models of macrophage biology

Commonly used model systems to study macrophage biology have included macrophage-like leukemic cell lines such as THP-1 (Tsuchiya et al., 1982), primary macrophages derived from model organisms and primary human macrophages differentiated from blood monocytes. Although these cells have provided important insights into macrophage-associated biology, they have some limitations. Immortalised cell lines often have accumulated multiple genetic aberrations and can exhibit functional defects compared to primary cells such as impaired cytokine production upon LPS stimulation (Adati et al., 2009; Schildberger et al., 2013), while multiple functional differences exist between macrophages from different species (Schroder et al., 2012). Additionally, human monocyte derived macrophages (MDMs) can be difficult to

obtain in sufficient numbers for repeated experimental assays and it is currently challenging to introduce targeted mutations into their genomes, limiting their utility in genetic studies.

1.3.1 Differentiating macrophages from human induced pluripotent stem cells

A promising alternative approach is to differentiate macrophages directly from human induced pluripotent stems cells (iPSCs). The key advantage of the iPSC-based system is that it is possible to produce large numbers of cells from almost any genetic background (both natural and engineered), provided that the genetic background does not interfere with macrophage differentiation itself. The simpler protocol that we have used throughout this thesis relies on spontaneous formation of embryoid bodies (EBs) followed by directed differentiation in the presence of interleukin-3 (IL-3) and macrophage colony stimulating factor (M-CSF) (Karlsson et al., 2008; Lachmann et al., 2015; van Wilgenburg et al., 2013). Alternative approaches avoid the EB formation step and directly differentiate macrophages from pluripotent stem cells using a combination of multiple factors (BMP4, VEGF, SCF, TPO, Flt3, bFGF, M-CSF) (Yanagimachi et al., 2013; Zhang et al., 2015).

Early studies established that macrophages differentiated from induced pluripotent stem cells (IPSDMs) recapitulated many aspects of primary macrophage biology. They exhibited a transcriptomic signature specific to myeloid cells and expressed many macrophage specific cell surface markers including CD14, CD16, CD206 and CD68 (Karlsson et al., 2008; van Wilgenburg et al., 2013). In addition, IPSDMs were able to endocytose low-density lipoprotein (LDL), phagocytose opsonised yeast particles, produce specific cytokines in response to LPS stimulation and respond differentially to IFNɣ and IL-4 stimulation (Karlsson et al., 2008; van Wilgenburg et al., 2013). Patient-derived IPSDMs have successfully been used to model many monogenic disorders such as chronic granulomatous disease (Jiang et al., 2012) and Tangier disease (Zhang et al., 2015). However, at the outset of this work it was not yet clear how similar were IPSDMs to MDMs on the transcriptome level.

1.4 Genome-wide profiling of gene expression and chromatin accessibility

1.4.1 RNA sequencing

RNA sequencing (RNA-seq) is a widely used method to measure genome-wide gene expression profiles (Marioni et al., 2008). Since the majority of the RNA in most cells is ribosomal, either ribosomal RNA (rRNA) depletion or poly-A pulldown is often used to enrich for messenger RNA, after which the RNA is fragmented, reverse transcribed, PCR-amplified and sequenced using short read technologies. Each step in the workflow can introduce its own set of biases, some of which have been quite well characterised. For example, rRNA depletion can lead to large variation in read coverage across gene bodies while poly-A pulldown tends to introduce 3' bias (Lahens et al., 2014). On the other hand, PCR often preferentially amplifies sequences with higher GC content in a manner that varies from sample to sample (Benjamini and Speed, 2012). Finally, RNA fragmentation process can lead to preferential sequencing of fragments with specific start and end positions (Roberts et al., 2011a) i.e. fragment start and end positions are not uniformly distributed across exons. While 3' bias can often be minimised experimentally by ensuring that the RNA is intact before sequencing, multiple computational approaches have been developed to estimate and correct for GC-content and fragment biases (Benjamini and Speed, 2012; Hansen et al., 2012; Roberts et al., 2011a).

Quantifying gene expression levels

The first step in RNA-seq analysis is the quantification of gene expression levels. This has traditionally been done by first aligning reads to the reference genome using a splice-aware short read aligner that is able to also align reads across known and novel splice junctions. One of the first splice-aware aligners was TopHat (Trapnell et al., 2009), but it has since been surpassed both in speed and accuracy by newer aligners such as STAR (Dobin et al., 2013) and HISAT (Kim et al., 2015). After alignment, reads overlapping known gene annotations from databases such as GENCODE (Harrow et al., 2012) can be counted using multiple available tools such as featureCounts (Liao et al., 2014) or HTSeq (Anders et al., 2015). Reference genome alignments are also useful for visualising read coverage across the gene body.

Quantifying alternative transcription

Many human genes express multiple alternative transcripts that can differ from each other in terms of function, stability or subcellular localisation of the protein product (Carpenter et al., 2014; Wang et al., 2008). Considering expression only at a whole gene level can hide some of these important differences. Alternative transcription includes alternative promoter usage, alternative splicing, where middle exons are selectively included or excluded, and alternative polyadenylation. Two complementary approaches are often used to quantify changes in alternative transcription. One approach is to estimate the relative expression levels of all known transcripts of the gene that can best explain the observed RNA-seq read patterns across the gene body. The first methods that adopted this strategy were Flux Capacitor (Montgomery et al., 2010), MISO (Katz et al., 2010) and cufflinks (Roberts et al., 2011b; Trapnell et al., 2013). These were later improved upon by more accurate methods such as mmseq (Turro et al., 2011) and BitSeq (Glaus et al., 2012) that outperformed their predecessor on independent benchmark datasets (Kanitz et al., 2015). A major limitation of these methods has been their computational complexity that can prevent them from being applied to studies with large numbers of samples. Newer quantification methods such as Sailfish (Patro et al., 2014), kallisto (Bray et al., 2016) and Salmon (Patro et al., 2016) omit the explicit reference genome alignment step and quantify gene expression levels directly using transcriptome sequences. This has been shown to dramatically reduce the time required for quantification.

Even though the computational requirements have largely been resolved, important biological challenges still remain. First, genes often have multiple annotated transcripts that only differ from each other by a small amount of sequence, making it challenging to accurately estimate their expression from short read sequencing data. Secondly, many transcript annotations in the most comprehensive Ensembl database (Yates et al., 2016) are still incomplete and have either their 3’ or 5’ ends missing. Finally, many genes still have missing transcripts that have not been annotated. For example, a long gene might have three alternative promoters, two alternatively spliced exons and four alternative 3’ ends. If we make the assumption that most of these events are regulated independently, then this gene should have 2*3*4 = 24 alternative transcripts, but usually only a subset of these are present in the database. The assumption of independence is not completely unrealistic, because for example promoter selection and alternative splicing are regulated by independent molecular mechanisms (Barash et al., 2010).

A commonly used alternative analysis is to ignore the full transcript annotations and try to identify individual alternative transcription events independently. Two of the pioneers of this approach were DEXSeq (Anders et al., 2012) and MISO (Katz et al., 2010). DEXSeq aims to identify individual exons that are differentially expressed within a gene and as a result does not require the alternative exons to be previously annotated. MISO estimates the relative expression of alternative transcription events consisting of annotated alternative exons and their neighbouring exons. As a result, it is limited to annotated alternative exons but it can also take advantage of informative reads mapping to exon-exon junctions that are ignored by DEXSeq. Finally, LeafCutter (Li et al., 2016b) detects and quantifies clusters of alternatively excised introns directly from the read alignments by focussing on reads mapping to exon-exon junctions. In principle, this can be done without using reference transcript annotations, although in practice reference transcripts are usually still used during the read alignment phase to aid the detection of exon-exon junctions.

Quantifying allele-specific expression

In addition to total gene expression level, RNA-seq data can also provide information about the relative expression of the gene from the maternal and paternal chromosomes. This is possible when an individual is heterozygous at sites within the gene body, making it possible to count the number of RNA-seq reads that come from each allele. Allele-specific expression has been shown to increase the power to detect gene expression quantitative trait loci (eQTLs) (van de Geijn et al., 2015; Kumasaka et al., 2016). However, a major challenge is reference mapping bias - reads containing the non-reference allele can be less likely to be mapped than reads containing the reference allele. This is because read alignment algorithms penalise mismatches and reads containing the alternative allele will have at least one mismatch by definition. The simplest approach is to use a set of *ad hoc* rules to filter out variants that are likely to exhibit strong reference bias (Castel et al., 2015). A second approach is to deal with the issue at the time of read alignment either by using personalised reference genomes (Rozowsky et al., 2011) or editing the reads (van de Geijn et al., 2015). Finally, it is possible to use computational methods such as RASQUAL (Kumasaka et al., 2016) that explicitly model reference mapping bias.

1.4.2 Chromatin state profiling

As highlighted above, gene expression is predominantly regulated by the binding of transcription factors (TFs) to the promoters and distal regulatory elements. TF binding to a specific site often

leads to increased chromatin accessibility at the site as well as to covalent modification of nearby histones (Henikoff and Shilatifard, 2011). Hence, TF binding can be measured either directly using ChIP-seq or indirectly by measuring the levels of histone modifications (ChIP-seq) or chromatin accessibility (DNAse-seq (Furey, 2012), ATAC-seq (Buenrostro et al., 2013)) at the locus.

ChIP-seq

Chromatin immunoprecipitation followed by sequencing is a technique to identify the binding locations of specific proteins on the DNA (Furey, 2012). It is commonly used to detect the DNA binding locations of either TFs or modified histones. In ChIP-seq, proteins are first crosslinked to the DNA using formaldehyde, the DNA is then sheared and antibodies against a specific protein are used to selectively enrich for fragments that are bound by the protein of interest. Finally, the fragments are constructed into a library and sequenced.

Chromatin accessibility

The classical method to locate accessible chromatin regions has been DNAse I digestion followed by sequencing (DNAse-seq) (Bell et al., 2011). However, a major limitation of DNAse-seq has been its requirement for large numbers of cells and laborious and complicated experimental protocols. Consequently, most existing DNAse data has been generated by large-scale projects such as ENCODE (Neph et al., 2012) and Roadmap Epigenomics (Roadmap Epigenomics Consortium et al., 2015) in a small number of labs. This has changed recently with the introduction of ATAC-seq technique, which can be reliably performed even at the single cell level, and takes only a single day to complete (Buenrostro et al., 2013, 2015). ATAC-seq relies on Tn5 transposase that is used to insert Illumina sequencing adaptors into native chromatin. When Tn5 is used on intact nuclei this results in sequencing adaptors being preferentially integrated into regions of accessible chromatin.

Data analysis

After the reads have been aligned to the reference genome, the first steps is identifying regions (‘peaks’) that show either more protein binding or chromatin accessibility than the genome-wide background. Many different peak calling algorithms exist, but one commonly used method is MACS2 (Zhang et al., 2008b). Once the regions have been identified, we can quantify total and allele-specific signal using the same approaches that are used for RNA-seq data.

1.5 Genetics of molecular traits

Genome wide association studies (GWAS) have identified thousands of genetic variants associated with various human traits and diseases. For example, as of 12 June 2016 the NHGRI-EBI GWAS catalog contains 21,941 unique variant-trait associations from 2457 studies (Welter et al., 2014). These variants lie predominantly in non-coding regions of the genome, making it difficult to identify the gene that is being affected as well as the relevant tissue and cell type for the disease (Maurano et al., 2012). However, GWAS variants are also enriched in gene regulatory elements (Farh et al., 2014; Maurano et al., 2012; Trynka et al., 2013) with different traits often showing enrichments in specific cell types and tissues, suggesting that many of the GWAS variants act by regulating the expression level of some nearby genes.

Moreover, emerging evidence suggests that the gene closest to the GWAS variant is not necessarily regulated by it. For example, a variant in the first intron of the FTO gene that has been associated with body mass index was only recently found to regulate the expression of IRX3 and IRX5 genes that are up to 1 Mb away from the variant (Claussnitzer et al., 2015). These long-range interactions can be quite common, as illustrated by a recent joint analysis of GWAS summary statistics for multiple traits and blood eQTL data from 5,311 individuals (Zhu et al., 2016). They identified 126 genes where the GWAS signal and eQTL signal where consistent with a shared causal variant, and found that in ~60% of the cases the regulated gene was not the one closest to the lead GWAS variant. Hence, for variants that are further away from genes, distance might not be reliable, and additional information is necessary to identify the most likely target genes. One promising approach for linking GWAS hits to their target genes has been eQTL mapping studies. Intuitively, if the same genetic variant is associated with both the expression level of gene A and the risk of disease B then this can provide a hypothesis that the genetic variant might influence disease B via gene A.

1.5.1 Genetics of gene expression

Large-scale eQTL mapping studies have revealed that common variants regulating gene expression are ubiquitous. One of the largest human studies involving whole blood RNA-seq data 922 individuals identified at least one eQTL for 79% of the genes with quantifiable expression level (Battle et al., 2014). However, it remains unclear why most of these variants do not seem to have deleterious effects on organismal fitness. One possibility is that many of the eQTLs are buffered at the protein level. In support of this theory, shared eQTLs and protein

QTLs (pQTLs) identified in human lymphoblastoid cell lines (LCLs) tend to have smaller effect sizes on the protein level (Battle et al., 2015). Similar buffering effects have also been observed for pQTLs identified in *Arabidopsis* (Fu et al., 2009) and mouse (Chick et al., 2016; Ghazalpour et al., 2011). Alternatively, high variability in the expression levels of some genes might be tolerated without significant effect on the organismal fitness (Keren et al., 2016).

Early on, it was identified that genetic variation influences gene expression in a cell type specific manner. Gene expression QTL mapping in three human tissues (adipose tissue, skin and LCLs) showed that on average 29% of the local eQTL were tissue-specific with substantial variation of sharing between different tissues (Nica et al., 2011). This has led to multiple individual eQTL mapping studies in various human cell types (monocytes (Fairfax et al., 2012), neutrophils (Naranbhai et al., 2015), B-cells (Fairfax et al., 2012), T-cells, to name a few) as well as large-scale consortium efforts such as the Genotype-Tissue Expression (GTEx) (The GTEx Consortium, 2015) project that aims to perform RNA and genome sequencing on 44 tissues collected from up to 500 post-mortem donors. The relatively high cell type specificity of eQTLs is perhaps unsurprising given that patterns of TF binding that regulate gene expressions are highly cell type specific as highlighted above and even the same biological processes can be regulated by distinct sets of regulatory elements in different cell types (Soucie et al., 2016).

However, an aspect that has gotten relatively less attention is that genetic effects can also be modulated by the environment that the cells are in. Early on, Smith and Kruglyak showed that many eQTLs in yeast were specific to the environment that the cells were grown in (ethanol *versus* glucose) (Smith and Kruglyak, 2008). Similar condition-specific genetic effects were later observed in mouse macrophages stimulated with either LPS or oxidized phospholipids (Orozco et al., 2012). The first human studies were performed on LCLs stimulated with glucocorticoids (N=114) (Maranville et al., 2011) and primary dendritic cells (N=65) infected with *Mycobacterium tuberculosis* (Barreiro et al., 2012). These have been followed by several studies involving different immune cells and additional stimuli (Table 1).

Table 1: Selection of eQTL studies looking at gene-environment interactions in stimulated human cells.

Study	Cell type	Stimulations	Sample size

(Maranville et al., 2011)	Lymphoblastoid cell lines (LCLs)	Glucocorticoids	114 individuals
(Barreiro et al., 2012)	Dendritic cells	*Mycobacterium tuberculosis*	65 individuals
(Fairfax et al., 2014)	Monocytes	LPS (2h), LPS (24h), IFNɣ (24h)	261-414 individuals
(Lee et al., 2014)	Dendritic cells	LPS (5h), influenza (10h), IFNβ (6.5h)	534 individuals
(Kim et al., 2014)	monocytes	LPS (1.5h)	137 individuals
(Çalışkan et al., 2015)	Peripheral blood mononuclear cells (PBMCs)	Rhinovirus infection	98 individuals

This area is still relatively underexplored given that for each human cell type there could be tens of relevant individual stimuli or combinations of stimuli that can modulate the effects of genetic variants on gene expression. Furthermore, the effect of a single stimulus can depend on the time when it was measured (Fairfax et al., 2014), thus increasing the number of relevant experimental conditions even further. With that many experimental conditions, obtaining enough cells from controlled genetic backgrounds becomes a major challenge. However, if efficient differentiation protocols are available, then iPSCs can be used to produce large numbers of differentiated cells from any cell type.

1.5.2 Genetics of chromatin states

A major limitation of eQTL mapping studies is that due to linkage disequilibrium we are mostly unable to identify the single most likely causal variant. This can severely hamper our ability to understand the principles of gene regulation and, as a consequence, means that even if we have a strong evidence of co-localisation between GWAS hit and an eQTL we might still not understand the molecular mechanism that gives rise to both of the traits.

A promising approach is to use the same QTL mapping approach to search for genetic variants that are associated with the activity of regulatory elements (i.e. regulatory QTLs). An advantage of regulatory QTLs is that they often reside within the same regulatory element, making it easier to predict the most likely causal variant (Degner et al., 2012; Ding et al., 2014). The activity of regulatory elements can be characterised by either measuring the levels transcription factor (TF) binding, histone modifications (both measured by ChIP-seq) or chromatin accessibility (measured by DNase-seq or ATAC-seq). Until recently, all of these approaches were limited by either complicated experimental protocols and/or the requirement of large number of cells, making it feasible to perform regulatory QTL mapping experiments only in LCL and in relatively small number of individuals. This has changed with the introduction of ATAC-seq technique that can be reliably performed on as few as 5,000 cells and takes only a single day to complete (Buenrostro et al., 2013).

TF binding as measured by ChIP-seq is the most specific measurement, but this also means a separate experiment needs to be performed for each TF of interest. In addition, not all TFs have reliable ChIP-seq antibodies available and generally a large number of cells are required for a successful experiment (>10 million). Profiling the levels of histone modifications hides the identity of specific TFs, but can still reveal if the regulatory element is in a repressed, poised or active state. Finally, DNase-seq or ATAC-seq only reveal which regions of the chromatin are open or closed, but require only a single experiment, and in the case of ATAC-seq work on a very small number of cells and generally have higher resolution than histone ChIP-seq experiments. A selection of recent chromatin QTL studies is presented in Table 1.2.

Table 1.2: summary of recent chromatin QTL mapping studies.

Study	Cell type	Phenotype	Sample size
(Kasowski et al., 2010)	LCL	NF-κB ChIP-seq RBP2 (Pol II) ChIP-seq	10 individuals
(Degner et al., 2012)	YRI LCL	DNAse-seq	70 individuals
(Kasowski et al., 2013)	LCL	H3K27ac, H3K4me1, H3K4me3, H3K36me3, and H3K27me3 CTCF SA1 (cohesin subunit)	19 individuals
(Kilpinen et al., 2013)	LCL	Histones: H3K4me1, H3K4me3, H3K27ac, H3K27me3 TFs: TFIIB, PU.1, and MYC RPB2 (Pol II)	2 trios + 8 individuals (subset of assays)
(McVicker et al., 2013)	YRI LCL	H3K4me1, H3K4me3, H3K27ac, and H3K27me3 Pol II	10 individuals
(Ding et al., 2014)	CEU LCL	CTCF ChIP-seq	51 individuals
(Kumasaka et al., 2016)	CEU LCL	ATAC-seq	24 individuals
(Grubert et al., 2015)	YRI LCL	H3K4me1, H3K4me3, H3K27ac	75 individuals
(Waszak et al., 2015)	CEU LCL	PU.1, RBP2 (Pol II) H3K4me1, H3K4me3, H3K27ac	47 individuals

1.5.3 Using eQTLs to interpret GWAS associations

If the same genetic variant is associated both with expression level of gene A and increased risk of disease B then this can provide a mechanistic hypothesis that the expression level of gene A influences the risk of disease B. However as highlighted above, eQTLs are extremely common and because of strong LD between variants there is often a large number of variants that are significantly associated with either gene expression level and/or disease risk. As a result, it is easy to get random overlaps between eQTLs and GWAS hits where the two associations are driven by different causal variants.

To overcome this limitation, different approaches have been developed that compare the association patterns of two traits across many variants and try to identify if they are likely to be driven by the same causal variant. Although the amount of molecular QTL studies has been steadily increasing, the number GWAS hits that can be readily explained by eQTLs has still remained relatively small. A study of 49 type 1 diabetes loci and monocyte eQTLs from 1,370 individuals identified 21 cases where the data was consistent with a shared causal variant driving both traits (Wallace et al., 2012). However, when a newer Bayesian colocalisation test (Giambartolomei et al., 2014) was applied to ten immune-mediated diseases and gene expression data from multiple immune cell types, it was able to identify only six confident colocalised associations (Guo et al., 2015). This is an active area of research and newer methods are continuously being developed and applied to ever larger data sets (Chun et al., 2016; Hormozdiari et al., 2016; Zhu et al., 2016).

Multiple factors might be responsible for the limited success of using eQTLs to interpret GWAS hits. One possible reason is that the disease relevant eQTLs might be active in very specific cell types and conditions and the limited eQTL studies that have been performed thus far have been unable to uncover them. Another reason is that if there are many variants that are in high LD with the causal variant, then even if the two traits have almost identical association profiles it is statistically impossible to distinguish if they are likely to be driven by the same causal variant or two different causal variants (Zhu et al., 2016). Finally, the disease-associated variants might affect other aspects of gene expression such as splicing, that are not captured by current eQTL mapping studies (Li et al., 2016c).

1.6 Outline of the thesis

The second chapter of the thesis focusses on establishing human iPSC-derived macrophages as a model system to study innate immune responses. To this end, I compared the transcriptomes of human monocyte-derived and iPSC-derived macrophages (IPSDMs) before and after stimulation with LPS. I showed that IPSDMs are broadly similar to MDMs and exhibit a conserved response to LPS. I also analysed alternative promoter usage and 3'UTR shortening in LPS response both in MDMs and IPSDMs.

The aim of the third chapter was to establish IPSDMs as a suitable model to study and discover the functions of common genetic variants. I first characterised the reliability and reproducibility of our macrophage differentiation protocol by analysing results from 138 macrophage differentiations from 123 different iPSC lines. Secondly, I characterised the sources of variation that have a strong effect on macrophage gene expression level so that they could be controlled for more effectively in future genomic studies. Finally, because flow cytometry is often used as a quality control step in cellular differentiation assays, I focussed on the factors that are responsible for variability in the expression of cell surface markers in IPSC-derived macrophages.

In the fourth chapter, I used IPSDMs to study the genetics of gene expression in macrophage immune response. We performed RNA-seq on macrophage differentiated from 84 donors in four experimental conditions: naive, IFNɣ stimulation (18 hours), *Salmonella* infection (5 hours) and IFNɣ stimulation followed by *Salmonella* infection. I used this data to answer three main questions: How condition-specific are the genetic effects on gene expression in the four conditions and what proportion of associations remain undetected when studying the naïve cells alone? How does common genetic variation affect other aspects of transcription such as alternative promoter usage, alternative splicing and alternative polyadenylation? What are the complex traits whose genetic risk variants are most enriched among macrophage eQTLs and alternative transcription QTLs?

Finally, in the fifth chapter we used ATAC-seq to measure chromatin accessibility in up to 42 individuals in the same four experimental conditions used in chapter 4. I then identified chromatin accessibility QTLs (caQTLs) and compared them to eQTLs from chapter 4 to explore, how condition-specific are genetic effect on chromatin accessibility compared to gene

expression. I also studied, how genetic effects propagate from chromatin accessibility to gene expression between experimental stimulations. Finally, I tested if caQTLs could be used to fine map causal variants underlying eQTLs and GWAS associations.

2 Comparison of monocyte-derived and iPSC-derived macrophages

Collaboration note

The work described in this chapter has been published as “Transcriptional profiling of macrophages derived from monocytes and iPS cells identifies a conserved response to LPS and novel alternative transcription” (Alasoo et al., 2015). I performed the iPSC-derived macrophage experiments and analysed the data. Fernando O. Martinez from the University of Oxford performed the monocyte-derived macrophage experiments. Subhankar Mukhopadhyay and Gordon Dougan were involved in designing and optimising the experiments and interpreting the results. RNA-seq library construction and sequencing was done by DNA Pipelines core facility at Sanger. I thank Kosuke Yusa and Mariya Chhatriwala for fruitful discussions on troubleshooting iPSC culture.

2.1 Introduction

Macrophages are key cells associated with innate immunity, pathogen containment and modulation of the immune response (Murray and Wynn, 2011; Wynn et al., 2013). Commonly used model systems for studying macrophage biology have included macrophage-like leukemic cell lines, primary macrophages derived from model organisms and primary human macrophages differentiated from blood monocytes. Although these cells have provided important insights into macrophage-associated biology, they have some limitations. Immortalised cell lines often have accumulated multiple genetic aberrations and can exhibit functional defects compared to primary cells such as impaired cytokine production upon inflammatory stimulation (Adati et al., 2009; Schildberger et al., 2013), while multiple functional differences exist between macrophages from different species (Schroder et al., 2012). Additionally, human monocyte derived macrophages (MDMs) can be difficult to obtain in sufficient numbers for repeated experimental assays and it is currently challenging to introduce targeted mutations into their genomes, limiting their utility in genetic studies. For example, introduction of foreign nucleic acid into the cytosol induces a robust antiviral response that may make it difficult to interpret experimental data (Muruve et al., 2008).

Recently, methods have been developed to differentiate macrophage-like cells from human induced pluripotent stem cells (IPSCs) that have the potential to complement current approaches and overcome some of their limitations (Karlsson et al., 2008; van Wilgenburg et al., 2013). This approach is scalable and large numbers of highly pure iPSC-derived macrophages (IPSDMs) can be routinely obtained from any human donor following establishment of an iPSC line. IPSDMs also share striking phenotypic and functional similarities with primary human macrophages (Karlsson et al., 2008; van Wilgenburg et al., 2013). Since human iPSCs are amenable to genetic manipulation, this approach can provide large numbers of genetically modified human macrophages (van Wilgenburg et al., 2013). Previous studies have successfully used IPSDMs to model rare monogenic defects that severely impact macrophage function (Jiang et al., 2012). However, it remains unclear how closely IPSDMs resemble primary human monocyte-derived macrophages (MDMs) at the transcriptome level and to what extent they can be used as an alternative model for functional assays.

Here, we provide an in-depth comparison of the global transcriptional profiles of naïve and lipopolysaccharide (LPS) stimulated IPSDMs with MDMs using RNA-seq. We found that their transcriptional profiles were broadly similar in both naïve and LPS-stimulated conditions. However, certain chemokine genes as well as genes involved in antigen presentation and tissue remodelling were differentially regulated between MDMs and IPSDMs. Additionally, we identified novel changes in alternative transcript usage following LPS stimulation suggesting that alternative transcription may represent an important component of the macrophage immune response.

2.2 Methods

2.2.1 Samples

Human blood for monocyte-derived macrophages was obtained from NHS Blood and Transplant, UK and all experiments were performed according to guidelines of the University of Oxford ethics review committee. All IPSDMs were differentiated from four iPSC lines: CRL1, S7RE, FSPS10C and FSPS11B. CRL1 iPSC line was originally derived from a commercially available human fibroblast cell line and has been described before (Vallier et al., 2009). S7RE iPSC line was derived as part of an earlier study from our lab (Rouhani et al., 2014). FSPS10C

and FSPS11B iPSC lines were derived as part of the Human Induced Pluripotent Stem Cell Initiative (Kilpinen et al., 2016). All iPSC work was carried out in accordance to UK research ethics committee approvals (REC No. 09/H306/73 & REC No. 09/H0304/77).

2.2.2 Cell culture and reagents

IPSCs were grown on Mitomycin C-inactivated mouse embryonic fibroblast (MEF) feeder cells in Advanced DMEM F12 (Gibco) supplemented with 20% KnockOut Serum Replacement (Gibco, cat no 10828-028), 2mM L-glutamine, 50 IU/ml penicillin, 50 IU/ml streptomycin and 50 μM 2-mercaptoethanol (Sigma M6250) on 10 cm tissue-culture treated dishes (Corning). The medium was supplemented with 4 ng/ml rhFGF basic (R&D) and changed daily (10 ml per dish). Prior to passage, the cells were detached from the dish with 1:1 solution of 1 mg/ml collagenase and 1mg/ml dispase (both Gibco). Human macrophage colony stimulating factor (M-CSF) producing cell line CRL-10154 was obtained from ATCC. The cells were grown in T150 tissue culture flasks containing 40 ml of medium (90% alpha minimum essential medium (Sigma), 10% FBS, 2mM L-glutamine, 50 IU/ml penicillin, 50 IU/ml streptomycin). On day 9 the supernatant was sterile-filtered and stored at -80°C.

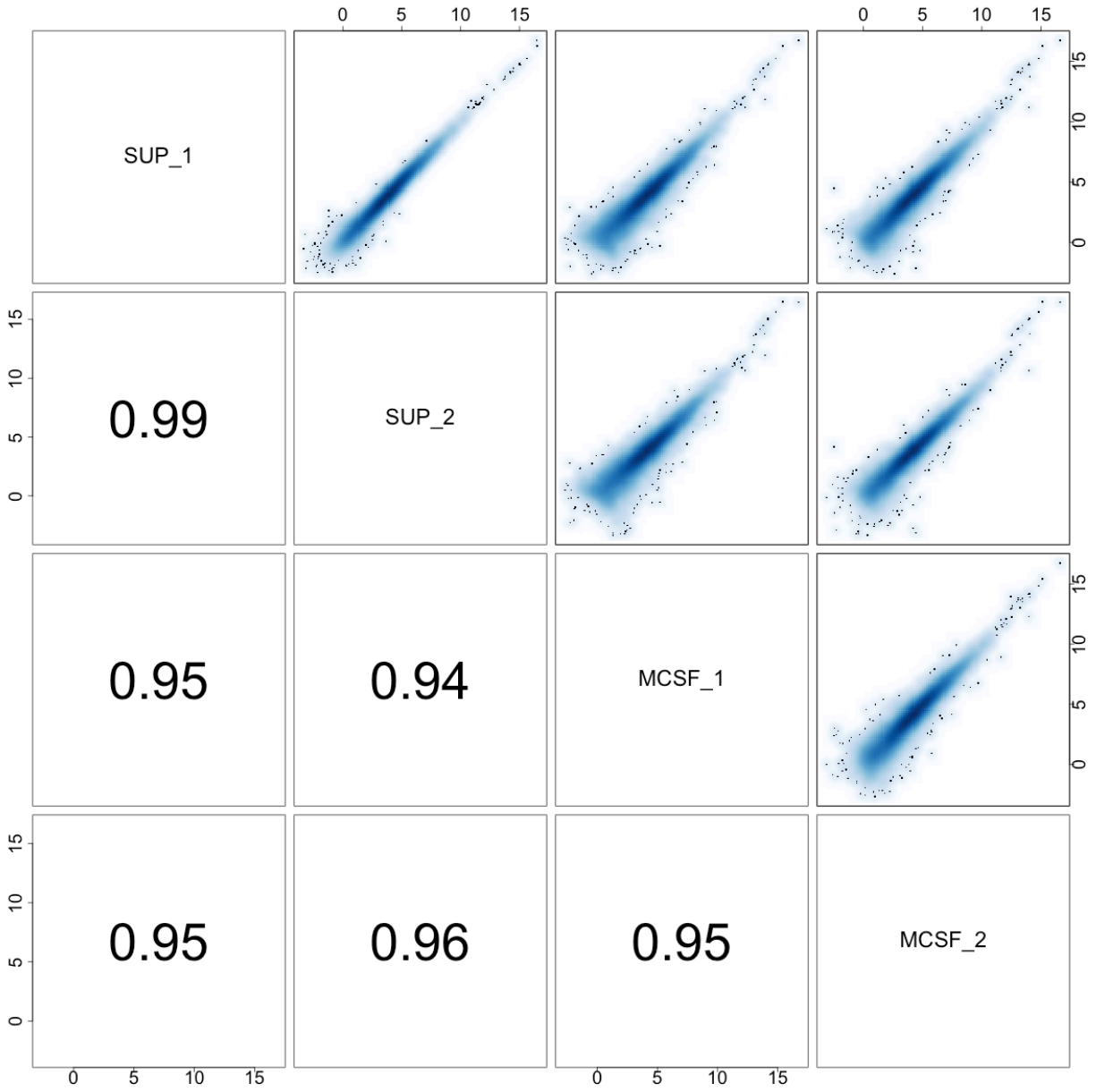

Figure 2.1. Biological reproducibility of IPSDM differentiation. Two biological replicates of FSPS10C-derived IPSDMs differentiated with either supernatant (SUP_1 and SUP_2) or recombinant M-CSF (MCSF_1 and MCSF_2). Above diagonal: pairwise scatterplots of expressed genes (transcripts per million (TPM) > 1) between all four samples. Below diagonal: pairwise Spearman's correlation of gene expression between all four samples.

IPSCs were differentiated into macrophages following a previously published protocol consisting of three steps: i) embryoid body (EB) formation, ii) production of myeloid progenitors from the EBs and iii) terminal differentiation of myeloid progenitors into mature macrophages (van

Wilgenburg et al., 2013). For EB formation, intact iPSC colonies were separated from MEFs using collagenase-dispase solution, transferred to 10 cm low-adherence bacteriological dishes (Sterilin) and cultured in 25 ml iPSC medium without rhFGF for 3 days. Mature EBs were resuspended in myeloid progenitor differentiation medium (90% X-VIVO 15 (Lonza), 10% FBS, 2mM L-glutamine, 50 IU/ml penicillin, 50 IU/ml streptomycin and 50 μM 2-mercaptoethanol (Sigma M6250), 50 ng/ml hM-CSF (R&D), 25 ng/ml hIL-3 (R&D)) and plated on 10 cm gelatinised tissue-culture treated dishes. Medium was changed every 4-7 days. After 3-4 weeks, floating progenitor cells were isolated from the adherent EBs, filtered using a 40 μm cell strainer (Falcon) and resuspended in macrophage differentiation medium (90 % RPMI 1640, 10% FBS, 50 IU/ml penicillin and 50 IU/ml streptomycin) supplemented with 20% supernatant from CRL-10154 cell line. Approximately 7×10^5 cells in 15 ml of media were plated on a 10 cm tissue-culture treated dish and cultured for 7 days until final differentiation. We observed that using supernatant instead of 100 ng/ml M-CSF as specified in the original protocol (van Wilgenburg et al., 2013) did not alter macrophage gene expression profile. The variation between cells differentiated with supernatant or M-CSF was comparable to the variation between two biological replicates of macrophages differentiated with M-CSF (Figure 2.1).

Human monocytes (90-95% purity) were obtained from healthy donor leukocyte cones (corresponding to 450 ml of total blood) by 2-step gradient centrifugation (Martinez, 2012; Martinez et al., 2006). The monocyte fraction in this type of preparation is on average 98% $CD14^+$, 13% $CD16^+$ by single staining. The isolated monocytes were cultured for 7 days in the same macrophage differentiation medium as IPSDMs. The same seeding density and tissue-culture treated plastic was used as for IPSDMs. Non-adherent contaminating cells were removed by vigorous washing before cell lysis at day 7.

On day 7 of macrophage differentiation, medium was replaced with either 10 ml of fresh macrophage medium (without M-CSF) or medium supplemented with 2.5 ng/ml LPS (E. coli). After 6 hours, cells were lifted from the plate using lidocaine solution (6 mg/ml lidocaine, PBS, 0.0002% EDTA), counted with haemocytometer (C-Chip) and lysed in 600 μl RLT buffer (Qiagen). All cells from a dish were used for lysis and subsequent RNA extraction.

2.2.3 Flow cytometry

Flow cytometry was used to characterise the IPSDM cell populations used in the experiments. Approximately 1×10^6 cells were resuspended in flow cytometry buffer (D-PBS, 2% BSA, 0.001%

EDTA) supplemented with Human TruStain FcX (Biolegend) and incubated for 45 minutes on ice to block the Fc receptors. Next, cells were washed once and resuspended in buffer containing one of the antibodies or isotype control. After 1 hour, cells were washed three times with flow cytometry buffer and immediately measured on BD LSRFortessa cell analyser. The following antibodies (BD) were used (cat no): CD14-Pacific Blue (558121), CD32-FITC (552883), CD163-PE (556018), CD4-PE (561844), CD206-APC (550889) and PE isotype control (555749). The data were analysed using FlowJo. The raw data are available on figshare (doi: 10.6084/m9.figshare.1119735).

2.2.4 RNA extraction and sequencing

RNA was extracted with RNeasy Mini Kit (Qiagen) according to the manufacturer's protocol. After extraction, the sample was incubated with Turbo DNase at 37°C for 30 minutes and subsequently re-purified using RNeasy clean-up protocol. Standard Illumina unstranded poly-A enriched libraries were prepared and then sequenced 5-plex on Illumina HiSeq 2500 generating 20-50 million 75bp paired-end reads per sample. RNA-seq data from six iPSC samples was taken from a previous study (Rouhani et al., 2014). Sample information together with the total number of aligned fragments are detailed in Table 2.1.

Table 2.1: General information about the RNA-seq samples. Library size column contains the total number of aligned fragments per sample.

Sample	Donor	Cell type	Treatment	Library size
S7_RE15	S7RE	IPSC	control	83280070
S7_RE11	S7RE	IPSC	control	72411619
S4_SF5	S4SF	IPSC	control	72167859
S4_SF3	S4SF	IPSC	control	72427265
S5_SF1	S5SF	IPSC	control	90998616
S5_SF3	S5SF	IPSC	control	83746320
CRL1_ctrl	CRL1	IPSDM	control	47052432
S7RE_ctrl	S7RE	IPSDM	control	25322078
FSPS10C_ctrl	FSPS10C	IPSDM	control	23443481
FSPS11B_ctrl	FSPS11B	IPSDM	control	19933949
CRL1_LPS	CRL1	IPSDM	LPS	33985920

S7RE_LPS	S7RE	IPSDM	LPS	24349911
FSPS10C_LPS	FSPS10C	IPSDM	LPS	24570506
FSPS11B_LPS	FSPS11B	IPSDM	LPS	24394255
B1_ctrl	B1	MDM	control	23381545
B4_ctrl	B4	MDM	control	47790764
B5_ctrl	B5	MDM	control	26056124
B2_ctrl	B2	MDM	control	20901894
B3_ctrl	B3	MDM	control	26059134
B1_LPS	B1	MDM	LPS	20748290
B4_LPS	B4	MDM	LPS	25538994
B5_LPS	B5	MDM	LPS	56227352
B2_LPS	B2	MDM	LPS	24456569
B3_LPS	B3	MDM	LPS	24075743

2.2.5 RNA-seq data analysis

Differential expression

Sequencing reads were aligned to GRCh37 reference genome with Ensembl 74 annotations using TopHat v2.0.8b (Kim et al., 2013). Reads overlapping gene annotations were counted using featureCounts (Liao et al., 2014) and DESeq2 (Love et al., 2014) was used to identify differentially expressed genes. Genes with FDR < 0.01 and fold-change > 2 were identified as differentially expressed. We used g:Profiler to perform Gene Ontology and pathway enrichment analysis (Reimand et al., 2011). For conditional enrichment analysis of the genes differentially regulated in LPS response we used all LPS-responsive genes as the background set. All analysis was performed on genes classified as expressed in at least one condition (TPM > 2) except where noted otherwise. The bedtools (Quinlan and Hall, 2010) suite was used to construct BigWig files with genome-wide read coverage. All downstream analysis was carried out in R and ggplot2 was used for figures.

Effect of genetic differences on differential expression analysis

To estimate the contribution that genetic differences between IPSDMs and MDMs might have on the differential expression analysis, I obtained gene level RNA-seq read counts from

lymphoblastoid cell lines (LCLs) from 84 British individuals from a previously published study (Lappalainen et al., 2013). To mimic our experimental design, I repeatedly (100 times) sampled 9 individuals from the pool of 84, assigned them randomly into two groups (four and five individuals) and used DESeq2 to estimate the number of differentially expressed genes between the groups that satisfied the same thresholds that I used in the main analysis (FDR < 0.01, fold change > 2).

Alternative transcript usage

To quantify alternative transcript usage, reads were aligned to Ensembl 74 transcriptome using bowtie v1.0.0 (Langmead et al., 2009). Next, I used mmseq and mmdiff to quantify transcript expression and identify transcripts whose proportions had significantly changed (Turro et al., 2011, 2014). For each transcript I estimated the posterior probability of five models (i) no difference in isoform proportion (null model), (ii) difference between LPS treatment and control (LPS effect), (iii) difference between IPSDMs and MDMs (macrophage type effect), (iv) independent treatment and cell type effects (both effects), (v) LPS response different between MDMs and IPSDMs (interaction effect). I specified the prior probabilities as (0.6, 0.1, 0.1, 0.1, 0.1) reflecting the prior belief that most transcripts were not likely to be differentially expressed. Transcripts with posterior probability of the null model < 0.05 were considered significantly changed.

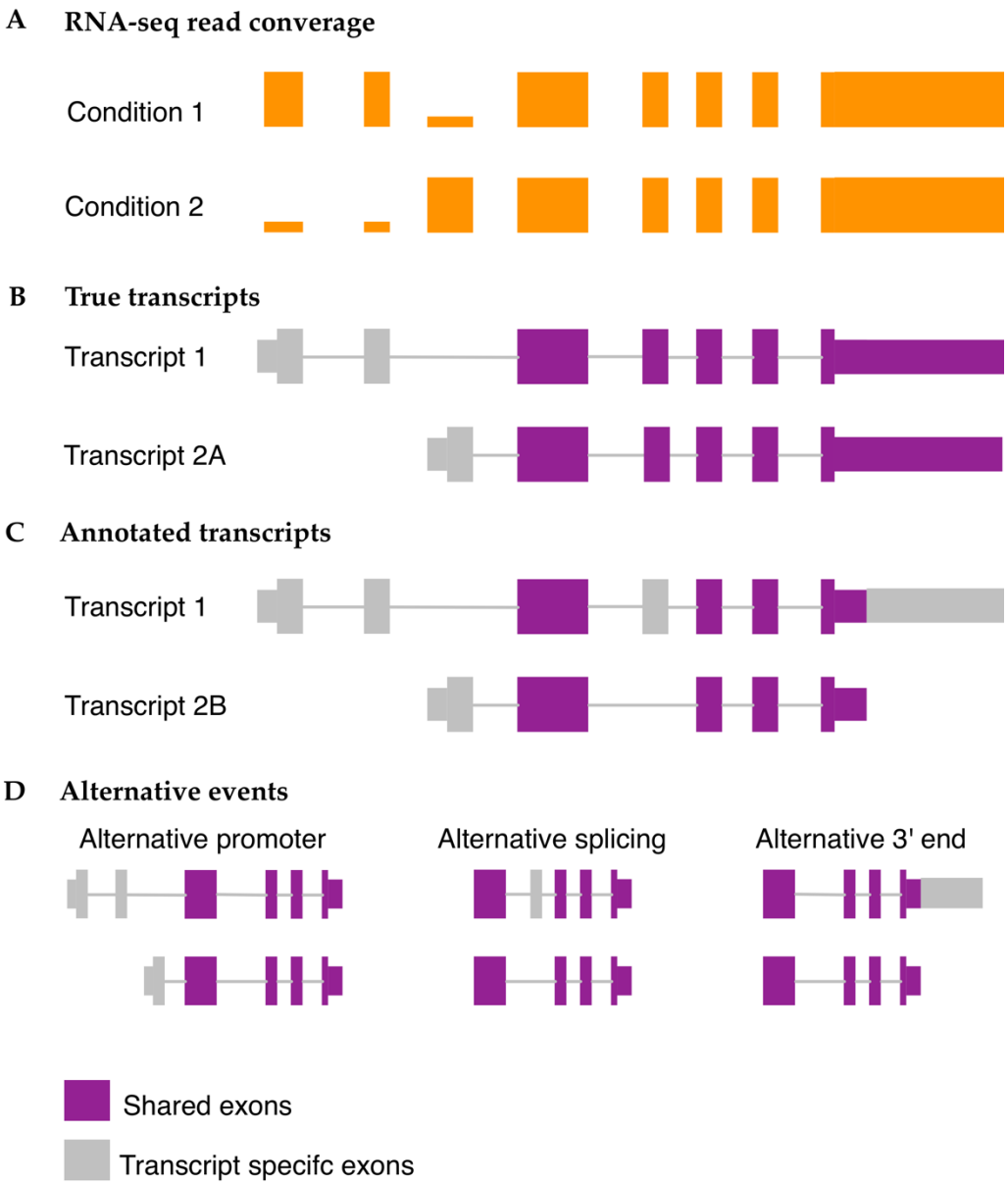

Figure 2.2. Constructing alternative transcription events from annotated transcripts. (A) Hypothetical RNA-seq read coverage over a gene indicating that there is switch from proximal to distal promoter between conditions 1 and 2. **(B)** True transcript annotations generating the read coverage observed on panel A. **(C)** Hypothetical reference transcripts detected to be differentially expressed between conditions 1 and 2. Note that the true transcript 2A from which the reads were generated was not present in the annotated transcripts. Consequently, different transcript 2B was detected to be differentially expressed that also had a skipped exon 4 and shorter 3′ UTR. Comparing transcript 1 to transcript 2B gives the wrong impression that exon 4 and the 3′ UTR are also differentially expressed although their read coverage has not changed between the conditions. **(D)** Three alternative transcription events constructed from transcripts 1

and 2B using the reviseAnnotations package. Estimating the differential expression of these alternative events separately correctly identifies that only the promoter usage changes between conditions.

Next, I used a two-step process to identify the exact alternative transcription events (alternative promoter usage, alternative splicing or alternative 3′ end usage) that were responsible for the observed changes in transcript proportions. First, to identify all potential alternative transcription events in each gene, I compared the transcript whose proportion changed the most between the two conditions to the most highly expressed transcript of the gene (Figure 2.2). This analysis revealed that for 93% of the genes the two selected transcripts differed from each other in more than one location, for example both the promoters and alternative 3′ ends were different between the two transcripts. However, visual inspection of the read coverage plots suggested that in majority of these cases there was only one change between the two transcripts and the other changes were false positives caused by missing or incomplete transcript annotations. To identify which one of the changes was responsible for the alternative transcription signal, I developed the reviseAnnotations R package (https://github.com/kauralasoo/reviseAnnotations) to split the two identified transcripts into individual alternative transcription events (Figure 2.2). Next, I reanalysed the RNA-seq data using exactly the same strategy as described above (bowtie + mmseq + mmdiff) but substituted Ensembl 74 annotations with the identified transcription events. Finally, I required events to change at least 10% in proportion between the two conditions to be considered for downstream analysis. This analysis revealed that instead of the 93% suggested by the transcript level analysis, only 4% of the genes had more than one event whose proportion changed at least 10%, indicating that transcript level analysis leads to a large number of false positives. Our event-based approach is similar to the one used by the Mixture of Isoforms (MISO) model (Katz et al., 2010).

Visualising alternative transcript usage

I developed the wiggleplotr R package (https://github.com/kauralasoo/wiggleplotr) to aid the visualisation of RNA-seq read coverage across alternative transcription events. A key feature of the software is that it allows introns to be shortened to constant width thus making it easier to see differences in read coverage between neighbouring exons in genes with long introns.

2.3 Gene expression variation between iPSCs, IPSDMs and MDMs

2.3.1 Global patterns of gene expression

RNA-seq was used to profile the transcriptomes of MDMs derived from five and IPSDMs derived from four different individuals (Methods). Identical preparation, sequencing and analytical methodologies were used for all samples. Initially, I used Principal Component Analysis (PCA) to generate a genome-wide overview of the similarities and differences between naïve and LPS-stimulated IPSDMs and MDMs as well as undifferentiated iPSCs. The first principal component (PC1) explained 50% of the variance and clearly separated iPSCs from all macrophage samples (Figure 2.3A) illustrating that IPSDMs are transcriptionally much more similar to MDMs compared to undifferentiated iPSCs. This was further confirmed by high expression of macrophage specific markers and low expression of pluripotency factors in IPSDMs (Figure 2.3B). The second PC separated naïve cells from LPS-stimulated cells and explained 16% of the variance, while the third PC, explaining 8% of the variance, separated IPSDMs from MDMs. The principal component that separated IPSDMs from MDMs (PC3) was different from that separating macrophages from iPSCs (PC1). Since principal components are orthogonal to one another, this suggests that the differences between MDMs and IPSDMs are beyond the simple explanation of incomplete gene activation or silencing compared to iPSCs.

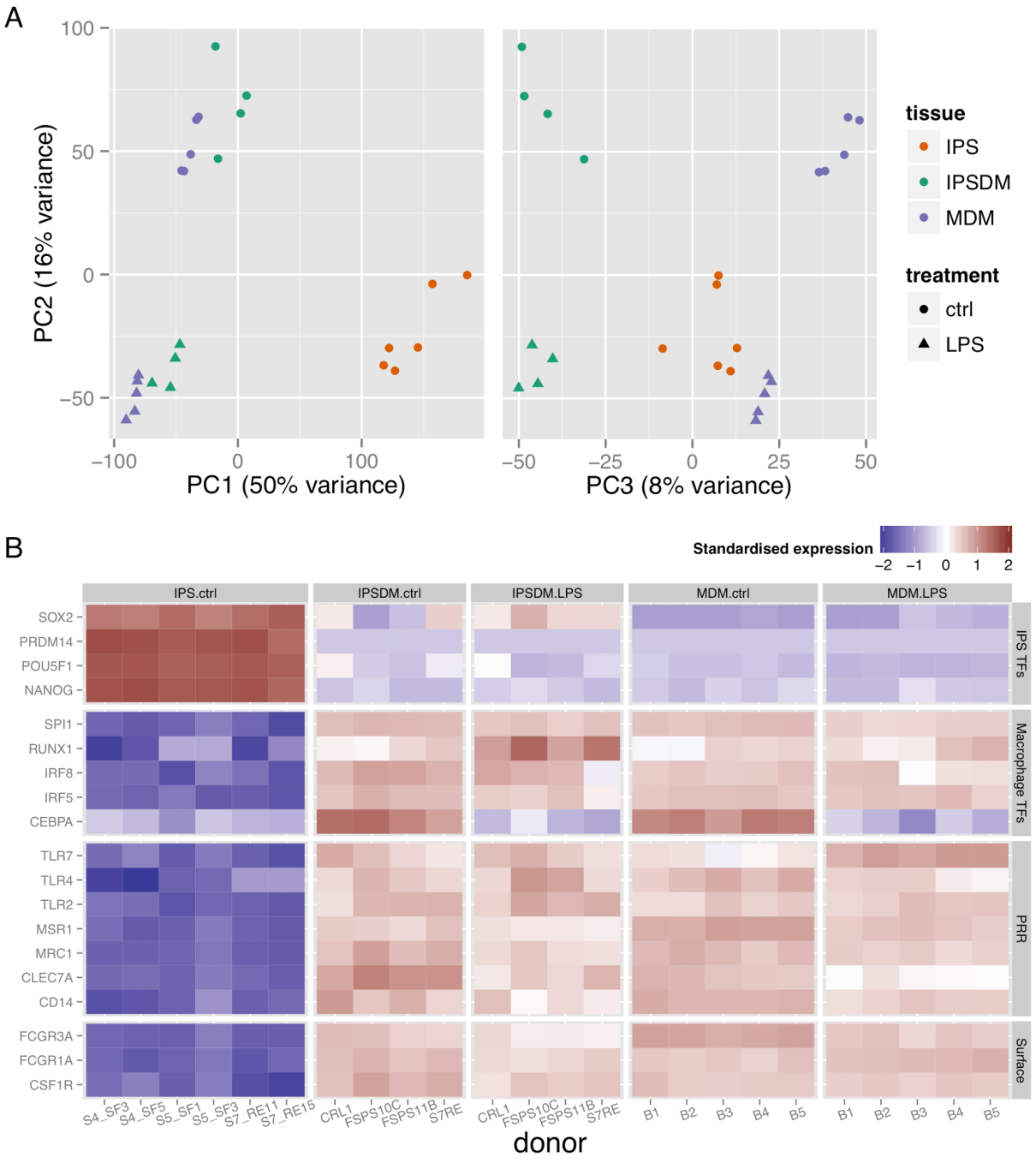

Figure 2.3. Gene expression variation between iPSCs, IPSDMs and MDMs. (A) Principal Component Analysis of expressed genes (TPM > 2) in iPSCs, IPSDMs and MDMs. **(B)** Heatmap showing the gene expression of selected iPSC-specific transcription factors (TFs), macrophage specific TFs, pattern recognition receptors (PRRs) and canonical macrophage cell surface markers. Rectangles correspond to measurements from independent biological replicates.

2.3.2 Differential expression analysis of IPSDMs vs MDMs

Table 2.2. Selection of enriched Gene Ontology terms and KEGG pathways for different groups of differentially expressed genes.

Upregulated in LPS response			
Term ID	**Domain**	**Term name**	**p-value**
GO:0045087	BP	innate immune response	7.31E-45
GO:0009617	BP	response to bacterium	2.42E-28
GO:0032496	BP	response to lipopolysaccharide	4.38E-28
KEGG:04668	ke	TNF signaling pathway	1.71E-20
KEGG:04064	ke	NF-kappa B signaling pathway	3.56E-14
Downregulated in LPS response			
Term ID	**Domain**	**Term name**	**p-value**
GO:0005096	MF	GTPase activator activity	1.01E-09
GO:0007264	BP	small GTPase mediated signal transduction	3.14E-09
More highly expressed in MDMs compared to IPSDMs			
Term ID	**Domain**	**Term name**	**p-value**
GO:0050778	BP	positive regulation of immune response	1.97E-21
GO:0003823	MF	antigen binding	2.55E-18
GO:0005764	CC	lysosome	1.42E-17
GO:0034341	BP	response to interferon-gamma	2.17E-16
GO:0042611	CC	MHC protein complex	3.67E-16
KEGG:04612	ke	Antigen processing and presentation	3.47E-13
KEGG:04145	ke	Phagosome	2.46E-11
More highly expressed in IPSDMs compared to MDMs			
Term ID	**Domain**	**Term name**	**p-value**
GO:0030198	BP	extracellular matrix organization	3.05E-45
GO:0016477	BP	cell migration	1.50E-40
GO:0001568	BP	blood vessel development	4.89E-36
GO:0016337	BP	cell-cell adhesion	6.27E-25
GO:0001525	BP	angiogenesis	1.34E-24

Although PCA provides a clear picture of global patterns and sources of transcriptional variation across all genes in the genome, important signals at individual genes might be missed. To better understand transcriptional changes at the gene level I used a two factor linear model implemented in the DESeq2 package (Love et al., 2014). The model included an LPS effect, capturing differences between unstimulated and stimulated macrophages and a macrophage

type effect capturing differences between MDMs and IPSDMs. Our model also included an interaction term that identified genes whose response to LPS differed between MDMs and IPSDMs. I defined significantly differentially expressed genes as having a fold-change of >2 between two conditions using a p-value threshold set to control our false discovery rate (FDR) to 0.01.

Using these thresholds, I identified 2977 genes that were differentially expressed between unstimulated IPSDMs and MDMs. Among these genes, 2080 were more highly expressed in IPSDMs and 897 were more highly expressed in MDMs (Figure 2.4A). Genes that were more highly expressed in MDMs such as HLA-B, LYZ, MARCO and HLA-DRB1 (Figure 2.4C), were significantly enriched for antigen binding, phagosome and lysosome pathways (Table 2.2). This result is consistent with a previous report that MDMs have higher cell surface expression of MHC-II compared to IPSDMs (Karlsson et al., 2008; van Wilgenburg et al., 2013). Genes that were more highly expressed in IPSDMs, such as MMP2, VEGFC and TGFB2 (Figure 2.4C) were significantly enriched for cell adhesion, extracellular matrix, angiogenesis, and multiple developmental processes (Table 2).

In the LPS response I identified 2638 genes that were differentially expressed in both MDMs and IPSDMs, of which 1525 genes were upregulated while 1113 were downregulated. As might be expected, Gene Ontology and KEGG pathway analysis revealed large enrichment for terms associated with innate immune and LPS response, NF-κB and TNF signalling (Table 2.2). I also identified 569 genes whose response to LPS was significantly different between IPSDMs and MDMs. The majority of these genes (365) responded in the same direction in both IPSDMs and MDMs, but the magnitude of change was significantly different. The remaining 229 genes showed a change in the opposite direction (8.7% of the LPS-responsive genes) (Figure 2.4B). This set of 229 were much weaker responders to LPS overall (2.3-fold compared to 4.7-fold). Additionally, I could not find convincing pathway or Gene Ontology enrichment signals in either gene set (229 and 569 genes) compared to all LPS-responsive genes. Overall, I found that the fold change of the genes that responded to LPS was highly correlated between MDMs and IPSDMs ($r = 0.82$, Figure 2.4B) indicating that the LPS response in these two macrophage types was broadly conserved. Interestingly, I also found that mean fold change was marginally (10%) higher in MDMs (4.95) compared to IPSDMs (4.43). The behaviour of some canonical LPS response genes is illustrated in Figure 2.4D.

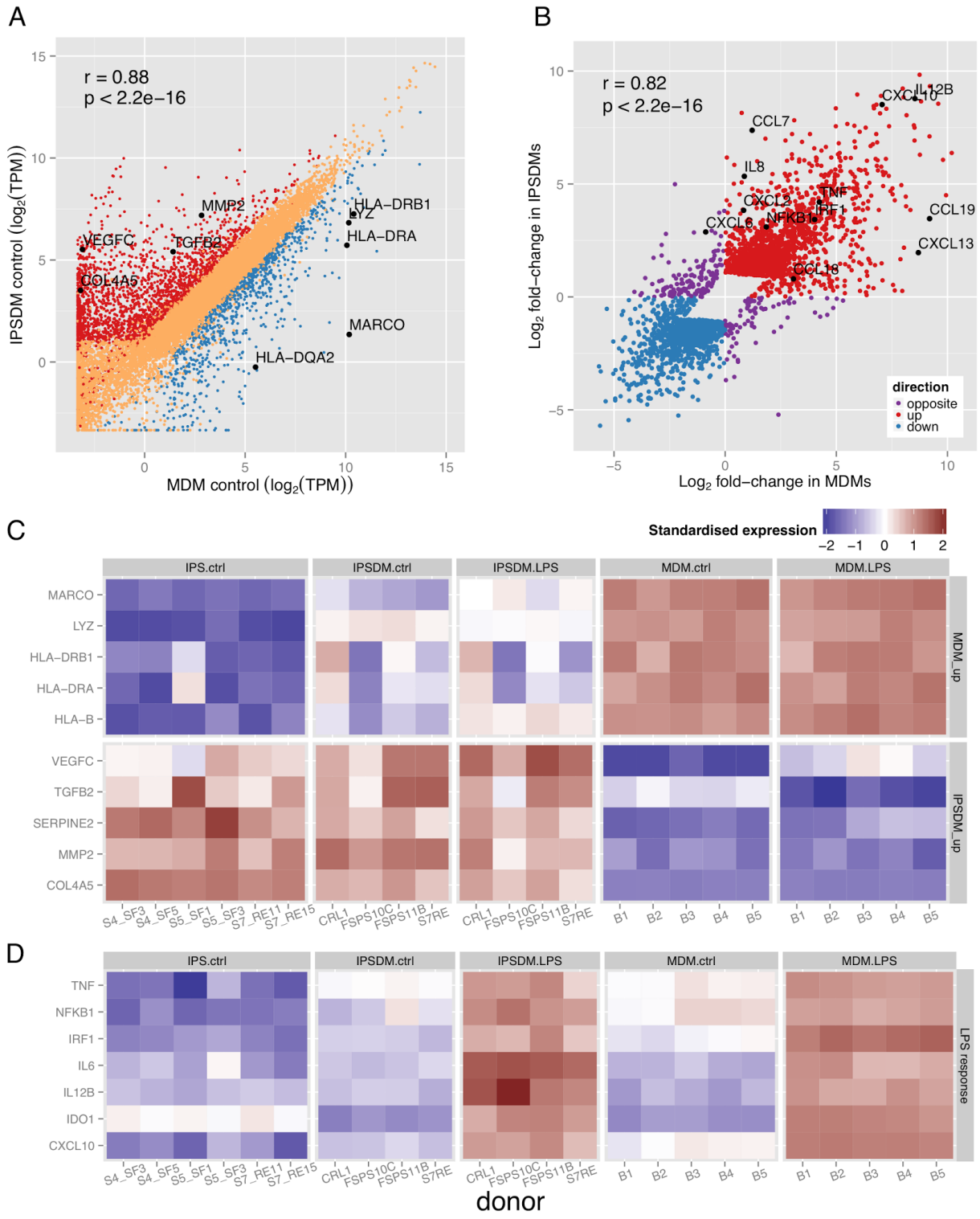

Figure 2.4. Differential expression analysis of IPSDMs vs MDMs. (A) Scatter plot of gene expression levels between MDMs and IPSDMs. Genes that are significantly more highly expressed in IPSDMs are shown in red and genes that are significantly more highly expressed in MDMs are shown in blue. **(B)** Scatter plot of fold change in response to LPS between MDMs

(x-axis) and IPSDMs (y-axis). Only genes with significant LPS or interaction term in the linear model are shown. Genes with LPS response fold change in the opposite direction between MDMs and IPSDMs are highlighted in purple. **(C)** Heatmap of genes differentially expressed between MDMs and IPSDMs. Representative genes from significantly overrepresented Gene Ontology terms (Table 1) include antigen presentation (HLA genes), lysosome formation (LYZ), angiogenesis (VEGFC, TGFB2), and extracellular matrix (SERPINE2, MMP2 COL4A5). The same genes are also marked in panel A. **(D)** Heatmap of example genes upregulated in LPS response.

Although genes with significantly different response to LPS between MDMs and IPSDMs were not enriched for particular Gene Ontology terms or pathways, IL8 and CCL7 mRNAs were more strongly upregulated in IPSDMs compared to MDMs (Figure 2.4B). Consequently, I looked at the response of all canonical chemokines in an unbiased manner. I observed relatively higher induction of further CXC subfamily monocyte and neutrophil attracting chemokines in IPSDMs (Figure 2.3). Moreover, five out of seven CXCR2 ligands (Zlotnik and Yoshie, 2012) were more strongly induced in IPSDMs (FDR < 0.1, fold-change difference between MDMs and IPSDMs > 2) which is significantly more than is expected by chance (Fisher's exact test $p = 4.5\times10^{-6}$) (Figure 2.5). These genes were also expressed at substantial levels (TPM > 100), with IL8 being one of the most highly expressed gene in IPSDMs after LPS stimulation. On the other hand, MDMs displayed relatively higher induction of three chemokines involved in attracting B-cells, T-cells and dendritic cells (CCL18, CCL19, CXCL13) (Figure 2.5).

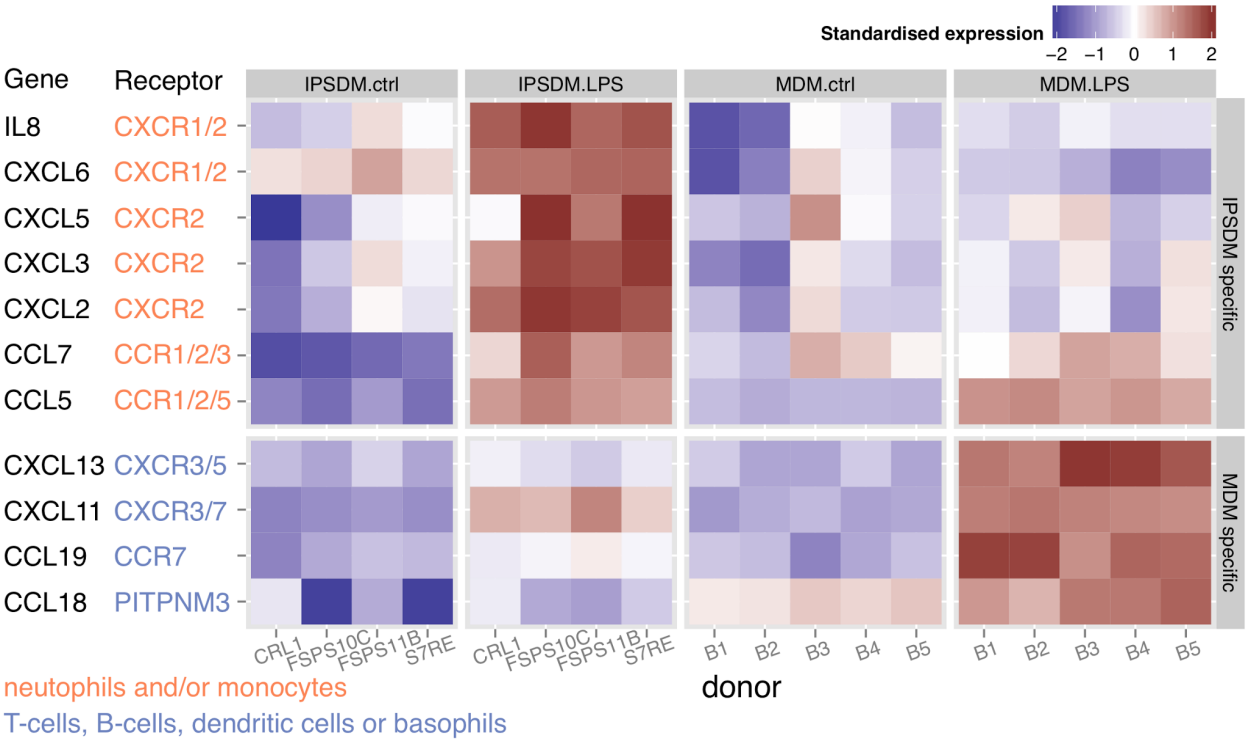

Figure 2.5. Chemokine genes that were particularly upregulated in either IPSDMs or MDMs in LPS response. Their annotated receptors and target cell types were taken from the literature (Soehnlein and Lindbom, 2010; Zlotnik and Yoshie, 2012).

2.3.3 Mechanisms underlying differences between MDMs and IPSDMs

To understand the mechanisms that might underlie the gene expression differences between MDMs and IPSDMs, I focussed on three hypotheses: (1) a minority contaminating cell population in IPSDM samples that is absent in MDMs, (2) genetic differences between donors from which the IPSDMs and MDMs were derived, and (3) incomplete differentiation from iPSCs resulting in developmentally immature macrophages that might exhibit some properties of the iPSCs. The high purity of our IPSDM samples (92-98%) (Table 2.3) and MDM samples (routinely 90-95% pure) suggested that there was no obvious contaminating cell type present that did not express the canonical macrophage markers. Furthermore, even the 99% pure IPSDM samples retained most of the differential expression with MDMs (Figure 2.6A) suggesting contamination is not a major source of IPSDM-MDM differences.

Table 2.3. Purity of iPSC-derived macrophages. We used flow cytometry to estimate the percentage of cells expressing five cell surface markers in IPSDMs differentiated from three IPSC lines.

Marker / Cell line	FSPS10C	FSPS11B	S7RE
CD14	98.6	90.4	91.2
CD206	99.5	85.1	
CD4	99.5	92.8	92.9
CD32	94.8		87.6
CD163	74.1	92	85.6

Alternatively, IPSDMs could be incompletely differentiated from iPSCs. Under this model, genes that are expressed in iPSCs but repressed in mature macrophages would be more highly expressed in IPSDMs compared to MDMs. Consistent with this hypothesis, genes that were more highly expressed in IPSDMs were often also expressed in iPSCs (Figure 2.4C, Figure 2.6A). Furthermore, while the majority of the genes that were more highly expressed in MDMs had mean expression > 2 TPM in both cell types, a large proportion of the genes that were more highly expressed in IPSDMs had mean expression < 1 TPM across both cell types (Figure 2.6B), suggesting that their expression level in IPSDMs might be too low to be functional. Moreover, the promoters of the upregulated genes were highly enriched for repressive H3K27me3 histone marks in CD14+ monocytes (The ENCODE Project Consortium, 2012) (Figure 2.6C), suggesting that these genes normally become silenced prior to monocyte-macrophage differentiation *in vivo* and may not have been completely silenced in IPSDMs.

Finally, it is possible that some of the differences between IPSDMs and MDMs could be confounded with genetic differences between the donors. For example, by chance, the different individuals from which the IPSDMs and MDMs were derived could be fixed for alternate alleles of a cis-regulatory variant that changes the expression of a given gene, which would appear to be differentially expressed between the two cell types. However, since all our IPSDM and MDM donors were randomly sampled from the same population, strong clustering of IPSDM and MDM samples in the PCA analysis (Figure 2.3A) suggests that genetics is not a major source of differences between these cell types. To address this quantitatively, I reanalysed an independent RNA-seq data from 84 British individuals (Lappalainen et al., 2013). I found only a median of three differentially expressed genes between any two random samples of 4 and 5

individuals (Figure 2.6D). This suggests that only a small fraction of the differences between MDMs and IPSDMs are likely to be due to genetics.

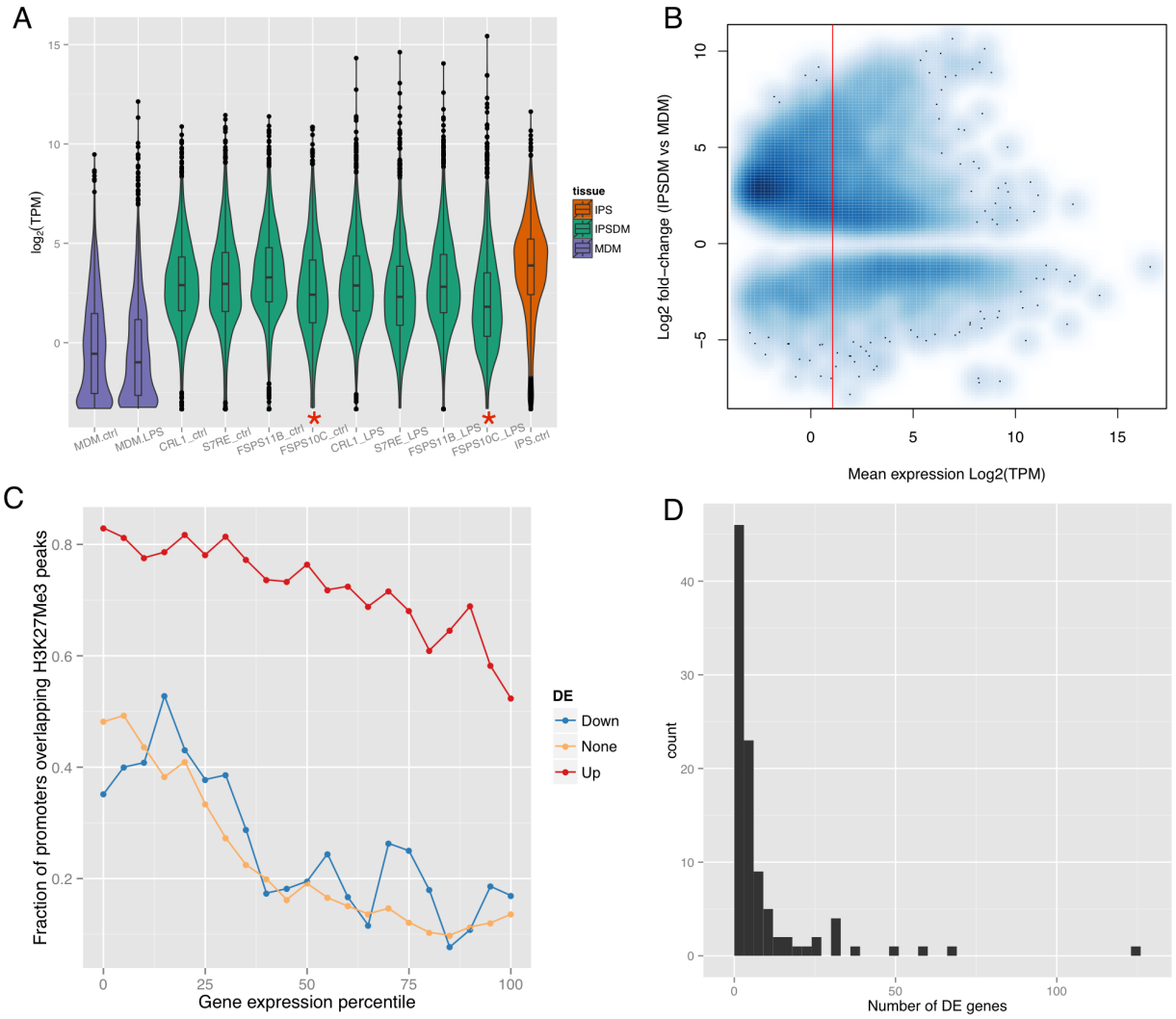

Figure 2.6: Mechanisms underlying differential expression between MDMs and IPSDMs. **(A)** Expression levels of genes that were more highly expressed in IPSDMs compared to MDMs (TPM > 2). Purple violin plots show the mean expression of these genes in MDMs and orange in IPS cells. Red asterisks mark IPSDM samples (FSPS10C) that stained > 99% positive for CD14, CD206 and CD4 while S7RE and FSPS11B samples were ~91% positive. **(B)** MA-plot of differentially expressed genes between MDMs and IPSDMs (without TPM cut-off). On the y-axis is the DESeq2 estimate of fold-change between MDMs and IPSDMs. Red line denotes the 2 TPM cut-off used in most analyses. **(C)** Fraction of gene promoters overlapping H3K27Me3 peaks in ENCODE CD14+ monocyte samples stratified by the percentile of gene expression

level. Up - genes upregulated in IPSDMs; Down - downregulated in IPSDMs; None - not differentially expressed between MDMs and IPSDMs. **(D)** Histogram of the number of differentially expressed genes between two groups of randomly selected individuals.

2.4 Global variation in alternative transcript usage

Many human genes express multiple transcripts that can differ from each other in terms of function, stability or subcellular localisation of the protein product (Carpenter et al., 2014; Wang et al., 2008). Considering expression only at a whole gene level can hide some of these important differences. Therefore, we sought to quantify how similar were naïve and stimulated IPSDMs and MDMs at the individual transcript expression level. Here, we first used mmseq (Turro et al., 2011) to estimate the most likely expression level of each annotated transcript that would best fit the observed pattern of RNA-seq reads across the gene. Next, we calculated the proportion of total expression accounted for by each transcript by dividing transcript expression by the overall expression level of the gene, only including genes that were expressed over two transcripts per million (TPM) (Wagner et al., 2012) in all experimental conditions (8284 genes). Since the proportions of all transcripts of a gene sum to one and most genes express one dominant transcript (Gonzàlez-Porta et al., 2013), I used the proportion of the most highly expressed transcript as a proxy to capture variation in transcript proportions within a gene. In this context and similarly to gene level analysis, the first PC explained 31% of the variance and clearly separated IPSCs from macrophages (Figure 2.7A). However, the second PC (11% of variance) not only separated unstimulated cells from stimulated cells but also IPSDMs from MDMs. One interpretation of this result is that the changes in transcript proportions between IPSDMs and MDMs, to some extent, also resemble those induced in the LPS response. Further analysis (below) highlighted that much of this variation can be explained by changes in 3′ untranslated region (UTR) usage.

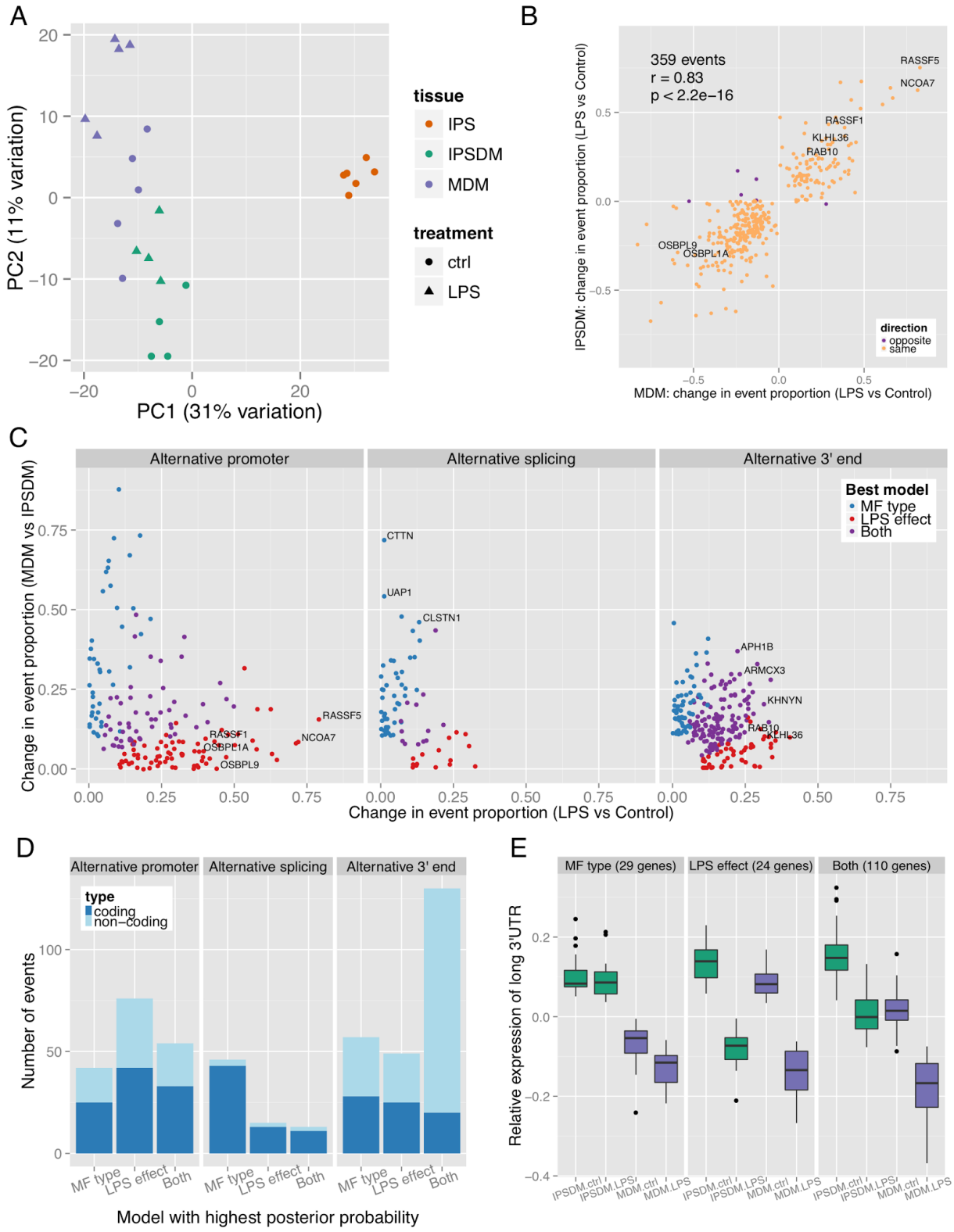

Figure 2.7. Alternative transcription in IPSDMs and MDMs. (A) PCA of relative transcript proportions in iPSCs, IPSDMs and MDMs. Only genes with mean TPM > 2 in all conditions were

included. **(B)** Alternative transcription events detected in LPS response. Each point corresponds to an alternative transcription event and shows the absolute change in the proportion of the most highly expressed transcript (across all samples) in LPS response in MDMs (x-axis) and IPSDMs (y-axis). **(C)** All detected alternative transcription events were divided into three groups based on whether they affected alternative promoter, alternative splicing or alternative 3′ end of the transcript. For each event, we plotted its change in proportion in LPS response (x-axis) against its change between macrophage types (y-axis). The events are coloured by the most parsimonious model of change selected by mmseq: LPS effect (difference between naïve and LPS-stimulated cells only); macrophage (MF) type (difference between IPSDMs and MDMs only); both (data support both MF type and LPS effects). **(D)** Number of alternative transcription events form panel C grouped by position in the gene (alternative promoter, alternative splicing, alternative 3′ end) and most parsimonious model selected by mmseq. (e) Relative expression of long alternative 3′ UTRs in genes showing a change between IPSDM and MDMs (MF type), between naïve and LPS-stimulated cells (LPS effect) and for genes showing both types of change.

2.4.1 Identification and characterisation of alternative transcription events

Alternative transcription can manifest in many forms, including alternative promoter usage, alternative splicing and alternative 3′ end choice, each likely to be regulated by independent biological pathways. Thus, I sought to characterise and quantify how these different classes of alternative transcription events were regulated in the LPS response, and between MDMs and IPSDMs. Using a linear model implemented in the mmdiff (Turro et al., 2014) package followed by a series of downstream filtering steps (Methods) we identified 504 alternative transcription events (ATEs) in 485 genes. Out of those, 145 events changed between unstimulated IPSDMs and MDMs (macrophage (MF) type effect) while 156 events changed between naive and LPS stimulated cells across macrophage types (LPS effect). Further 197 events had different baseline expression between macrophage types, but also changed in the same direction after LPS stimulation (Both effects). Finally, only 6 events change in the opposite direction after LPS stimulation between MDMs and IPSDMs (Figure 2.7B). Next, I focussed on the 359 events that changed in the LPS response in at least one macrophage type (156 + 197 events with LPS response in the same direction and 6 events with LPS response in the opposite direction). I found that the LPS-induced change in the proportion of the most highly expressed transcript was highly correlated between MDMs and IPSDMs (Pearson r = 0.83) (Figure 2.7B), further confirming that the LPS response in both macrophage types is conserved.

Perhaps surprisingly, although the transcriptional response to LPS at the whole gene level is relatively well understood, the effect of LPS on transcript usage has remained largely unexplored. Therefore, I decided to investigate the types of alternative transcription events identified in LPS response as well as between MDMs and IPSDMs (See Methods for details). Most protein coding changes in LPS response were generated by alternative promoter usage (Figure 2.7C-D). In total, I identified 180 alternative promoter events, 51 of which changed the coding sequence by more than 100 bp in LPS response. Strikingly, alternative promoter events displayed larger change in proportion than other events so that often the most highly expressed transcript of the gene changed between cell types and conditions (Figure 2.7C). Alternative promoter usage for three example genes is illustrated on Figure 2.8.

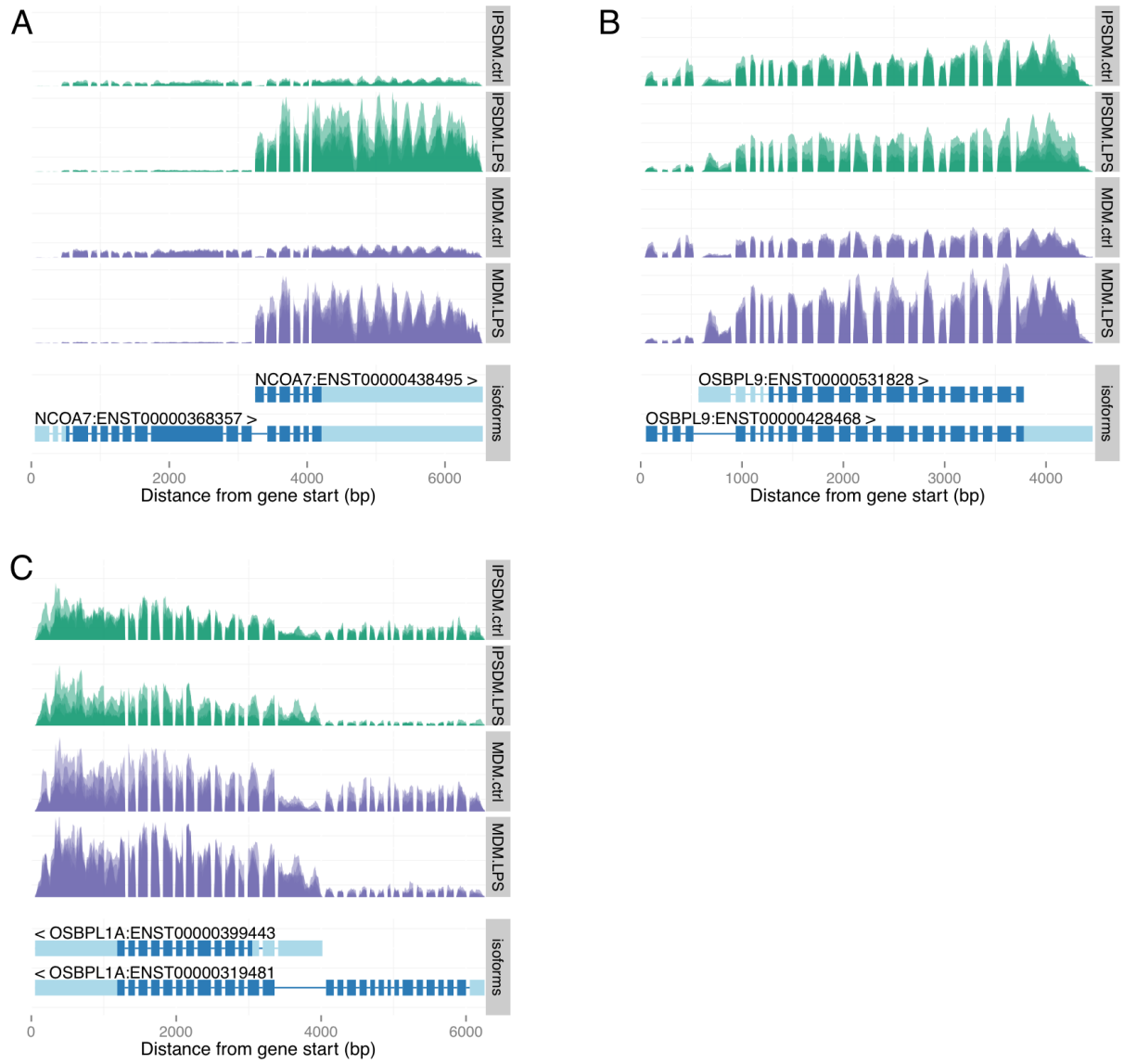

Figure 2.8. Examples of alternative promoter usage in LPS response. Each plot shows normalised read depth across the gene body in IPSDMs (green) and MDMs (purple) with gene structure in the panel beneath each plot. Introns have been compressed relative to exons to facilitate visualisation. (A-C) Alternative promoter usage in NCOA7, OSBPL9 and OSBPL1A genes.

I also observed widespread alternative 3′ end usage both in the LPS response as well as between MDMs and IPSDMs (Figure 2.7C-D). In contrast to alternative promoters, most of the 3′ end events only changed the length of the 3′ UTR and not the coding sequence (Figure 2.7D). Changes in 3′ UTR usage were strongly asymmetric, with longer 3’ UTRs being more highly

expressed in IPSDMs relative to MDMs, and in unstimulated cells relative to stimulated cells (Figure 2.7E, Figure 2.9A). Notably, I also observed that the decrease in 3′ UTR length correlated with the second principal component of relative transcript expression (Figure 2.7A). Consistent with this observation, I found that genes with 3′ UTR events were enriched for high absolute weights in PC2 ($p < 2.2\times10^{-16}$, chi-square goodness-of-fit test), (Figure 2.9B) indicating that part of the transcriptional variation captured by PC2 manifests as changes in 3′ UTR usage. I found no convincing pathway or Gene Ontology enrichment signal in genes with alternative 3′ UTR events.

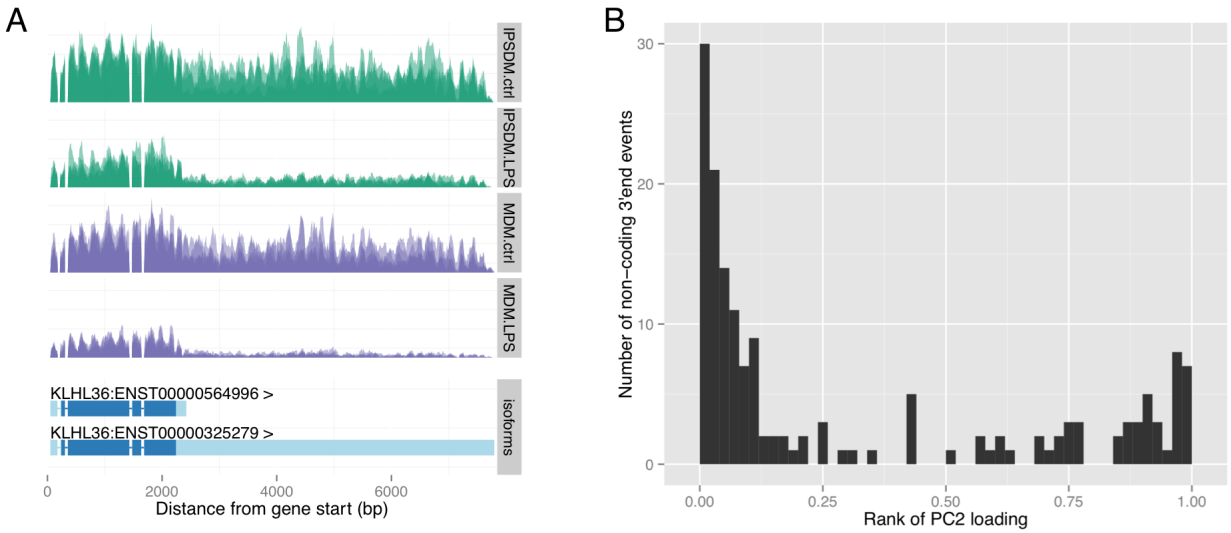

Figure 2.9. 3′ UTR shortening in LPS response. (A) Examples of 3′ UTR shortening in LPS response. The plot shows normalised read depth across the gene body in IPSDMs (green) and MDMs (purple) with gene structure in the panel beneath the plot. Introns have been compressed relative to exons to facilitate visualisation. **(B)** All genes were ranked based on their weights in PC2 (Figure 2.7A) and the relative ranks of the 162 genes with 3'UTR events are displayed on the histogram. The ranks of a randomly sampled set of genes should be uniformly distributed whereas genes that contribute strongly to the PC should be enriched for high and low relative ranks (corresponding to large positive and negative weights on the PC).

Finally, I detected only a small number of alternative splicing events influencing middle exons, most of which occurred between MDMs and IPSDMs rather than in the LPS response (Figure 2.7C-D). Three of the events with largest changes in proportion affected cassette exons in UAP1, CTTN and CLSTN1 genes (Figure 2.10A-C). The inclusion of these exons has previously

been shown to be regulated by RNA-binding protein RBFOX2 that was also significantly more highly expressed in IPSDMs (Figure 2.10D) (Lambert et al., 2014; Venables et al., 2013).

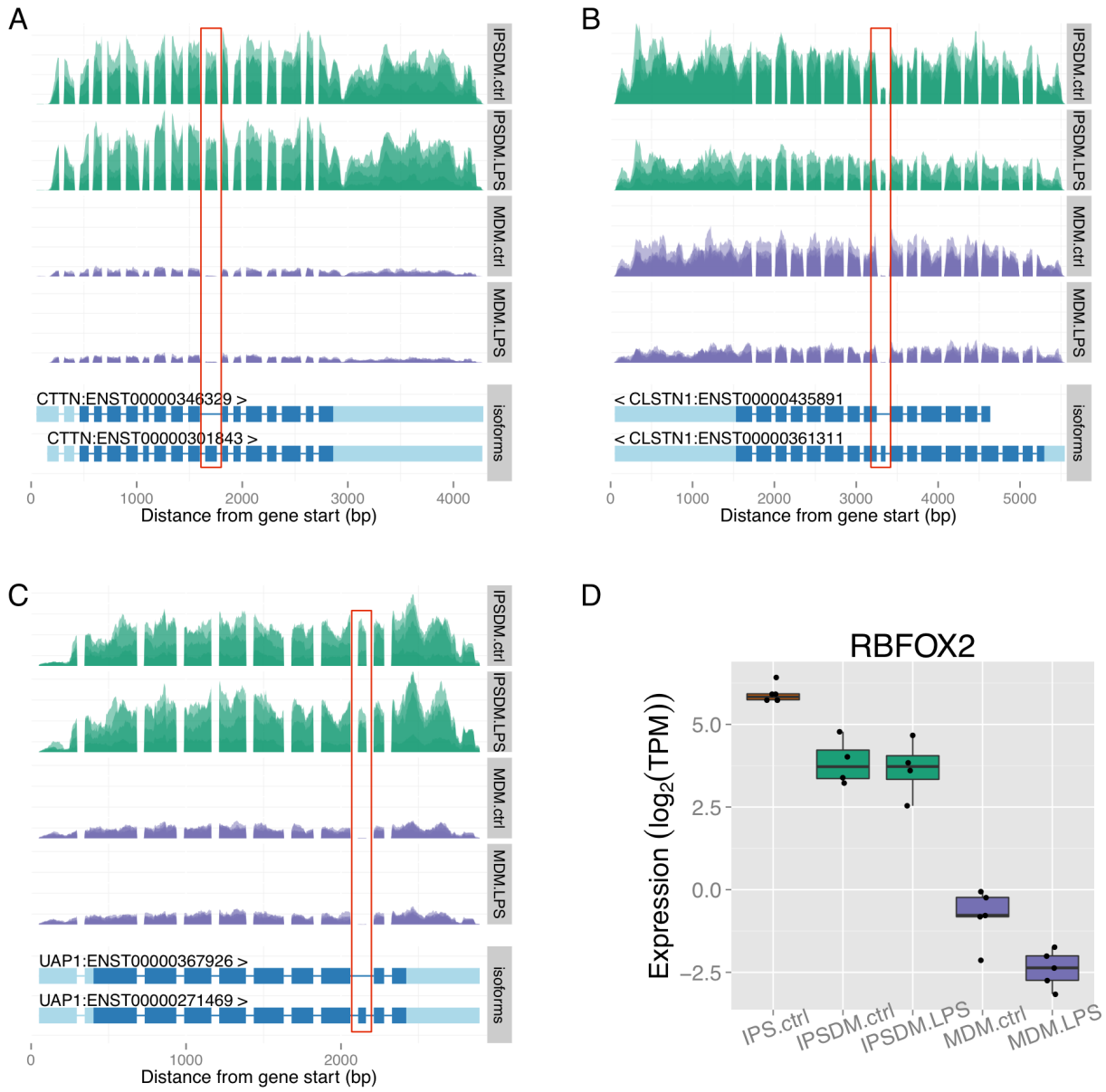

Figure 2.10. Alternative splicing between IPSDMs and MDMs. (A-C) Examples of alternative splicing between MDMs and IPSDMs. The alternatively spliced exon is marked with the red rectangle. **(D)** Expression of RBFOX2 gene in iPSCs, IPSDMs and MDMs.

2.5 Discussion

In this study, we used high-depth RNA-seq to investigate transcriptional similarities and differences between human monocyte and iPSC-derived macrophages. Our principal findings are that, relative to differences between MDMs and iPSCs, the transcriptomes of naïve and LPS stimulated MDMs and IPSDMs are broadly similar both at the whole gene and individual transcript levels. Concurrently with our study, another paper using a different macrophages differentiation protocol came to the same broad conclusion (Zhang et al., 2015). Although we have only examined steady-state mRNA levels, conservation of transcriptional response to LPS implies that the major components of regulatory network that coordinate LPS response on the protein level are likely to also be similarly conserved. We did, however, also observe intriguing differences in expression in specific sets of genes, including those involved in tissue remodelling, antigen presentation and neutrophil recruitment, suggesting that IPSDMs might possess some phenotypic differences from MDMs. Our analysis also revealed a rich diversity of alternative transcription changes suggesting widespread fine-tuning of regulation in macrophage LPS response.

We also looked at the mechanisms that might be underlying the observed differences between MDMs and IPSDMs. We were able to rule out genetic differences between MDMs and IPSDMs or contamination by some other cell type not expressing macrophage specific cell surface markers as a major source of these differences. However, we did find some evidence that IPSDMs might be developmentally less mature than MDMs. This was illustrated by the fact that IPSDMs expressed residual amounts of genes what were substantially more highly expressed in iPSCs and almost completely silenced in MDMs. Furthermore, we found that promoters of these genes were usually actively silenced by H3K27Me3 histone modifications in CD14+ monocytes suggesting that this silencing might be incomplete in IPSDMs.

Alternatively, IPSDMs might share some features with tissue resident macrophages that are developmentally and phenotypically distinct from MDMs (Gautier et al., 2012; Ginhoux et al., 2010; Gosselin et al., 2014; Lavin et al., 2014). In support of that, higher expression of tissue remodelling and neutrophil recruitment genes has previously been associated with tissue and tumour associated macrophages (Cailhier et al., 2005; Mantovani et al., 2013; Schmieder et al., 2012; Soehnlein and Lindbom, 2010). On the other hand, higher expression of antigen presentation genes in MDMs is consistent with the specialised role of monocyte-derived cells in

immune regulation and antigen presentation (Gundra et al., 2014; Jakubzick et al., 2013; Soehnlein and Lindbom, 2010). This is consistent with a previous study suggesting a shared developmental pathway between IPSDMs and foetal macrophages (Klimchenko et al., 2011). Nevertheless, it is likely that the exact characteristics of IPSDMs can be shaped by the addition of cytokines and other factors during differentiation and this could be an important area for further exploration.

In addition to showing that LPS response was broadly conserved between MDMs and IPSDMs both on gene and transcript level, we also identified hundreds of individual alternative transcription events, highlighting an important, but potentially overlooked, regulatory mechanism in innate immune response. A small number of the events have known functional consequences. For example, the LPS-induced short isoform of the NCOA7 (Figure 2.8A) gene is known to be regulated by Interferon β-1b and it is suggested to protect against inflammation-mediated oxidative stress (Yu et al., 2014) whereas the long isoform is a constitutively expressed coactivator of oestrogen receptor (Shao et al., 2002). Similarly, the two isoforms of the OSBPL1A gene (Figure 2.8C) have distinct intracellular localisation and function (Johansson et al., 2003) while the LPS-induced short transcript of the OSBPL9 gene (Figure 2.8B) codes for an inhibitory isoform of the protein (Ngo and Ridgway, 2009). Thus, alternative promoter usage has the potential to significantly alter gene function in LPS response and these changes can be missed in gene level analysis.

Widespread shortening of 3′ UTRs has previously been observed in proliferating cells and cancer as well as activated T-cells and monocytes (Mayr and Bartel, 2009; Sandberg et al., 2008). The functional consequences of 3′ UTR shortening are unclear, but extended 3′ UTRs are often enriched for binding sites for miRNAs or RNA-binding proteins that can regulate mRNA stability and translation efficiency (Gupta et al., 2014; Sandberg et al., 2008). The role of miRNAs in fine-tuning immune response is well established (O'Neill et al., 2011). Furthermore, interactions between alternative 3′ UTRs and miRNAs have recently been implicated in the brain (Miura et al., 2013; Wehrspaun et al., 2014). Therefore, it might be interesting to explore how 3′ UTR shortening affects miRNA-dependent regulation in LPS response.

In summary, we have performed an in depth comparison of an iPSC-derived immune cell with its primary counterpart. Our study suggests that iPSC-derived macrophages are potentially valuable alternative models for the study of innate immune stimuli in a genetically manipulable,

stable cell culture system. The ability to readily derive and store iPSCs potentially enables in-depth future studies of the innate immune response in both healthy and diseased individuals. A key advantage of this model will be the ability to study the impact of human genetic variation, both natural and engineered, in innate immunity.

3 Large-scale differentiation of macrophages from human iPSCs

Collaboration note

The macrophage differentiation work in this chapter was performed in collaboration with Julia Rodrigues who was a research assistant in Daniel Gaffney's lab at the time. I designed the experiments, performed *Salmonella* infection and IFNɣ stimulation assays, took care of sample logistics and performed all of the data analysis. Julia was mainly responsible for tissue culture required for macrophage differentiation and preparing cells for stimulation experiments. Julia also prepared and stained the cells for flow cytometry experiments. Subhankar Mukhopadhyay and Gordon Dougan provided valuable feedback in designing and optimising *Salmonella* infection and IFNɣ stimulation conditions. RNA-seq library construction and sequencing was done by DNA Pipelines core facility at Sanger.

3.1 Introduction

Human induced pluripotent cells (iPSCs) can be derived from almost any individual with many differentiation protocols available for different cell lineages, including macrophages (van Wilgenburg et al., 2013), neurons (Rigamonti et al., 2016) and cardiomyocytes (Kempf et al., 2015). However, typical published differentiation protocols have been developed and used on a few iPSC lines. Hence, the expected range of normal variability between iPSC lines regarding many aspects of these protocols including success rate, duration of differentiation, yield, and purity of the differentiated cells is generally not well understood. If iPSCs are to be used for studying the functions of common genetic variation in differentiated cell types, differentiation protocols need to be robust enough to facilitate large-scale studies in tens or hundreds of lines. However, for most differentiation protocols systematic studies of critical iPSC differentiation parameters are not available.

The factors that influences iPSC differentiation success and yield are not well understood. In one of the largest studies to date, (Koyanagi-Aoi et al., 2013) performed neural differentiation from 10 human embryonic stem cell lines (ESCs) and 40 human iPSC lines. They observed that

7/40 iPSC lines showed aberrant gene expression profiles that correlated with defects in neural differentiation. A smaller study looking at five human ESC lines and 12 iPSC lines observed that iPSCs showed higher variability in their potency to differentiate into neurons compared to ESCs, but was unable uncover a specific cause (Hu et al., 2010). A study of 28 iPSC lines found that variations in hepatic differentiation could largely be attributed to differences between donors (Kajiwara et al., 2012) and other work has found that the method used to form embryoid bodies can have a large effect on differentiation propensity (Paull et al., 2015). Finally, a study in mouse iPSCs showed that the cell type of origin might influence differentiation propensity in early passage iPSCs, but these effects disappeared after 10-16 passages (Polo et al., 2010). Thus, there are many factors influencing differentiation success and their relative importance is likely to vary between protocols.

Additionally, when we differentiate iPSCs into a cell type of interest, we typically have a specific phenotype of interest, such as difference in gene expression level between two conditions, that we want to measure. Ideal experimental design should control for all other sources of variability in differentiation to maximise the chance of detecting the signal of interest. However, controlling all potential sources of variability is often impractical or even unfeasible. Hence, there is great interest in knowing which sources of variability have a strong effect on the phenotype (and should be controlled for) and which are so weak that they can be ignored. Variance component analysis is an effective approach to understand the relative contribution of both technical and biological factors on a phenotype of interest such as gene expression levels (’t Hoen et al., 2013; Rouhani et al., 2014). For example, two recent studies have used this approach to highlight the importance of genetic differences between donors as a major factor underlying gene expression variation in human iPSCs (Kilpinen et al., 2016; Rouhani et al., 2014).

We performed 138 macrophage differentiation attempts from 123 IPSC lines selected randomly from the HipSci project (Kilpinen et al., 2016), making it one of the largest directed differentiation study from human iPSCs. However, some of the differentiated lines did not produce enough macrophages to perform all of the experimental assays or the cells were not pure enough to be used in stimulation experiments. In total, we sequenced the RNA from 84 of these lines in four experimental conditions. We focussed on three questions: (i) how reliable and reproducible was the macrophage differentiation protocol (ii) which sources of variation had a strong effect on macrophage gene expression levels (iii) because flow cytometry is often used as a quality

control step in cellular differentiation assays, what factors are responsible for variability in the expression of cell surface markers in iPSC-derived macrophages.

We were able to successfully differentiate macrophages from 101/123 iPSC cell lines, with an overall success rate of 82%. Combining gene expression data with extensive sample metadata, we were able to estimate the relative proportion of gene expression variance explained by different experimental factors. Our results highlight the importance of maintaining high purity and constant cell density of the differentiated cells. We also showed that using live bacteria can lead to larger stimulation-specific batch effects than using well-defined molecular stimuli such as IFNɣ. Finally, we have shown that expression of CD14 and CD16 cell surface markers can be highly variable between genetically distinct cell lines and in the case of CD14, most of this variation can be attributed to a genetic variant upstream of the CD14 gene. This highlights the importance of accounting for genetic differences when comparing primary and iPSC-derived cells from different individuals.

3.2 Methods

3.2.1 Cell culture and reagents

Donors and cell lines

Human induced pluripotent stem cells (iPSCs) from 123 healthy donors (72 females and 51 males) were obtained from the HipSci project (Kilpinen et al., 2016). Of these lines, 57 were initially grown in feeder-dependent medium and 66 were grown in feeder-free E8 medium.

Feeder-free iPSC culture

Feeder-free iPSCs were grown on tissue culture treated plates coated with vitronectin (VTN-N) (Gibco, cat. no. A14700) in Essential 8 (E8) medium (Gibco). The cells were dissociated from the plates using Gentle Cell Dissociation Buffer (Stemcell Technologies, cat. no. 07174) and passaged every 3-5 days. Prior to macrophage differentiation, the feeder-free iPSCs were first transferred to feeder-dependent media and propagated for at least two passages. This step was necessary because multiple attempts to differentiate macrophage directly from feeder-free iPSCs with our protocol failed.

Feeder-dependent iPSC culture

Feeder-dependent iPSCs were grown on irradiated CF-1 mouse embryonic fibroblast (MEF) feeder cells (AMS Biotechnology) in Advanced DMEM-F12 (Gibco) supplemented with 20% KnockOut Serum Replacement (KSR) (Gibco), 2mM L-glutamine (Sigma), 50 IU/ml penicillin (Sigma), 50 IU/ml Streptomycin (Sigma) and 50μM β-Mercaptoethanol (Sigma M6250). The media was supplemented with 4 ng/ml recombinant human fibroblast growth factor (rhFGF) basic (R&D, 233-FB-025) to maintain pluripotency and was changed daily. MEFs were seeded on 0.1% gelatine-coated tissue-culture treated plates (Corning 6-well or 10 cm plates) 24 hours prior to passaging iPSCs at a cell density of 2 million cells per 6-well or 10-cm plate in Advanced DMEM-F12 supplemented with 10% FBS (Labtech), 2mM L-glutamine (Sigma), 50IU/ml Penicillin & 50IU/ml Streptomycin (Sigma). Prior to passaging or embryoid body formation, iPSCs were dissociated from the plates using 1:1 mixture of collagenase (1 mg/ml) and dispase (1 mg/ml) (both Gibco).

Macrophage differentiation

iPSCs were differentiated into macrophages using a previously published protocol (van Wilgenburg et al., 2013) involving 3 stages: i) embryoid body (EB) formation, ii) generation of monocyte-like myeloid progenitors from the EBs and iii) terminal differentiation of the progenitors into macrophages. For EB formation, iPSC colonies were treated with 1:1 mixture of collagenase (1 mg/ml) and dispase (1 mg/ml) and intact colonies were transferred to low-adherence plates (Sterilin). The colonies were cultured in feeder-dependent iPSC medium without rhFGF for 3 days. On day 3, the EBs were harvested and transferred to gelatinised tissue-culture treated 10 cm plates in serum-free haematopoietic medium (Lonza X-VIVO 15), supplemented with 2 mM L-glutamine (Sigma), 50 IU/ml penicillin, 50 IU/ml streptomycin (Sigma), 50 μM β-Mercaptoethanol (Sigma M6250), 50 ng/ml macrophage colony stimulating factor (M-CSF) (R&D) and 25 ng/ml interleukin-3 (IL-3) (R&D). EBs were maintained in these plates with media changes every 3-5 days for 4-6 weeks until the progenitor cells appeared in the supernatant. Progenitor cells were harvested from the supernatant, filtered through a 40μm cell strainer (BD 352340), centrifuged at 1200 rpm for 5 minutes, counted, and plated in RPMI 1640 (Gibco) supplemented with 10% FBS (Labtech), 2mM L-glutamine (Sigma) and 100 ng/ml hM-CSF (R&D) at a cell density of 150,000 cells per 6-well plate or 1,000,000 cells per 10 cm plate and differentiated for another 7 days.

3.2.2 Macrophage stimulation assays

After harvesting, macrophage progenitors were seeded on 6-well plates at 150,000 cells/well. Two wells were used per condition to ensure sufficient amount of RNA. On day 6 of macrophage differentiation, medium was changed for all wells with half of the wells receiving macrophage differentiation media (with M-CSF) and half of the cells receiving macrophage differentiation media supplemented with 20 ng/ml IFNɣ (R&D) and M-CSF. After 18 hours, cells from two wells of the naive and IFNɣ conditions were harvested for RNA extraction. The remaining two wells from each condition were additionally infected with *Salmonella* Typhimurium SL1344 (hereafter 'SL1344') for 5 hours. For RNA extraction, cells were washed once with PBS and lysed in 300 µl of RLT buffer (Qiagen) per one well of a 6-well plate. Lysates from two wells were immediately pooled and stored at -80°C. RNA was extracted using RNA Mini Kit (Qiagen) following manufacturer's instructions and eluted in 35 µl nuclease-free water. RNA concentration was measured using NanoDrop and RNA integrity was measured on Agilent 2100 Bioanalyzer using RNA 6000 Nano total RNA kit.

Two days before infection, *Salmonella* Typhimurium SL1344 culture was inoculated in 10 ml low salt LB broth and incubated overnight in a shaking incubator (200 rpm) at 37°C. Next morning, the culture was diluted 1:100 into 10 ml of fresh LB broth and incubated again in a shaking incubator. In the afternoon the culture was diluted once more 1:100 into 45 ml of LB broth and kept overnight in a static incubator. In the morning before infection, the culture was centrifuged at 4000 rpm for 10 minutes, washed once with 4°C PBS and re-suspended in 30 ml of PBS. Subsequently, optical density at 600 nm was measured and *Salmonella* was diluted in macrophage differentiation media (without M-CSF) at multiplicity of infection (MOI) 10 assuming 300,000 cells per well. To infect the cells, old media was removed and replaced with 1 ml of media containing *Salmonella* for 45 minutes. Subsequently, the cells were washed twice with PBS and replaced in fresh medium with 50 ng/ml gentamicin (Sigma) to kill extracellular bacteria. After 45 minutes, the medium was changed once again to fresh medium containing 10 ng/ml gentamicin.

3.2.3 RNA sequencing

All of the RNA-seq libraries were constructed using poly-A selection. The first 120 RNA-seq libraries from 30 donors were constructed manually using the Illumina TruSeq stranded library preparation kit. The TruSeq libraries were quantified using Bioanalyzer and manually pooled for

sequencing. For the remaining 216 samples, we used an automated library construction protocol that was based on the KAPA stranded mRNA-seq kit. The KAPA libraries were quantified using Quant-iT plate reader and pooled automatically using the Beckman Coulter NX-8. The first 16 samples were sequenced on Illumina HiSeq 2500 using V3 chemistry and multiplexed at 4 samples/lane. All of the other samples were sequenced on Illumina HiSeq 2000 using V4 chemistry and multiplexed at 6 samples/lane.

RNA-seq pre-processing and quality control

I aligned RNA-seq data to the GRCh38 reference genome and Ensembl 79 transcript annotations using STAR v2.4.0j (Dobin et al., 2013). I then used VerifyBamID v1.1.2 (Jun et al., 2012) to detect and correct any potential sample swaps and cross-contamination between donors. I did not detect any cross-contamination, but I did identify one sample swap between two donors. I used featureCounts v1.5.0 (Liao et al., 2014) to count the number of uniquely mapping fragments overlapping GENCODE (Harrow et al., 2012) basic annotation from Ensembl 79. I excluded short RNAs and pseudogenes from the analysis leaving 35,033 unique genes of which 19,796 were protein coding. I only used 15,797 genes with mean expression in at least one of the conditions greater than 0.5 transcripts per million (TPM) (Wagner et al., 2012) in all downstream analyses. I also quantile-normalised the data and corrected for sample-specific GC content bias using the conditional quantile normalisation (cqn) (Hansen et al., 2012) R package as recommended previously (Ellis et al., 2013). To detect hidden confounders in gene expression, I applied PEER (Stegle et al., 2012) on each condition separately allowing for at most 10 hidden factors. I found that the first 3-5 factors explained the most variation in the data and the others remained close to zero.

Variance component analysis

I used a linear mixed model implemented in the lme4 (Bates et al., 2015) package to estimate the proportion of variance explained by various biological and technical factors in the expression levels of 15,797 genes across 336 samples. The 14 factors that I included in the model are listed below. Continuous variables were binned into a small number of categories as described.

1. **Salmonella** - Salmonella infection status (yes or no) (binary)
2. **IFNɣ** - IFNɣ stimulation status (yes or no) (binary)
3. **IFNɣ:Salmonella** - interaction term between Salmonella and IFNɣ stimulations (binary)
4. **Line** - the iPSC cell line from which the macrophages were derived. All lines used in the analysis were from 84 unique donors. This component should capture genetic

differences between donors, but can also capture line and differentiation specific effects. (84 categories)

5. **Cell density** - I used mean RNA concentration across the four conditions as proxy of the total number of cells on a plate, because counting the cells prior to lysis and RNA extraction was not feasible. (categorical: 0-100 ng/ul, 100-200 ng/ul, 200-300 ng/ul, 300-500 ng/ul)
6. **Library type** - type of the RNA library construction method used (manual or automatic) (binary)
7. **Sex** - sex of the donor (binary)
8. **Purity** - purity of the differentiated macrophages as quantified by flow cytometry. This is a noisy measurement, because RNA-seq and flow cytometry were not performed from the same plate of cells and they were often performed on different days (up to 2 weeks apart) due to logistical reasons (categorical: 90-95%, 95-97.5%, 97.5-100%).
9. **Chemistry** - chemistry of the Illumina RNA-seq protocol (V3 or V4).
10. **Stimulation date** - date of the stimulation assays and cell lysis (categorical: 32 levels)
11. **Library pool** - RNA-seq library construction batch (categorical: 10 levels)
12. **RNA extraction** - RNA extraction batch (categorical: 31 levels)
13. **Differentiation duration** - Number of days from the start of the differentiation until cell lysis (5 categories: 20-30 days, 31-40 days, 41-50 days, 51-60 days, 61+ days).
14. **Passage** - passage of the IPSC line at the start of the differentiation (4 categories: 0-25, 26-35, 36-45, 46-60)

First, I analysed all 15,797 expressed genes from all of the 336 samples across the four conditions together using a single linear mixed model with all of the 14 factors included as random effects. The following model was fit to each gene independently, using `lme4`:

```
expression ~ (1|Salmonella) + (1|IFNɣ) + (1|IFNɣ:Salmonella) + (1| Line) +
      (1| Cell_density) + (1| Library_type) + (1| Stimulation_date) +
      (1| Sex) + (1| Chemistry) + (1| Purity) + (1| Passage) +
      (1| Diff_duration) + (1| Library_pool) + (1| RNA_extraction)
```

To better understand the relative contribution of weaker technical factors and how their effects might vary between conditions, I also performed variance component analysis in each condition separately by only including the ten technical factors in the model as random effects:

```
expression ~ (1| Cell_density) + (1| Library_type) + (1| Stimulation_date) +
      (1| Sex) + (1| Chemistry) + (1| Purity) + (1| Passage) +
      (1| Diff_duration) + (1| Library_pool) + (1| RNA_extraction)
```

Next, I used the `VarCorr` function from the `lme4` package to calculate the amount of variance attributed to each of the factors. I then estimated the proportion of variance explained by each factor by dividing the variance attributed to each factor by the total variance of the gene. As a result, for each factor I obtained a distribution of the proportion of variance explained estimates across 15,797 genes.

3.2.4 Flow cytometry

Measuring macrophage cell surface marker expression using flow cytometry

We used flow cytometry to measure the cell surface expression of three canonical macrophage markers: CD14, CD16 (FCGR3A/FCGR3B) and CD206 (MRC1). Macrophages were cultured in 10 cm tissue-culture treated plates and detached from the plates by incubation in 6 mg/ml lidocaine-PBS solution (Sigma L5647) for 30 minutes followed by gentle scraping. From each cell line we harvested between 300,000-500,000 cells. Detached cells were washed in media, centrifuged at 1200 rpm for 5 minutes and resuspended in flow cytometry buffer (2% BSA, 0.001% EDTA in D-PBS) and split into two wells of a 96-well plate. Nonspecific antibody binding sites were blocked by incubating cells with Human TruStain FcX (Biolegend) for 45 minutes and washing with flow cytometry buffer. Half of the cells were stained for 1 hour with the PE-isotype control (BD 555749) antibody. The other half of the cells were co-stained for 1 hour with following three antibodies: CD14-Pacific Blue (BD 558121), CD16-PE (BD 555407), CD206-APC (BD 550889). After staining, the cells were washed three times. Resuspended cells were filtered through cell strainer cap tubes (BD 352235) and measured on the BD LSRFortessa Cell Analyzer.

Flow cytometry data analysis

I used the flow cytometry data for two purposes: to estimate the proportion of cells expressing macrophage surface markers CD14, CD16 and CD206 and to quantify the relative intensity of these markers compared to unstained cells. I imported the raw FCS files into R using the OpenCyto (Finak et al., 2014) package. First, I logicle-transformed (Herzenberg et al., 2006) the

intensity values for all three channels in both stained and isotype control samples using the `estimateLogicle` function. I then performed two automated gating steps to exclude debris and identify the main cell population using the `mindensity` (max = 150,000) and `flowCust` (K=2, target=c(1e5,5e4), level=0.9) functions. For pure macrophage samples the distribution of intensity values for all three cell surface markers looked bimodal with stained and unstained cells in two separate peaks (Figure 3.1). Samples with moderate contamination had an additional low intensity peak both in stained and unstained cells (Figure 3.1) corresponding to the contaminating cells. Since all of the peaks were approximately normally distributed, I decided to model the data for each mark as a mixture of Gaussian distributions and used the mclust (Fraley and Raftery, 1999) R package to estimate the optimal number of components (2 or 3) as well as the mean and standard deviation of each component. I used the Bayesian Information Criterion to choose between two or three components. I then compared the mean of the highest intensity peak ($\mu_{stained}$) to the mean of the second highest intensity peak ($\mu_{unstained}$) to estimate the relative fluorescent intensity of each cell surface marker (Figure 3.1). I also measured sample purity by estimating the proportion of cells whose intensity was greater than the threshold $t = \mu_{stained} - 3\times\sigma_{stained}$ (Figure 3.1), where $\sigma_{stained}$ is the standard deviation of the stained population.

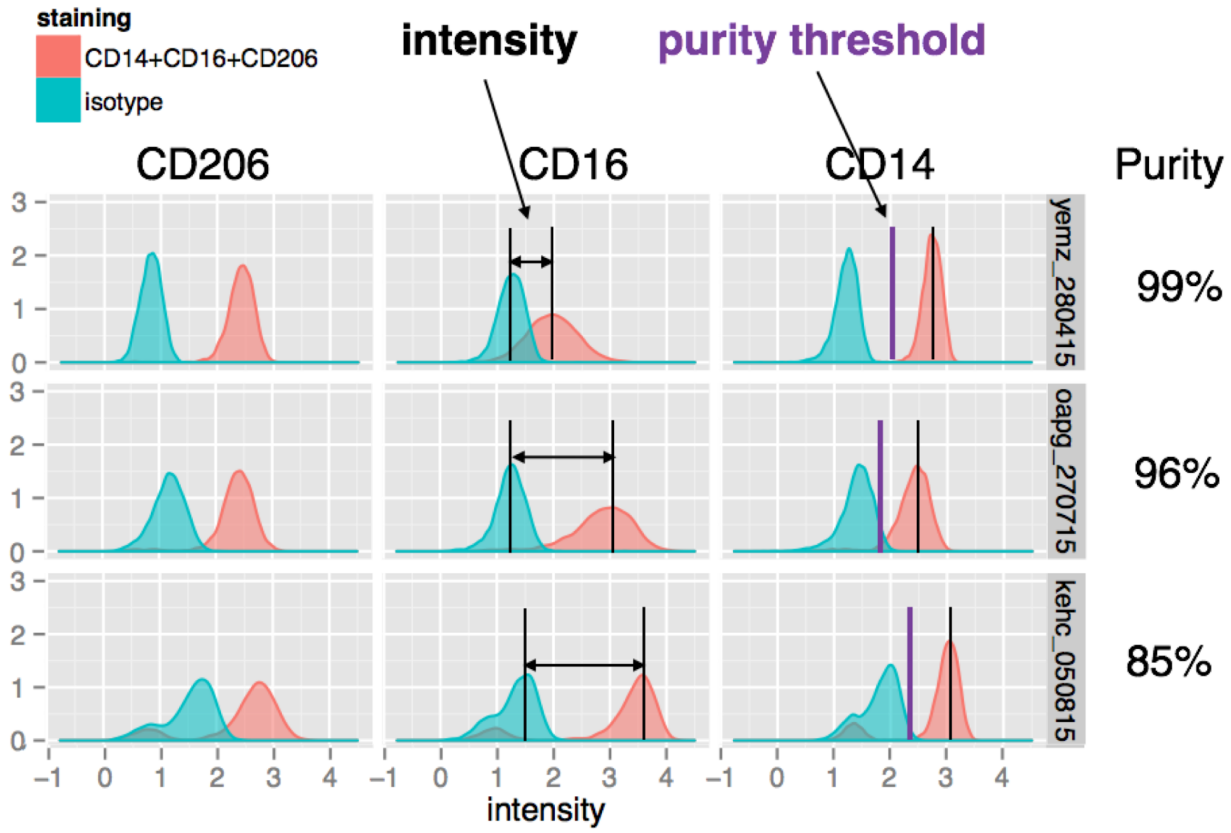

Figure 3.1: Quantifying cell purity and relative fluorescent intensity of macrophage CD206, CD16 and CD14 markers from the flow cytometry data. The rows correspond to three different iPSC lines and the columns represent three macrophage markers. X-axis shows the logicle-transformed absolute intensity values from the flow cytometer and values on the y-axis correspond the density of the cells with that intensity value. Red designates cells stained with antibodies against the three markers, blue indicates cells stained with isotype control (unstained). Marker relative fluorescent intensity is defined as the difference in mean intensity between the stained and unstained cell populations (middle panel). Purity is measure by estimating the proportion of stained cells (red) whose intensity is greater than the purity threshold (purple) $t = \mu_{stained} - 3 \times \sigma_{stained}$.

Variance component analysis and QTL mapping

I used linear mixed model implemented in `lme4` (Bates et al., 2015) package in R to characterise the observed variation in the relative fluorescent intensity measurements of the three macrophage markers. For each marker, I estimated the proportion of variance explained by differences between the iPSC lines (hereafter 'line effect') as well as the batch effect represented by the date when the cells were harvested, stained and measured on the flow

cytometer ('date effect'). I used the following `lme4` model specification: `intensity ~ (1|date) + (1|line)`.

I used FastQTL (Ongen et al., 2016) to test for association between relative fluorescent intensity and common genetic variants (minor allele frequency > 0.05, IMP2 > 0.7) in the +/- 200kb region around the corresponding genes: CD14, FCGR3A and FCGR3B for CD16, and MRC1 for CD206. I used measurements from 95 unique lines (donors) for QTL mapping. If a particular line had multiple measurements, then I picked one randomly. After permutation testing (n=10,000), I identified significant cis QTLs for CD14 and CD16 markers. Subsequently, I redid the variance component analysis for each marker and included the lead QTL variants into the model: 'intensity ~ (1|date) + (1|line) + (1|rs2569177) + (1|rs4657019)'.

3.3 Large-scale differentiation of macrophages for genomics assays

We aimed to develop a robust and standardised differentiation pipeline that would allow us to produce at least 3 million macrophages from each donor for four different experimental assays: (1) Flow cytometry (this chapter), *Salmonella* RNA-seq (Chapter 4), *Salmonella* ATAC-seq (Chapter 5) and acLDL RNA-seq (not described here). We relied on a previously published macrophage differentiation protocol (van Wilgenburg et al., 2013) that I compared to monocyte-derived macrophages in Chapter 2. The timeline of the differentiation protocol is illustrated in Figure 3.2 and the full details of the protocol are given in the Methods. Briefly, the main steps of the differentiation are (1) expansion of iPSCs in feeder-dependent medium (median 19 days), (2) embryoid body (EB) formation (3 days), (3) differentiating EBs into macrophage progenitors (median 27 days) and (4) harvesting and final differentiation of progenitors into macrophage (7 days). One attractive feature of this system is that differentiated EBs can be kept in culture for prolonged period of time and progenitors can be harvested in every 4-5 days making it possible to perform additional assays on the cells without increasing the amount of tissue culture needed for the initial steps of the differentiation (van Wilgenburg et al., 2013).

Although other protocols exist that can be used to differentiate macrophages in a shorter period of time (Zhang et al., 2015), a major advantage of our protocol is that the bulk of the differentiation and maintenance is performed in single medium containing only two cytokines (interleukin-3 (IL-3) and macrophage colony stimulating factor (M-CSF)) and the exact timing

between medium changes can be varied without significantly influencing differentiation success. This property made the protocol scalable to differentiating many iPSC lines in parallel without a large increase in complexity, because all of the dishes receive the same media and medium changes could be conveniently scheduled.

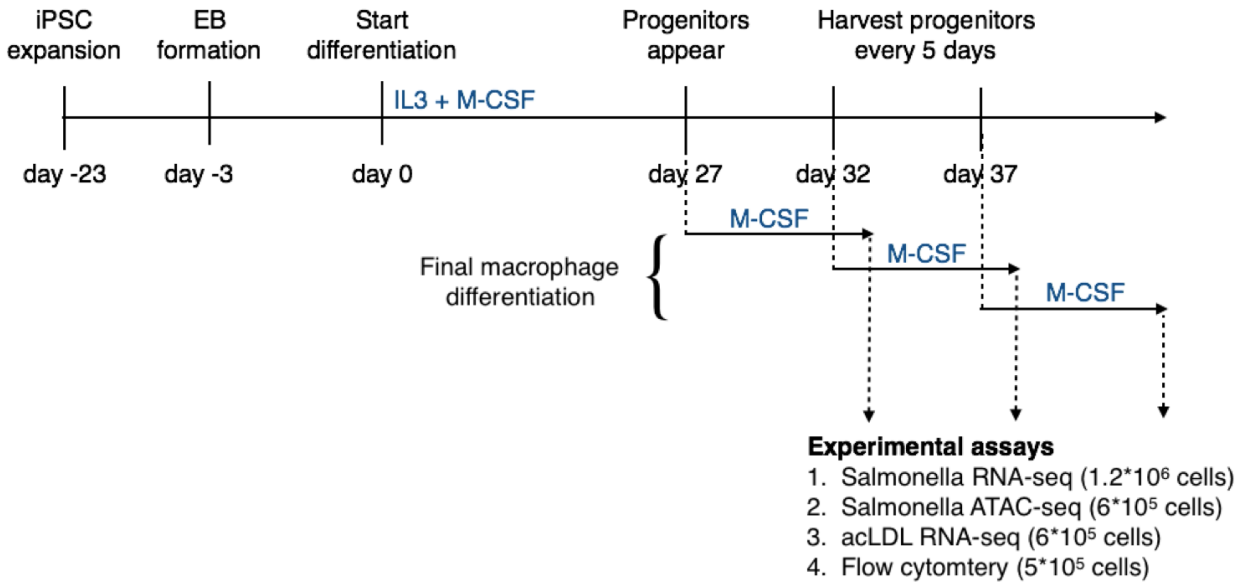

Figure 3.2: Timeline of macrophage differentiation from iPSCs. The protocol starts with the expansion of IPSCs followed by embryoid body formation. The bulk of the differentiation is performed in X-VIVO 15 media supplemented with IL-3 and M-CSF cytokines. The differentiation takes usually 4-5 weeks (median 27 days) until macrophages progenitors appear. During this time the medium has to be changed in every 4-5 days. Once the macrophage progenitors appear, they are harvested at every medium change and differentiated in the presence of M-CSF for another 7 days until the cells are ready for experimental assays.

We differentiated macrophages from batches of multiple iPSC lines in parallel. In addition to logistical convenience, this approach enabled us to estimate and control for batch-to-batch variation in gene expression and differentiation success measurements.

3.3.1 Variability in success rate

We performed 138 macrophage differentiation attempts from 123 different HipSci iPSC lines. We were able to successfully differentiate macrophages from 101/123 (82%) of the iPSC lines. Here successful differentiation is defined as obtaining at least some proportion of cells that exhibited characteristic spindle-like macrophage morphology. For 97/101 lines, we further

confirmed the expression of CD14, CD16 and CD206 macrophage cell surface markers with flow cytometry.

To understand what was responsible for the failed differentiation, we tried to re-differentiate 8 iPSC lines that had failed on the first attempt. Surprisingly, 7/8 failed also at the second attempt. This was over 6 times higher than the global 18% failure rate observed across all lines (Fisher's exact test $p = 0.002$), suggesting that there might be a line specific bias against macrophage differentiation. However, six of these lines (all of which failed) were all re-differentiated in the same month (January 2015), meaning that this observation might have also arisen from a shared batch effect. We note though, that 3/4 lines cultured concurrently with the 6 failed lines differentiated successfully into macrophages. Hence, this suggests that there might be a line-specific (or donor-specific) bias against macrophage differentiation but further experiments on more iPSC lines are needed to confirm this.

3.3.2 Variability in the duration of the differentiation

Throughout our experiments we observed considerable variation in the time from initial iPSC culture to the production of mature macrophages. This variation was influenced by a variety of experimental factors, most importantly whether the differentiation was started from live or frozen cells. Initially, we received live cells in feeder dependent media from Wellcome Trust Sanger Institute core facilities. These live cell cultures required only a single passage before EB formation could be initiated (Figure 3.3A). Subsequently, however, for operational reasons we switched to cryopreserved cells cultured either on feeder-dependent or feeder-free E8 medium. Since our attempts to differentiate macrophages directly from feeder-free iPSCs were not successful, we had to transfer feeder-free cells to feeder-dependent medium for at least two passages. This added approximately 7-10 days to the time required for initial iPSC culture and expansion. However, the total time needed for iPSC expansion was comparable for feeder-free and feeder-dependent cryopreserved cells, because thawing feeder-dependent iPSCs generally took much longer than thawing feeder-free iPSCs (Figure 3.3A). We did not observe any discrepancy in the differentiation success rate between IPSCs initially grown either on feeder-dependent or feeder-free media.

The median time from the start of the differentiation (3 days after EB formation) until the appearance of first macrophage progenitors was 27 days (Figure 3.3B), and 96% of the lines that successfully differentiated into macrophages did so within 40 days. Thus, for this protocol, a

40-day threshold provided a useful guideline for deciding when a differentiation attempt had failed and should be aborted. Final macrophage differentiation added another 7 days to the protocol and for logistical reasons we were not always able to perform the stimulation assays on the first batch of cells that we harvested. This increased the median time from differentiation start to cell lysis to 38 days (Figure 3.3C). We recorded this information for each cell line to assess retrospectively if the time spent in culture had an effect on the macrophage transcriptome.

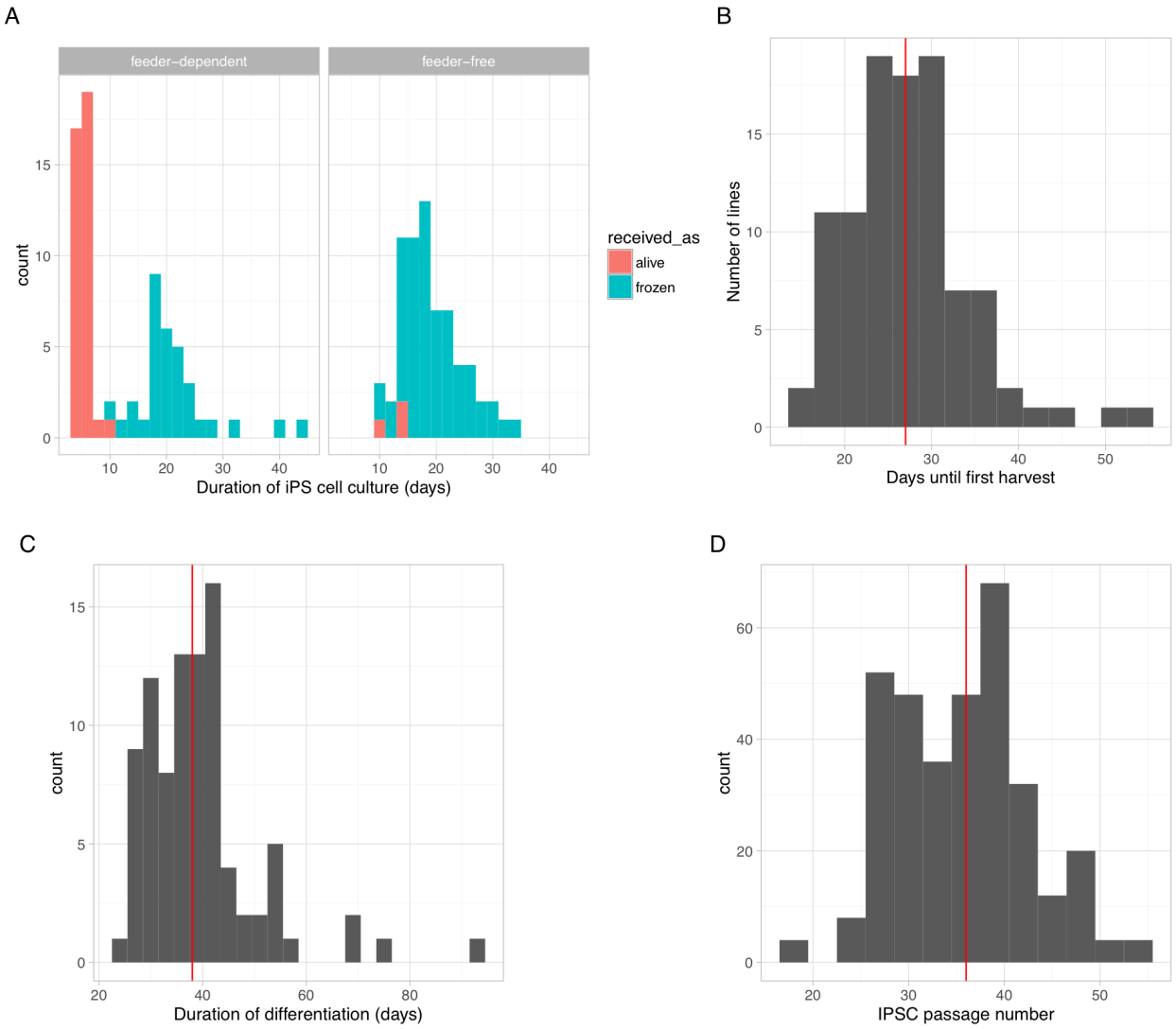

Figure 3.3: Variation in the duration of macrophage differentiation. (A) Duration of iPSC culture prior to the start of the differentiation. The two panels correspond to iPSC lines that were initially either on feeder-dependent medium or feeder-free medium. The colour represents whether the cell lines were received as live culture or cryopreserved stock. (**B)** Number of days

from the start of the differentiation until the harvest of first macrophages. Red line corresponds to the median of the distribution (27 days). **(C)** Histogram of the number of days from the start of the differentiation until the *Salmonella* infection experiment and cell lysis (median 38 days). **(D)** Histogram of the number of passages iPSCs had been propagated prior to the start of the differentiation.

3.3.3 Variability in cell numbers

Although we differentiated all lines in the same number of tissue culture plates, we observed an order of magnitude variation between lines in the mean number of macrophage progenitors produced per harvest (min 3×10^5, median 3×10^6, max 15×10^6) (Figure 3.4A). Most of the variation was likely caused by differences in the size and number of EBs per line, which was challenging to control during differentiations. Our approach to deal with this variation was to use more than minimally required cells for EB formation, thus ensuring that even differentiation with lower yield would produce enough cells for all of the planned experimental assays.

For the final macrophage differentiation, we always seeded 150,000 progenitors into a single well of a 6-well plate. However, due to variation in the fraction of adherent cells and their proliferation rate between iPSC lines, we observed substantial variation in the numbers of cells on the plate at the time of the stimulation assays. Since this variation was hard to control for experimentally (macrophages are strongly adherent cell type making them difficult to replate), we decided to measure the mean RNA concentration for each line as a proxy of the cell count (Figure 3.4B).

3.3.4 Variability in macrophage purity

Finally, we examined the purity of the differentiated macrophages. Despite not using cell sorting or other methods to experimentally enrich for macrophages, we found that 88% of the differentiations produced macrophages that were >90% pure based on the cell surface expression of CD14, CD16 and CD206 markers (Figure 3.4C). Although we did not use flow cytometry to directly select samples for RNA sequencing (flow cytometry was often performed after RNA had been collected), we found that only 4/84 of the selected samples had purity below 90% (Figure 3.4C).

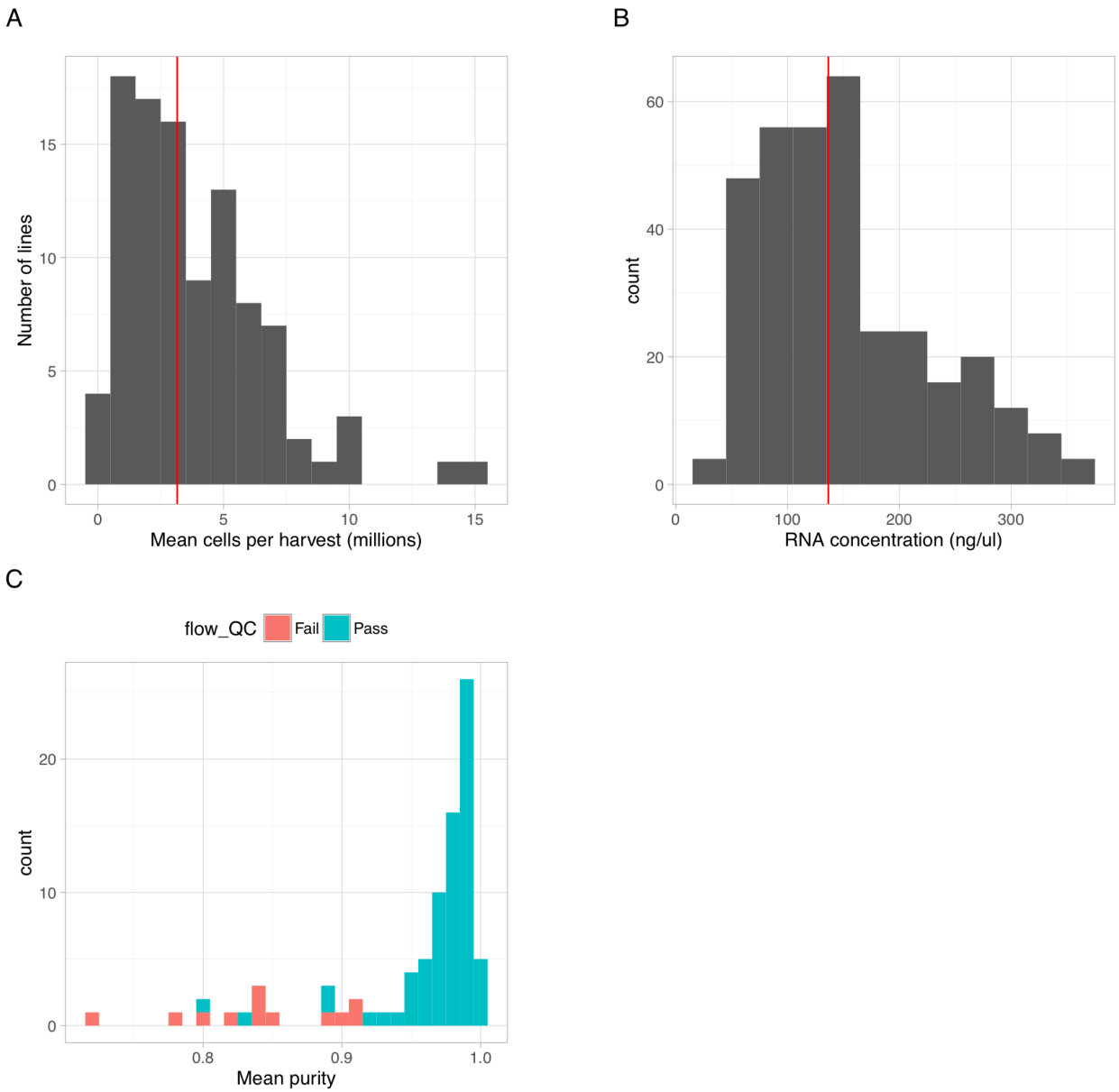

Figure 3.4: Distributions of some the key experimental variables the could influence the transcriptomes of differentiated macrophages. (A) Histogram of the mean number of cells obtained per harvest for all of the iPSC lines that successfully differentiated into macrophages. Red line corresponds to the median (3.16 million). **(B)** Histogram of the mean RNA concentration values for all of the cell lines across four experimental conditions. **(C)** Differentiated cells were stained with antibodies for CD14, CD16 and CD206 and the proportion

of cells staining positive for each of the three marks was estimated. The mean of the three marks was used as the purity score for each cell line. The figure shows the histogram of the purity scores across iPSC lines. Samples represented in green were used for RNA-seq experiments.

3.4 Variability in gene expression data

While many aspects of the differentiation might be variable between iPSCs lines, not all of them will have a significant effect on downstream macrophage gene expression levels. Thus, I decided to use variance component analysis to estimate the relative contribution of various biological and technical factors on macrophage gene expression levels.

3.4.1 Technical variability between RNA-seq samples

In addition to biological variability in the differentiation protocol described above, macrophage gene expression levels could also be influenced by technical variability in the way RNA samples were processed. The potential sources of variability that we identified were RNA extraction batch, RNA integrity, library construction batch, method of library construction used (manual or automatic) and sequencing chemistry used.

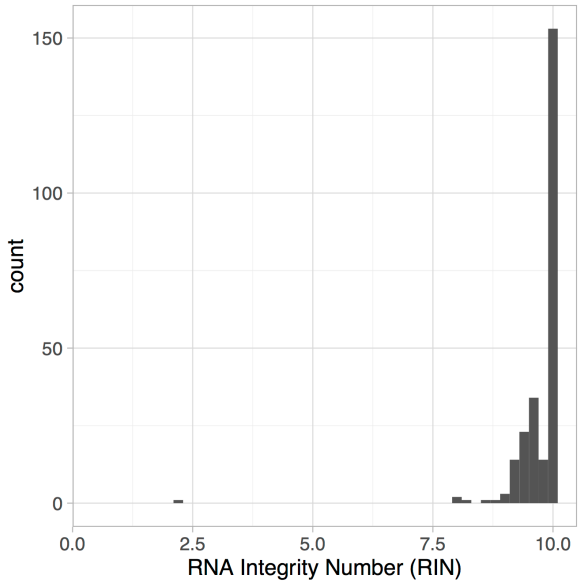

Figure 3.5: Distribution of RNA integrity (RIN) values for a subset of the RNA-seq samples.

Fortunately, I observed very little variation in RNA integrity between samples and for the vast majority of the samples the RNA integrity number (RIN) was greater than 9 out of 10 (Figure 3.5). However, I did observe some differences between automatic and manual library construction methods. First, the variability in total read coverage between samples was greatly reduced when the automatic protocol was used (Figure 3.6A). I also found that the automated protocol had lower GC content bias than the manual protocol which showed slight preference for low GC content fragments over high GC content fragments (Figure 3.6B).

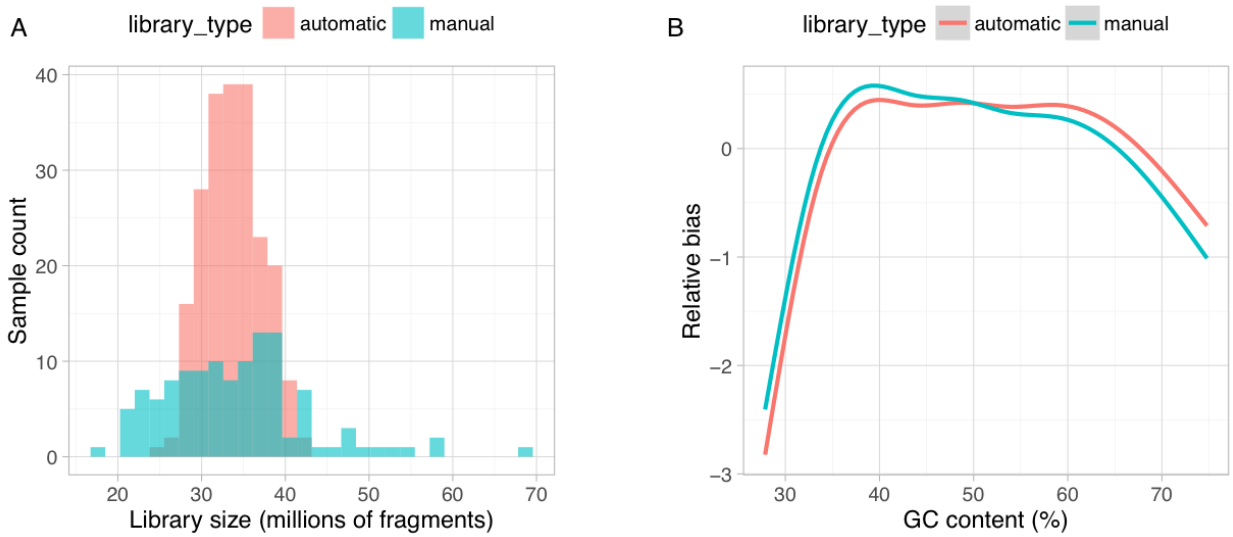

Figure 3.6: Comparison of manual and automated RNA-seq library construction protocols. (A) Histogram of total library size distribution for samples prepared either with manual or automatic protocol. **(B)** Mean GC content bias for the two library construction protocols. GC content bias was estimated from the raw read counts using the cqn (Hansen et al., 2012) package in R.

3.4.2 Variance component analysis of the RNA-seq data

Variance component analysis is a powerful approach to estimate the relative importance of various known experimental factors in an unbalanced experimental design (Rouhani et al. 2014; Kilpinen et al. 2016). When applied to our dataset, variance component analysis revealed that most of the variance in gene expression was explained by the three experimental stimuli: *Salmonella* infection (32.9%), IFNɣ stimulation (15.5%) and interaction between the two (11.4%) (Figure 3.7), highlighting the plasticity of the macrophage transcriptome in response to strong immunological stimuli. The second largest amount of variance explained (7.7%) was attributed

to differences between cell lines (hereafter ‘line effect’) while all of technical factors explained significantly less variance

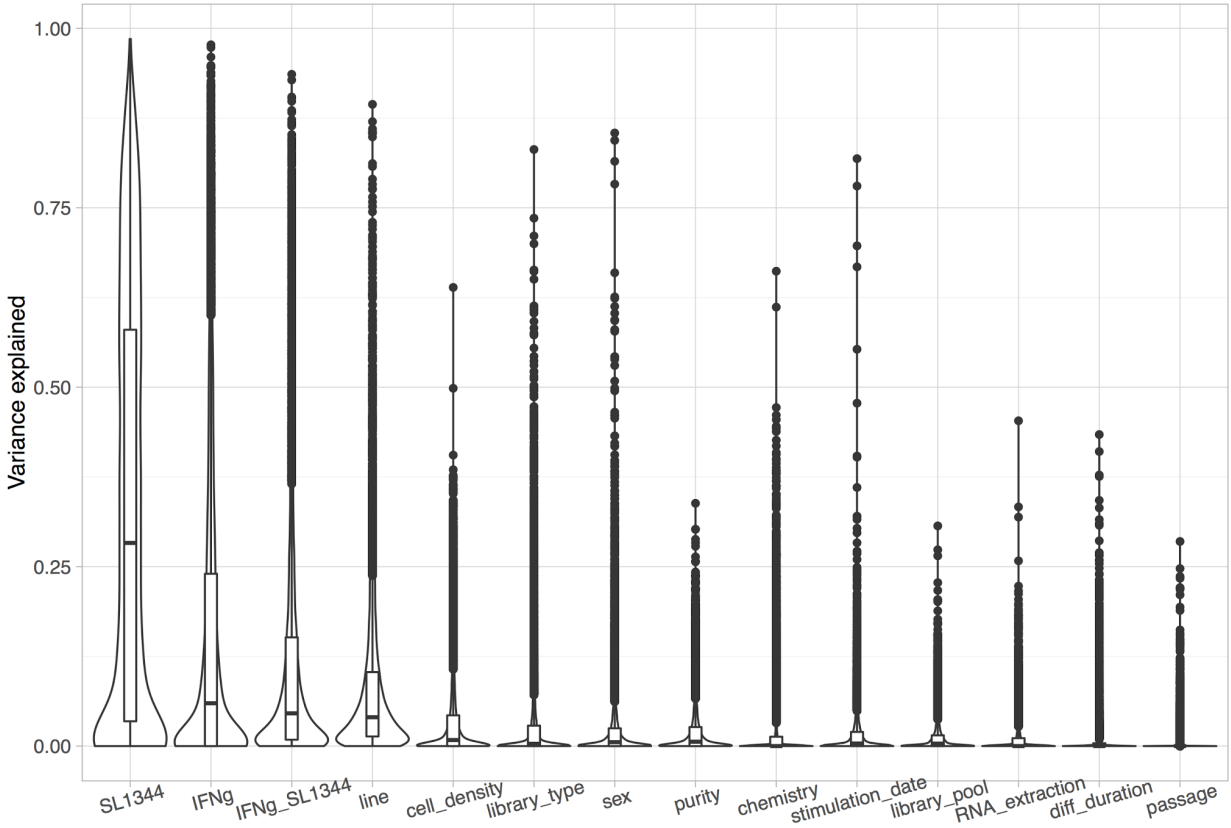

Figure 3.7: Variance component analysis of all four conditions in a joint model. We used a linear mixed model to estimate the proportion of variance explained by 14 different factors in the expression levels of 15,797 expressed genes (see Methods). For each factor on the x-axis, the violin plot shows the distribution of variance explained by that factor across all expressed genes. Factors are ordered by the mean variance explained across all genes.

To see how the relative contribution of the weaker technical factors varied between conditions, I performed variance component analysis in each of the four conditions separately. Now that the differences between stimulations were controlled for, most of variance was explained by RNA-seq library type (automatic vs manual), cell density, sex and purity of the macrophage population (Figure 3.8A) and the estimates for most of the factors were similar in all four conditions. The large contribution of library type is likely to be at least partially explained by differences between GC bias reported above (Figure 3.6B). The date when macrophages were

stimulated with IFNɣ and infected with *Salmonella* ('stimulation date') explained almost double the variance in *Salmonella* and IFNɣ + *Salmonella* conditions than in naive and IFNɣ conditions (Figure 3.8B). This is probably because live *Salmonella* culture was prepared fresh for each day of infections whereas IFNɣ originated from single-use frozen aliquots. Indeed, both of the *Salmonella* conditions had an excess of highly variable genes compared to naive and IFNɣ conditions (Figure 3.8C), indicating that *Salmonella* batch introduced additional variability into the data. Finally, the passage number of the iPSCs prior to differentiation (Figure 3.3D) and the total duration of the differentiation (Figure 3.3C) (after accounting for differences in purity) explained less than 3% of the variance, suggesting that controlling for these factors during differentiation is less important.

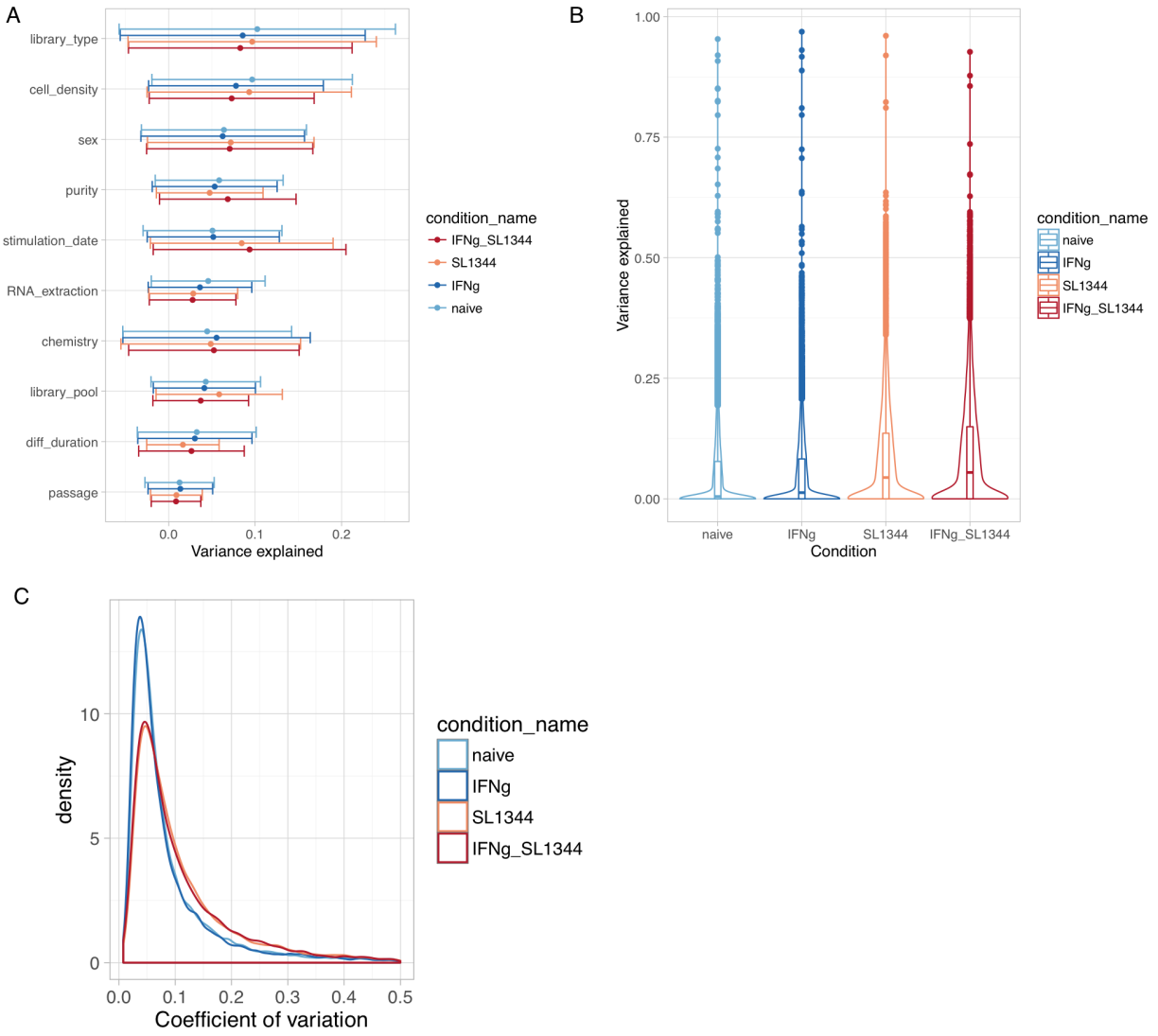

Figure 3.8: Variance component analysis of each of the four conditions separately. (A) Variance explained by ten technical factors in each of the four conditions. Points correspond to mean across all genes and horizontal lines represent standard deviation. Note that the lines are not true confidence intervals, because variance explained cannot be negative. **(B)** Proportion of variance explained by the date when the macrophages were stimulated with IFNɣ and infected with *Salmonella* ('stimulation date' effect). The violin plots show the distributions of the variance explained estimates across all genes in four experimental conditions. **(C)** Distribution of coefficient of variation (standard deviation / mean) for 10,000 most highly expressed genes in each condition.

3.4.3 Detecting hidden sources of variation

A complementary approach to dissect sources of variability in a large gene expression dataset are latent variable models (Leek and Storey, 2007; Stegle et al., 2010). Latent variable models are especially useful when the relevant covariates are not known beforehand or when they have not been measured accurately (Parts et al., 2011; Stegle et al., 2010). I applied PEER (Stegle et al., 2010) to the RNA-seq data from each condition to detect hidden sources of variation that affect many genes at the same time. I then calculated the proportion of variance explained by each hidden factor in each of the four experimental conditions (Figure 3.9). Note that PEER does not report residual variance and as a result these estimates are not directly comparable to the estimates from variance component analysis above. I found that in the naive cells 90% of the variance captured by PEER was explained by the first factor. Although macrophage purity and cell density (mean RNA concentration) were both correlated with the first factor (Figure 3.10A-B), in a joint linear model these two known covariates could explain only 42% of the variance captured by the first factor. This illustrates that PEER is able to capture additional variability beyond what can be explained by known covariates.

The second PEER factor explained an additional 3.1% of the variance and was correlated with the RNA-seq library type (Figure 3.10C) ($r = 0.79$, $p < 2.2 \times 10^{-16}$). However, as shown on Figure 3.6B, one of the differences between automatic and manual RNA-seq protocol was difference in GC bias. The quantile normalised gene expression values that we used as input to PEER were already corrected for sample-specific differences in GC bias. Therefore, the amount of variance explained by the second PEER factor might be higher in uncorrected samples.

I noticed that in stimulated conditions factors 2-5 explained much more variance than they did in the naive condition (Figure 3.9). This was especially prominent after *Salmonella* infection where ~40% of variance was explained by factors 2-5 (7.5% in naive). One possible interpretation is that stimulating cells introduces additional independent sources of variability ('batch effects') that are then captured by PEER as additional factors. This is consistent with the excess of highly variable genes observed after *Salmonella* infection (Figure 3.8C) and more variance explained by stimulation date (Figure 3.8B) reported above.

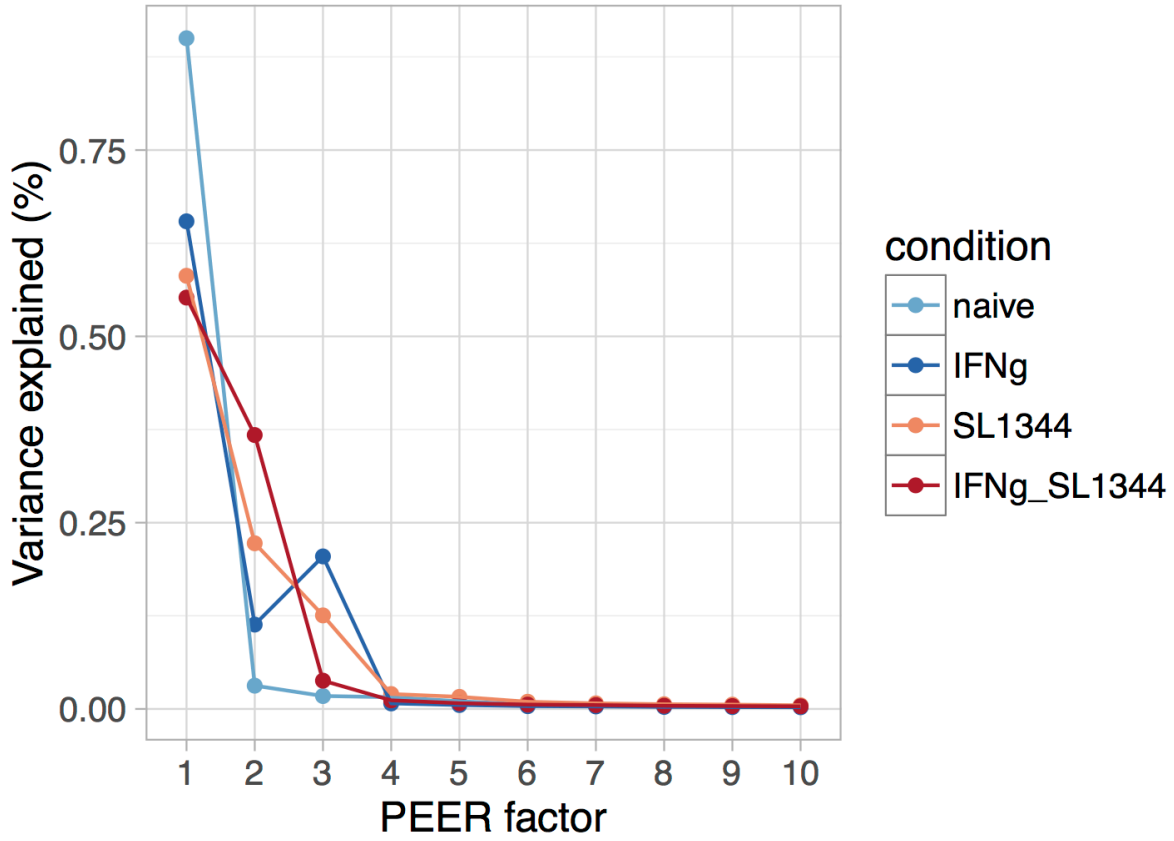

Figure 3.9: Proportion of variance explained by the first 10 PEER factors in each experimental condition. PEER was run on each condition separately, which mean that the factor names do not necessarily correspond to the same sources of variation in each condition.

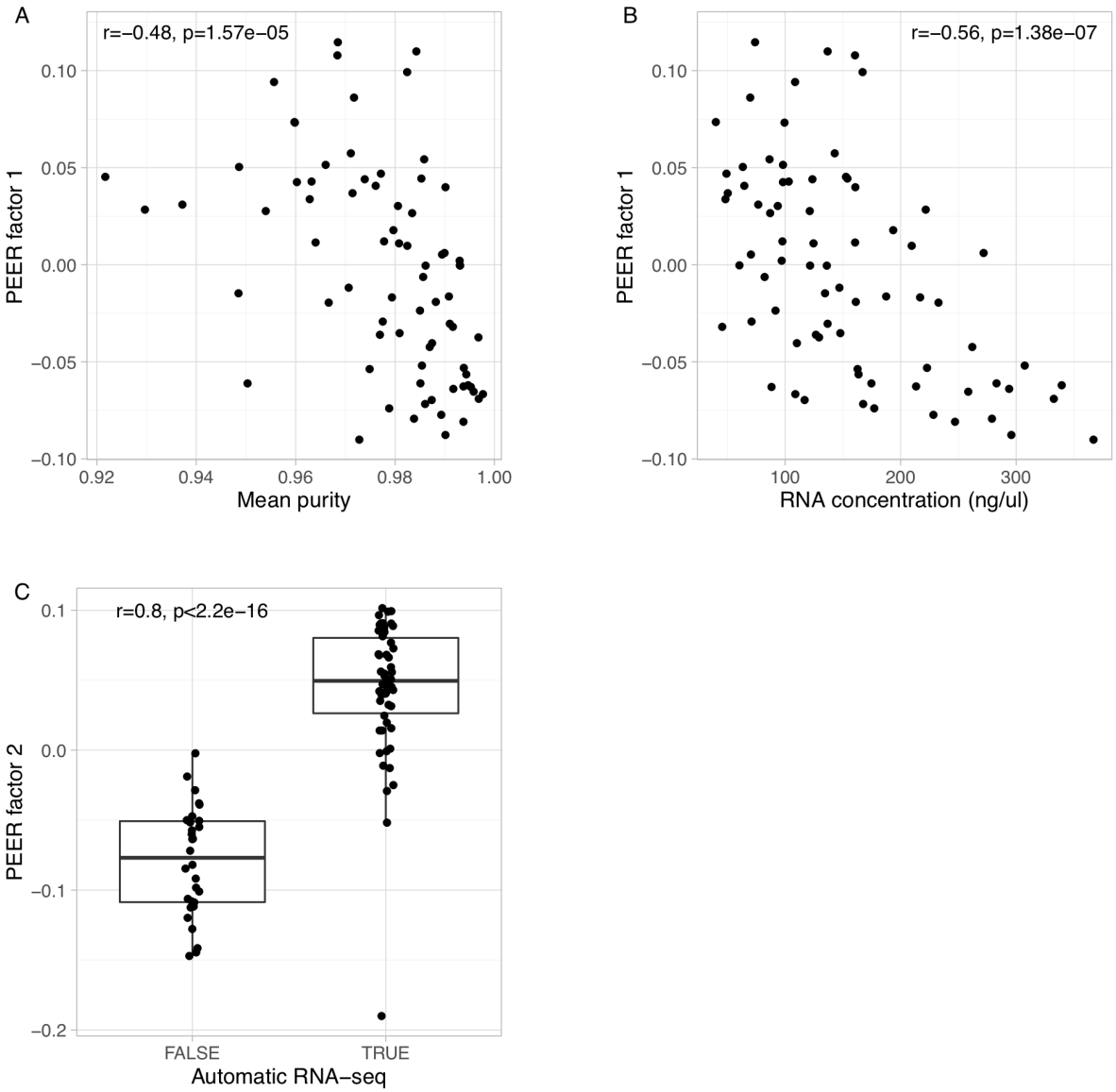

Figure 3.10: Correlation between first two PEER factors and experimental variables. (A) Correlation between mean purity and PEER factor 1 **(B)** Correlation between RNA concentration and PEER factor 1 **(C)** Correlation between RNA-seq library construction protocol (automatic vs manual) and PEER factor 2.

3.4.4 Reproducibility of differentiation

To assess how reproducible gene expression profiles were between differentiations, I analysed RNA-seq data from multiple independent differentiations from three different donors (three differentiations from donor ffdj and two from donors fpdl and ougl). We performed the same

stimulation experiments and RNA-seq on all of the samples. Although the differentiations were performed over the course of 10 months, I found that in the naive condition the samples clearly clustered together by donor (Figure 3.11), indicating that donor specific effects on gene expression are reproducible between differentiations. The clustering was not as clear in the stimulated conditions, possibly because of stronger batch effects induced by stimulation. For the ffdj donor we know that two of the differentiations were from the same iPSC line (samples fpdj_A and fpdj_A_2) whereas the third (nibo_A) was from a different line. For ougl and fpdl donors we unfortunately do not know if the two differentiations were from the same line or different lines, because we received these lines twice due to accidental sample swaps upstream and we only discovered the duplicate samples after matching genotypes in the RNA-seq data to the VCF file.

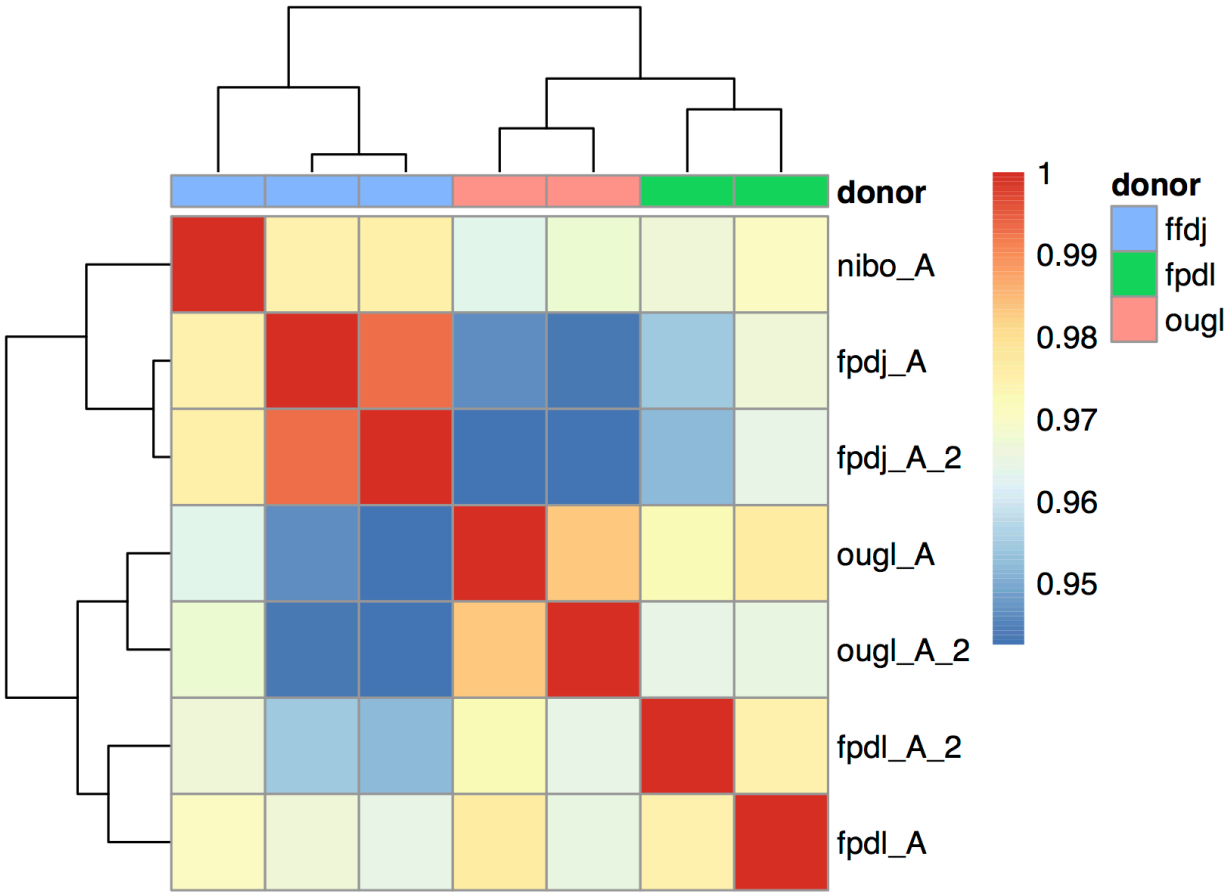

Figure 3.11: Gene expression reproducibility between independent differentiations. The heat map shows the Spearman correlation of gene expression profiles from 7 independent differentiations from 3 different donors.

3.5 Variability in cell surface marker expression

In addition to gene expression data, we also wanted to understand what is responsible for the variance in the cell surface expression of macrophage markers. Specifically, we wanted to know if, on average, the fluorescent intensity measurements of the same line on different days were more similar to each other than the measurements of different lines on the same day.

We measured cell surface expression of CD14, CD16 and CD206 macrophage markers in 97 different cell lines. This included 19 cell lines where duplicate measurements were obtained on different days. We processed and measured a median of four cell lines in a single batch on the same day. This allowed me to use a linear mixed model to estimate the relative proportion of variance explained by cell line and date of the assay (batch) for each of the three markers. I found that 62% of the variance in CD14 surface expression and 52% of the variance in CD16 surface expression was explained by the line effect and almost no variance was attributed to the date of the assay (Figure 3.12A). On the other hand, 64% of the variation in CD206 measurement was explained by the date of the assay and there was almost no line effect, suggesting that this antibody might have been more susceptible to technical variation. Between 25-50% of the variance remained unexplained for all three marks.

Next, I tested whether there was a genetic basis for the observed variation in the surface expression of CD14, CD16 and CD206 by performing QTL mapping for each of the three markers in +/- 200kb cis window around the corresponding genes (CD14, FCGR3A and FCGR3B for CD16, and MRC1 for CD206). I detected a very strong association between CD14 surface expression and rs2569177 variant (MAF = 0.24) located 19 kb upstream of the CD14 gene (permutation FDR = 2.7×10^{-11}) (Figure 3.12B). I also detected a weak association between CD16 expression and rs4657019 variant (MAF = 0.28) located 120 kb upstream of the FCGR3A gene (permutation FDR = 0.047) (Figure 3.12C). There was no significant QTL for CD206 consistent with only a small fraction of the variance being attributed to differences between lines. I redid the variance component analysis with the two CD14 and CD16 QTL SNPs included in the model. I found that the CD14 QTL explained most of the variation in CD14 intensity that was previously attributed to line effect (3.12D). On the other hand, the CD16 QTL explained only ~⅓ of the CD16 line effect, suggesting that there might be additional cis or trans QTLs for this protein that we were unable to detect because of our small sample size (Figure 3.12D).

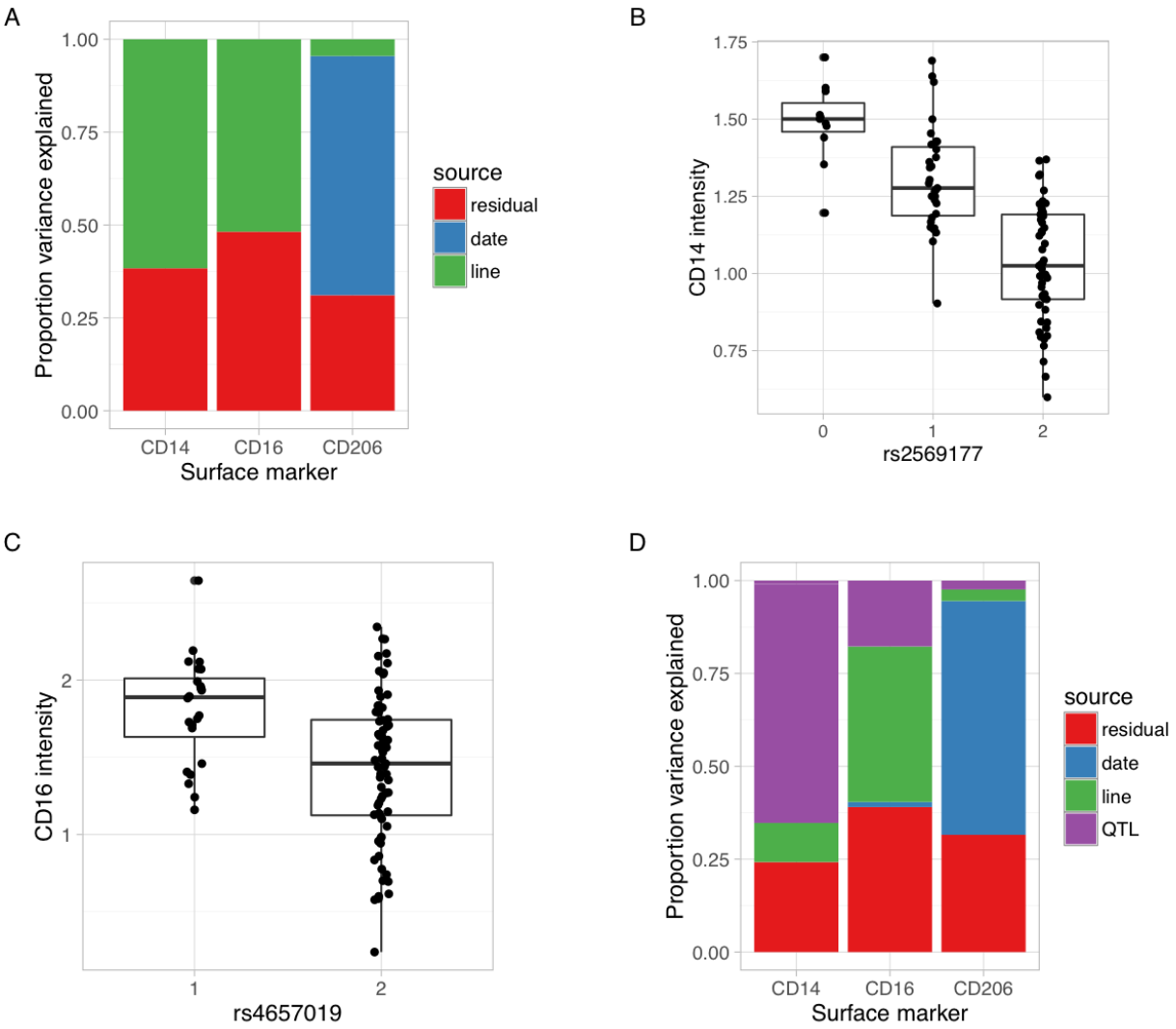

Figure 3.12: Variance of macrophage cell surface marker expression. (A) Variance of cell surface expression of CD14, CD16 and CD206 partitioned into three components: (1) iPSC line from which the macrophages were differentiated; (2) date of the flow cytometry assay; (3) residual variation. **(B)** Fluorescent intensity of CD14 cell surface expression stratified by the genotype of the lead QTL variant (FDR < 2.7×10^{-11}). **(C)** Fluorescent intensity of CD16 cell surface expression stratified by the genotype of the lead QTL variant (FDR < 0.048). **(D)** Variance partitioning after including CD14 and CD16 lead QTL variants into the model.

3.6 Discussion

In this chapter, we performed 138 macrophages differentiation attempts from 123 unique iPSC lines and we were able to successfully differentiate 101 (82%) of them. This makes our study

one of the largest long term directed differentiations of human iPSCs into another cell type. Extensive documentation of the differentiation attempts allowed us to characterise the extent of normal variation in multiple aspects of the differentiation protocol such as success rate, duration, yield and purity of the resulting macrophage population. We have shown that this differentiation protocol is highly reproducible at the level of gene expression, works on most iPSC lines and can be scaled to differentiate large numbers of cells in parallel.

An important open question is what underlies variability in iPSCs differentiation potential; are these genetic differences between donors, differences between clonal iPSC lines from the same donor or technical batch effects between independent differentiation attempts. Our experimental design of differentiating only one line per donor was optimised for detecting the maximal number of gene expression QTLs. As a result, we were not able to distinguish between donor and line effects. However, our observation that repeated differentiations are much more likely to fail for lines that failed the first differentiation than for lines that succeeded the first differentiation does suggest that there are some differences between iPSC lines (either genetic or epigenetic) that influence differentiation success.

We also collected RNA-seq data from most of the differentiated lines in four experimental conditions. Combining gene expression data with extensive metadata from the differentiations in a linear mixed model allowed us to identify important factors contributing to gene expression variation in iPSC-derived macrophages. In particular, we highlighted the importance of controlling for cell density and cell purity when performing genomics assays on iPSC-derived cells. The large effect of macrophage purity was unexpected, because the majority of the samples were already over 95% pure and we had discarded all samples that were less than 90% pure prior to RNA sequencing. On visual inspection the contaminating cells seemed larger than macrophages, and thus could have contributed relatively more RNA to the pool. We also observed that the date of stimulation explained double the variance in conditions where live *Salmonella* was used to infect cells compared to naive and IFNɣ conditions, highlighting an important trade-off between physiologically more accurate live infections and inherently less variable stimulations with well-defined molecular signals (such as IFNɣ and LPS).

Finally, we showed that variation in the intensity of expression of commonly used macrophage markers CD14 and CD16 on the cell surface is driven by common genetic variants. This was especially pronounced for CD14, where we identified a common genetic variant 19 kb upstream

of the gene that could explain almost all of the line-to-line variation in CD14 expression. Thus, it is important to take into account natural genetic variation when comparing the expression of cell type specific markers between primary cells, iPSC-derived cells and embryonic stem cell-derived cells. This is especially important because these different cell types can rarely be obtained from genetically matched donors. For example, CD14 has previously been highlighted as variably methylated gene in human ESCs and variably expressed in differentiated macrophages (Bock et al., 2011). The authors attributed this variability to defective methylation in some ESCs that interfered with macrophage differentiation. However, our results suggest that much of this variability is caused by segregation of a common genetic polymorphism. Flow cytometry on cell surface markers is also commonly used to quantify the relative abundance of different cell types in a tissue (such as blood). It is therefore important to take the natural variation in the expression of these markers into account when designing the experiments and setting threshold values so as not to mistake differences in cell surface expression of marker gene as differences of cell type proportion.

An important area for future work will be to optimise the differentiation protocol to work directly on feeder-free iPSCs without transferring them to feeder cells. This has the potential to greatly reduce the time and work needed for iPSC expansion prior to differentiation which currently takes ~20 days. With newer RNA-seq and chromatin assay requiring fewer cells, there is also potential to miniaturise the differentiation protocol making it feasible to differentiate hundreds of IPSCs in parallel. Here, alternative embryoid body formation protocols can be trialled (e.g. AggreWell plates (van Wilgenburg et al., 2013)) that have the potential to reduce variability in macrophage yield between differentiations.

4 Genetics of gene expression in macrophage immune response

Collaboration note

The macrophage differentiation work in this chapter was performed in collaboration with Julia Rodrigues who was a research assistant in Daniel Gaffney's lab at the time. I designed the experiments, performed *Salmonella* infection and IFNɣ stimulation assays, took care of sample logistics and performed all of the data analysis. Julia was mainly responsible for tissue culture required for macrophage differentiation and preparing cells for stimulation experiments. Subhankar Mukhopadhyay and Gordon Dougan provided valuable feedback in designing and optimising *Salmonella* infection and IFNɣ stimulation conditions.

4.1 Introduction

Genetic differences between individuals can have a major impact on how immune cells respond to environmental stimuli, such as the amount of cytokines they produce after infection (Li et al., 2016a). A number of studies have looked at the impact of genetic variation on cellular responses to different (immunological) environmental stimuli via the regulation of gene expression. Most studies have used either primary monocytes purified from peripheral blood (Fairfax et al., 2014; Kim et al., 2014) or monocyte-derived dendritic cells (Barreiro et al., 2012; Lee et al., 2014). While powerful, one limitation of primary cells is that the amount of material that can be obtained from a single individual is limited. This in turn limits both the number of assays that can be performed on cells from a single individual as well as the number of stimuli that can be studied. This is especially important because for any given cell type there can be tens of different relevant stimuli or combinations of stimuli, each one potentially revealing a different set regulatory variants that are otherwise hidden in the unstimulated state.

A major advantage of cell lines is that the number of cells is essentially unlimited meaning different phenotypes can be collected from the same set of individuals over time. In this respect,

human lymphoblastoid cell lines (LCLs) have been very powerful. For example, over the years LCLs from the Yoruban population have been profiled on many different levels including RNA sequencing (Pickrell et al., 2010), ribosome profiling (Battle et al., 2015), proteomics (Battle et al., 2015), DNase-seq (Degner et al., 2012) and ChIP-seq (Grubert et al., 2015; McVicker et al., 2013) and in multiple cases integrating old data sets with new ones has provided new biological insight (Li et al., 2016c). However, since LCLs are immortalised by infection with Epstein-Barr virus they are not a suitable model to study the response to different immunological stimuli.

A promising approach to overcome the limitations of LCLs are human induced pluripotent stem cells (iPSC) that have recently been derived from large collection of unrelated individuals (Kilpinen et al., 2016). In Chapter 3, we showed that iPSCs can be reliably differentiated into macrophages on a scale necessary for QTL mapping studies. The aim of this chapter is to first characterise how well iPSC-derived macrophage are able to recapitulate known aspects of macrophage response to *Salmonella* infection and IFNɣ stimulation. Subsequently, I want to identify common genetic variant that influence gene expression and mRNA processing (promoters, splicing, poly-adenylation) in each of the four conditions and assess how condition specific they are.

We obtained RNA-seq data from 84 iPSC-derived macrophage lines in four immunological conditions: (1) naive, (2) 18-hour IFNɣ stimulation, (3) 5-hour *Salmonella* infection (4) 18-hour IFNɣ stimulation followed by 5-hour *Salmonella* infection. We chose these stimuli, because they are known to activate distinct downstream signalling pathways. Lipopolysaccharide (LPS) and other components on the surface of *Salmonella* cell wall are recognised by macrophage Toll-like receptors (TLRs) that lead to activation of NF-kB and AP-1 signalling pathways (Takeuchi and Akira, 2010). TLR4 activation by LPS also leads to specific activation of the interferon response factor 3 (IRF) transcription factor and downstream antiviral response genes (Doyle et al., 2002). IFNɣ, on the other hand, is specifically recognised by the IFNɣ receptor that leads to phosphorylation and activation of the STAT1 transcription factor (Platanias, 2005). Moreover, pre-stimulating macrophages with IFNɣ prior to bacterial infection leads to enhanced microbial killing and stronger activation of inflammatory response by Toll-like receptors (TLRs) (Hu and Ivashkiv, 2009; Qiao et al., 2013; Su et al., 2015). There are at least two potential mechanisms that could be responsible for the enhanced response: (1) IFNɣ pre-stimulation can prime certain enhancers so that they can now be bound by *Salmonella*-activated TFs (Qiao et al., 2013), (2) IFNɣ priming can change the pool of active TFs available in the cell, this can facilitate new types

of collaborative binding between Salmonella-activated TFs and IFNɣ-activated TFs similarly to PU.1 binding to latent enhancers in mouse macrophages activated by IFNɣ stimulation (Ostuni et al., 2013).

With 84 samples, we were also highly powered to detect differential expression between the four conditions. By comparing the differentially expressed genes to the literature, I was able to show that iPSC-derived macrophages predominantly activated expected genes and pathways in response to the three stimuli, indicating that they are a suitable model to study human macrophage immune response. The main aim of the chapter was to uncover genetic variants that regulate gene expression on gene and transcript level. I used two complementary models to identify gene expression quantitative trait loci (eQTLs) and assess their condition specificity. I also developed a novel approach to pre-process transcript annotations prior to transcript ratio QTL (trQTL) mapping that increased interpretability of trQTLs and allowed me to detect more independent trQTLs per gene than established methods. I identified thousands of eQTLs and trQTLs across conditions and estimated that ~25% of them were condition specific. Consequently, a large proportion of the condition-specific QTLs were 'hidden' in the naive state, highlighting the importance of studying many different stimuli to uncover potential QTLs underlying disease associations. Although I was able to detect similar numbers of eQTLs and trQTLs across conditions, I found that eQTLs and trQTLs for the same genes were largely independent from each other, indicating that ignoring transcript-level variation can miss many genetic effects. Finally, I uncovered considerable heterogeneity in the QTLs discovered by different computational approaches. This was especially true for trQTLs because alternative transcripts are still poorly annotated. I was able to show that both macrophage eQTLs and trQTLs were enriched for GWAS hits for Alzheimer's disease, lipid traits and multiple autoimmune disorders. Together, these results highlight that iPSC-derived macrophages are a promising cell culture-based system to study condition-specific regulatory variation.

4.2 Methods

4.2.1 Gene expression analysis

Full details of the macrophage differentiation protocol, stimulation assays, RNA-seq experimental procedures, read alignment and gene expression quantification are presented in Chapter 3. I used the quantile normalised gene expression values from the cqn (Hansen et al.,

2012) package for clustering, eQTL mapping with linear models as well as for visualisation. For count-based methods such as DESeq2 (Love et al., 2014) and RASQUAL (Kumasaka et al., 2016) I used the raw read count data directly.

Differential expression analysis

I included 15,797 genes whose mean expression in at least one of the conditions was greater than 0.5 transcripts per million (TPM) into our differential expression analysis. For each gene, I used likelihood ratio test (test = “LRT”) implemented in DESeq2 (Love et al., 2014) v1.10.0 to test if a model that allowed different mean expression in each condition was a better fit to the data than a null model assuming the same mean expression across conditions. I used 1% Benjamini-Hochberg FDR threshold to identify differentially expressed genes. I further filtered the genes by requiring them to be at least 2-fold differentially expressed between the naive condition and one of the stimulated conditions resulting in 8758 differentially expressed genes.

To identify differentially expressed genes with specific expression patterns, I calculated mean quantile-normalised expression level in each condition and standardised the mean expression values across conditions to have zero mean and unit variance. I then used c-means fuzzy clustering implemented in MFuzz v.2.28 (Kumar and E Futschik, 2007) package with parameters ‘c = 9, m = 1.5, iter = 1000’ to assign the genes into 9 clusters. The number of clusters was chosen iteratively by trialling different numbers and observing which ones led to stable clustering results from independent runs. I ranked the genes in each cluster by their fold change and used g:Profiler (Reimand et al., 2016) R packages to identify pathways and Gene Ontology (GO) categories enriched in each cluster.

Detecting hidden confounders with PEER

To detect hidden confounders in gene expression, I applied PEER (Stegle et al., 2012) on each condition separately allowing for at most 10 hidden factors. As discussed in Chapter 3, I found that the first 3-5 factors explained the most variation in the data and the others remained close to zero.

4.2.2 Gene expression QTL mapping

Preparing genotype data

I obtained imputed genotypes for all of the samples from the HipSci project (Kilpinen et al., 2016). I used CrossMap (Zhao et al., 2014) v0.1.8 to convert variant coordinates from GRCh37 reference genome to GRCh38. Subsequently, I filtered the VCF file with bcftools v.1.2 (http://samtools.github.io/bcftools/) to contain only bi-allelic variants (both SNPs and indels) with IMP2 score > 0.4 and minor allele frequency (MAF) > 0.05 in our 84 samples. This VCF file was used for all subsequent analyses. The genotype data for 52 managed access lines is available from the European Genome-phenome Archive (EGA) (EGAD00010000773), the data for the remaining 34 open access lines is deposited in the European Nucleotide Archive (ENA) (PRJEB11749). The VCF file was imported into R using the SNPRelate (Zheng et al., 2012) R package.

Detecting eQTLs using linear model

I used linear regression implemented in the fastQTL (Ongen et al., 2016) software to map cis eQTLs in each experimental condition. I used the "--permute 100 10000" option to obtain permutation p-values for each gene. The size of the cis windows was set to +/-500 kb around the gene. I used sex and the first six PEER factors as covariates in the model. I picked single most significantly associated variant for each gene and used Benjamini-Hochberg correction to identify genes with at least one significant eQTL at 10% FDR level ('eGenes').

Quantifying allele-specific expression

I used ASEReadCounter (Castel et al., 2015) from the Genome Analysis ToolKit (GATK) to count the number of allele-specific fragments overlapping each variant. I used the following flags with ASEReadCounter: '-U ALLOW_N_CIGAR_READS -dt NONE --minMappingQuality 10 -rf MateSameStrand'. I removed indels from the VCF file prior to quantifying allele-specific expression because they are not supported by the RASQUAL model.

Detecting QTLs using RASQUAL

I wrote a collection of python scripts and a rasqualTools R package to simplify running RASQUAL on large number of samples and work with large RASQUAL output files. This software is available on GitHub (https://github.com/kauralasoo/rasqual). I used the vcfAddASE.py script to add allele-specific counts calculated in the previous step into the VCF

file. I ran RASQUAL (Kumasaka et al., 2016) independently for each experimental condition using sex and first two PEER factors as covariates. In contrast to standard linear model, covariates seemed to have only a minor effect on the number of eQTLs detected by RASQUAL. I only included variants that were either in the gene body or within 500 kb upstream or downstream of the gene. I specified '--imputation-quality > 0.7'. As a result, variants with imputation quality of < 0.7 were used as feature SNPs in allele-specific analysis but were not considered as possible causal variants. I also used RASQUAL's GC correction option to correct for sample-specific GC bias in the gene-level read count data. To correct for multiple testing, I picked one minimal p-value per gene, used eigenMT (Davis et al., 2016) to estimate the number of independent tests performed in the cis-region of each gene and then performed Bonferroni correction to obtain the corrected p-value. I further performed Benjamini-Hochberg FDR correction on the Bonferroni-corrected p-values to account for multiple testing between features and defined associations with FDR < 0.1 as significant.

Comparing RASQUAL and FastQTL results

To compare RASQUAL and FastQTL, I focussed on genes that were not filtered out by RASQUAL because of zero read count. Since performing thousands of genome-wide permutations was not feasible for RASQUAL, I only computed nominal p-values for the lead eQTL variant for each gene from both methods. I estimated the number of independent variants in the cis region of each gene with eigenMT (Davis et al., 2016) and then performed Bonferroni correction on gene level using the eigenMT estimates. Subsequently, I used Benjamini-Hochberg FDR correction to account for the number of genes tested and identified the genes that had a significant eQTL at 10% FDR. The eigenMT based FDR threshold was more conservative than permutation-based FDR normally used for FastQTL as reported in the eigenMT paper (Davis et al., 2016).

Detecting condition-specific QTLs with a linear model

In each condition, I first identified all features (genes or intron clusters) and corresponding lead variants that displayed significant association at 10% FDR level. These were identified either using RASQUAL (gene expression) or linear regression (intron excision ratios). For each feature, I then only kept independent lead variants (R^2 < 0.8). Finally, I used all independent pairs of features and corresponding lead variants to test if the QTL effect size was significantly different between conditions. This was equivalent to testing the significance of the interaction

term between condition and lead QTL variant for each feature. Specifically, I used ANOVA to compare two models for each gene-lead SNP pair:
H_0: expression ~ genotype + condition + covariates
H_1: expression ~ genotype + condition + genotype:condition + covariates
I calculated the p-value of rejecting H_0 and performed Benjamini-Hochberg FDR correction to identify condition-specific QTLs that were significant at 10% FDR level. For both gene expression and alternative transcription analysis, I used the same normalised data sets and covariates that were used for QTL mapping in each condition separately.

Filtering and clustering QTLs based on effect size

I extracted the RASQUAL eQTL effect size estimates π for each gene-variant pair in each condition and converted them to $\log_2$ fold changes between the two homozygotes using the formula $\log_2FC = -\log_2(\pi/(1-\pi))$. For an eQTL to be considered condition specific I required the difference in $\log_2FC$ between naive and any one of the stimulated conditions to be greater than 0.32 (~1.25 fold). I used k-means clustering to identify groups of eQTLs that had similar condition-specific patterns. For each eQTL, I divided the $\log_2FC$ values in each condition by the maximal $\log_2FC$ value observed across conditions. This scaling was necessary to make eQTLs with different absolute effect size comparable to each other for the k-means algorithm.

4.2.3 Alternative transcription analysis

I used three complementary approaches to quantify transcript expression in our samples. First, I quantified the expression levels of all known Ensembl transcripts. Secondly, I constructed alternative transcription events from known transcript annotations and quantified their relative expression. Finally, I used an annotation-free approach to quantify the rates of intron excision. All of these quantification approaches were subsequently used to identify transcript ratio QTLs (trQTLs).

Quantifying the expression of annotated alternative transcripts

I downloaded the Ensembl 85 gene annotations in FASTA format from the Ensembl website. I then used Salmon (Patro et al., 2016) v0.7.2 to quantify the expression levels of 178,136 transcripts from 39,037 genes. I specified the following options: '--useVBOpt --seqBias --gcBias --libType ISR'. The '--seqBias' option quantified the extent of sample specific fragment bias for each gene and adjusted the normalised transcript expression levels accordingly. Similarly, '--gcBias' option quantified the extent of sample specific GC content bias and corrected the

normalised transcript expression levels accordingly. I expected the '--gcBias' option to be important given the difference in GC content bias between automatic and manual library construction methods that I identified in Chapter 3.

Constructing alternative transcription events from known annotations

In the second approach, I modified the reviseAnnotations (https://github.com/kauralasoo/reviseAnnotations) code introduced in Chapter 2 to construct alternative transcription events from known annotated transcripts. I downloaded the Ensembl 85 transcript coordinates as well as transcript metadata using the biomaRt (Durinck et al., 2005) R package. I focussed the analysis on 71,991 protein coding and lincRNA transcripts from 16,762 genes, only including genes that had at least two annotated transcripts. I also extracted transcript tags from the Ensembl 85 GTF file downloaded from the Ensembl website. Importantly, the tags contained information if the 3′ or 5′ end of the coding sequence (CDS) was incomplete for any given transcript. In total, I found that the coding sequence was incomplete for 20,966/65,140 (32%) of the protein coding transcripts. The truncated transcripts of the IRF5 gene are illustrated on Figure 4.1A. To overcome potential bias caused by incomplete transcript annotations, I first decided to extend the truncated transcripts by using exons from transcript with the furthermost 3′ or 5′ end (depending on which end of the transcript was incomplete). The extended transcripts of the IRF5 gene are illustrated on Figure 4.1B.

Figure 4.1: Extending truncated transcripts of the IRF5 gene. (A) Protein coding transcripts of the IRF5 gene from the Ensembl 85 gene set. The transcripts with annotated incomplete 3′ ends are marked with red asterisks. **(B)** Truncated transcripts have been extended using the exons from the transcript with the furthermost 3′ end (ENST00000249375). Transcript annotations have been plotted using wiggleplotr (https://github.com/kauralasoo/wiggleplotr) R package and introns have been rescaled to constant length to facilitate visualisation.

In Chapter 2 I observed that different types of alternative transcription are often regulated independently, but this complexity is not well represented by current transcript annotation. After extending the truncated transcripts, I modified the reviseAnnotations (https://github.com/kauralasoo/reviseAnnotations) code to split the full transcripts into alternative transcription events. Briefly, I first identified the set of exons that were shared by all transcripts of the gene. Then I went through all of the individual transcripts of the gene and identified all the exons of the transcript that were either upstream, between or downstream of the shared exons. Finally, I appended the transcript-specific exons to the shared exons to construct alternative transcription events corresponding to alternative promoters, alternative middle exons and alternative transcript ends. With this approach I was able to identify seven different alternative promoters, one alternative middle exon and four alternative transcript ends from the original 11 different transcripts of the IRF5 gene (Figure 4.2). If there were no shared exons between all of the transcripts of the gene, I first split the transcripts into multiple groups of overlapping transcripts and then constructed alternative events in each group separately. The approach described here is best suited for disentangling changes in alternative promoters from changes in alternative transcript ends. Due to high complexity in transcript annotations, the alternative promoter and alternative transcript end events identified with this approach can still contain alternative middle exons (Figure 4.2).

I used the rtracklayer (Lawrence et al., 2009) package to export the alternative transcript annotations in GFF format and used to gffread tool from cufflinks v2.2.1 (Trapnell et al., 2010) to extract the alternative event sequences from the GRCh38 reference genome sequence. Finally, I quantified the expression of each alternative transcription event with Salmon using identical parameters that I used for full transcript analysis. I used separate Salmon index for the three different types of events (alternative promoters, middle exons and transcript ends) to avoid any bias caused by shared exons common to all of these events.

Figure 4.2: Alternative transcription events constructed from the 11 annotated transcripts (Figure 4.1B) of the IRF5 gene. Exons shared by all alternative events are highlighted in green and exons specific to some events are shown in blue.

Quantifying rates of intron retention

I used LeafCutter (Li et al., 2016b) to identify 38,725 clusters of intron excision events corresponding to a total of 142,030 alternatively excised introns. In each sample, I counted the number of reads supporting each intron excision event in a cluster as well as the total number of reads in a cluster.

4.2.4 Transcript ratio QTL mapping

Data normalisation

All three quantification approach described above (Ensembl 85, reviseAnnotations, and LeafCutter) allowed me to calculate the relative expression of a single event (transcript, transcription event or intron) relative to all other events in the same cluster (gene, part of a gene or intron cluster). In the case of transcripts, this can be interpreted as the proportion of the total expression of the gene that can be attributed to a single transcript. For transcripts and transcription events I used the Salmon TPM estimates to calculate the relative expression values. For intron excision events identified by LeafCutter I used the raw read counts overlapping exon junctions.

In some samples the relative expression of an event was not defined because the total expression of the group was zero. In those cases, I replaced the missing relative expression values with the mean value from all present samples. Finally, I quantile normalized the relative expression levels for each event across samples to a standard normal distribution. While conservative, this approach was efficient against two types of artefacts in intron excision ratios: (i) excess of values very close to 0 and 1 and (ii) excess of outlier excision ratios caused by very low estimated expression level for some events.

Detecting transcript ratio QTLs

I applied FastQTL to the quantile normalised transcript ratios from the three quantification approaches described above. I used the first six principal components of the phenotype matrix as covariates for the transcript ratio QTL (trQTL) mapping. I limited the cis region to +/- 100kb around the group of transcripts and obtained permutation p-values for each transcript. For each group, I took the p-value of the most significantly associated transcript and used Bonferroni correction to correct for the number of transcripts in a group. This approach was conservative as

the alternative events in a group are not independent from each other. Finally, I used Benjamin-Hochberg FDR correction on the Bonferroni-corrected p-values to identify all trQTLs at 10% FDR level.

4.2.5 Overlap analysis with the NHGRI-EBI GWAS catalogue

I downloaded the latest version of the NHGRI-EBI GWAS catalogue v1.0.1 from the EBI website on 2 March 2016 (Welter et al., 2014). I only retained studies that were conducted in European populations and where the sample size exceeded 1,000. For each trait, I performed LD pruning to only keep independent associations ($R^2 < 0.8$). After filtering, the catalogue contained 10,727 independent associations for 807 different traits. I considered an QTL to overlap a GWAS hit if the distance between the lead QTL variant and the GWAS hit was less than 1 Mb and R^2 between the variants was greater than 0.8.

4.2.6 QTL replicability between conditions

For the Storey's π_1 analysis (Nica et al., 2011), I identified eGenes at 10% FDR in one condition, took their permutation-based lead variant p-values in the other condition and used the qvalue (Dabney et al., 2010) package to estimate the proportion of non-null p-values. For the lead variant concordance analysis, I identified eGenes together with their lead variants at 1% FDR in one condition, extracted their lead variants in the other condition and counted how often R^2 between the two lead variants of the same gene was > 0.8.

4.3 Quantifying gene expression and alternative transcription

We collected a total of 336 RNA-seq samples from macrophages differentiated from 84 iPSC lines in four experimental conditions. After quantifying gene expression levels (See Methods), I used Principal Component Analysis (PCA) to assess the quality of the data. PCA revealed four distinct clusters with the first principal component (PC1) explaining 44% of the variance and roughly corresponding to *Salmonella* infection status and PC2 (explaining 15% of the variance) roughly corresponding to IFNɣ stimulation (Figure 4.3). PC5 that was most strongly correlated with the RNA-seq library construction method (manual or automatic) explained only 1.6% of the variance in the data.

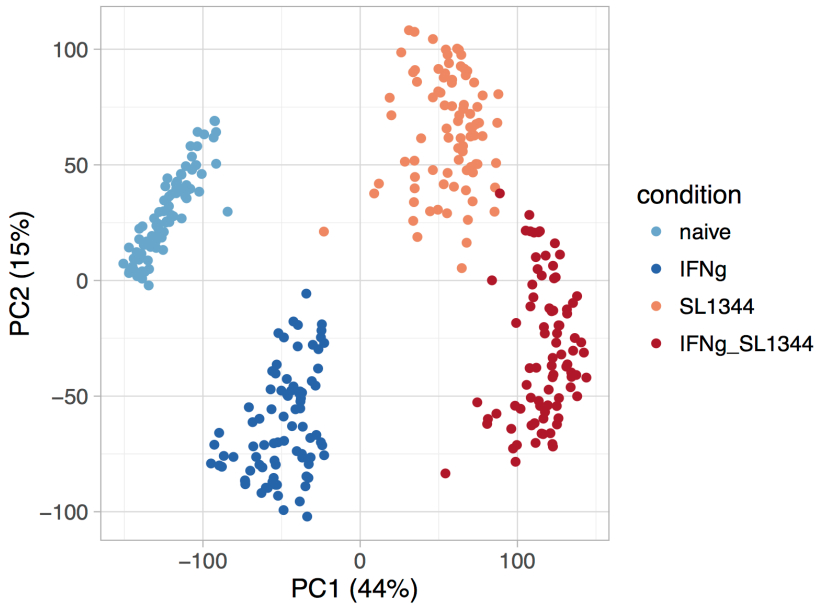

Figure 4.3. Principal component analysis of normalised and standardised gene expression data.

In addition to gene level analysis, I also quantified the relative expression of individual transcripts from the Ensembl 85 reference annotations and used the ratio between the transcript expression and total gene expression as the phenotype of interest. However, as highlighted in Chapter 2, reference annotations are still incomplete and often miss many transcripts expressed by the cells. To overcome this limitations, I used a modified version of the reviseAnnotations tool that I developed in Chapter 2 to split reference transcripts into individual alternative transcription events and subsequently quantified the relative expression of each event. I also used LeafCutter (Li et al., 2016b) to identify and quantify the relative excision ratios of 50,538 alternative introns. These three complementary quantification approaches are referred to as Ensembl 85, reviseAnnotations, and LeafCutter in the following text. More details on each of these approaches is given in the Methods section.

In the LeafCutter data, the first two PCs only explained ~9% of the variance, indicating that there was less structure in the intron excision measurements (Figure 4.4A) compare to the gene expression levels. Moreover, while PC1 (explaining 5% of the variance) still corresponded to *Salmonella* infection, the second PC was now strongly correlated with the method of RNA library preparation (manual vs automatic) (Figure 4.4A). Finally, PC3 (2% variance explained) corresponded to IFNɣ stimulation (Figure 4.4B). In Chapter 3 I showed that there was a

difference in GC-content bias between manual and automatic RNA-seq library construction protocols. This suggests that intron excision ratios that are based on a small number of reads from a short region are more susceptible to GC-content bias than gene expression measurements that are aggregated over a longer region.

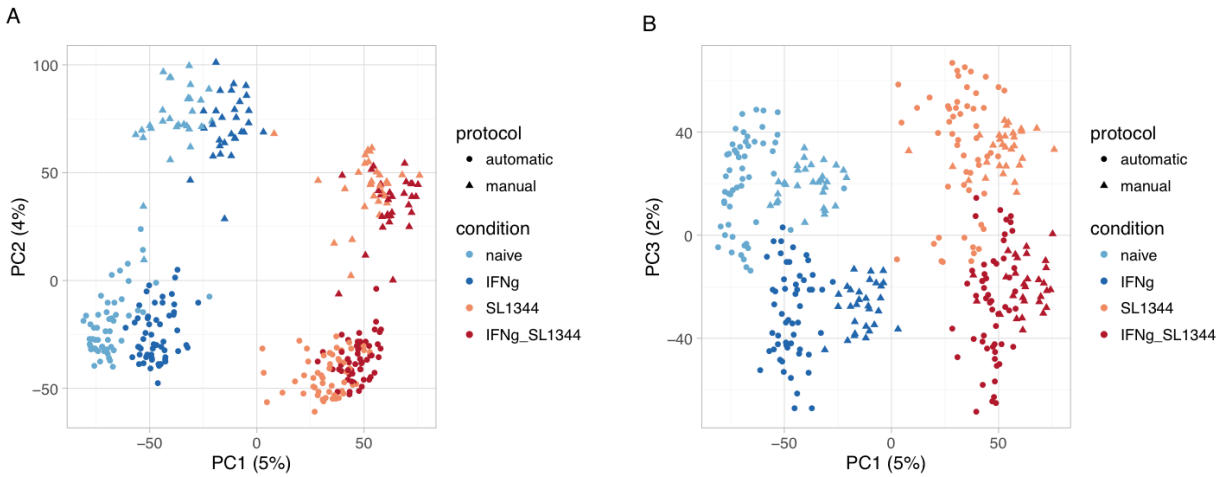

Figure 4.4: Principal component analysis of normalised intron excision ratios. (A) PC1 plotted against PC2. **(B)** PC1 plotted against PC3. **Protocol** - type of RNA-seq library construction protocol used, either manual or automatic.

4.3.1 Differential expression analysis reveals expected pathways

First, I wanted to verify that our iPSC-derived macrophages are a suitable model to study genetics of gene expression in immune response. Fortunately, macrophage response to IFNɣ and bacterial stimuli (such as LPS) have been extensively studied and most of the pathways involved in the response have been identified. I therefore sought to verify that the expected pathways are also activated in iPSC-derived macrophages after corresponding stimuli.

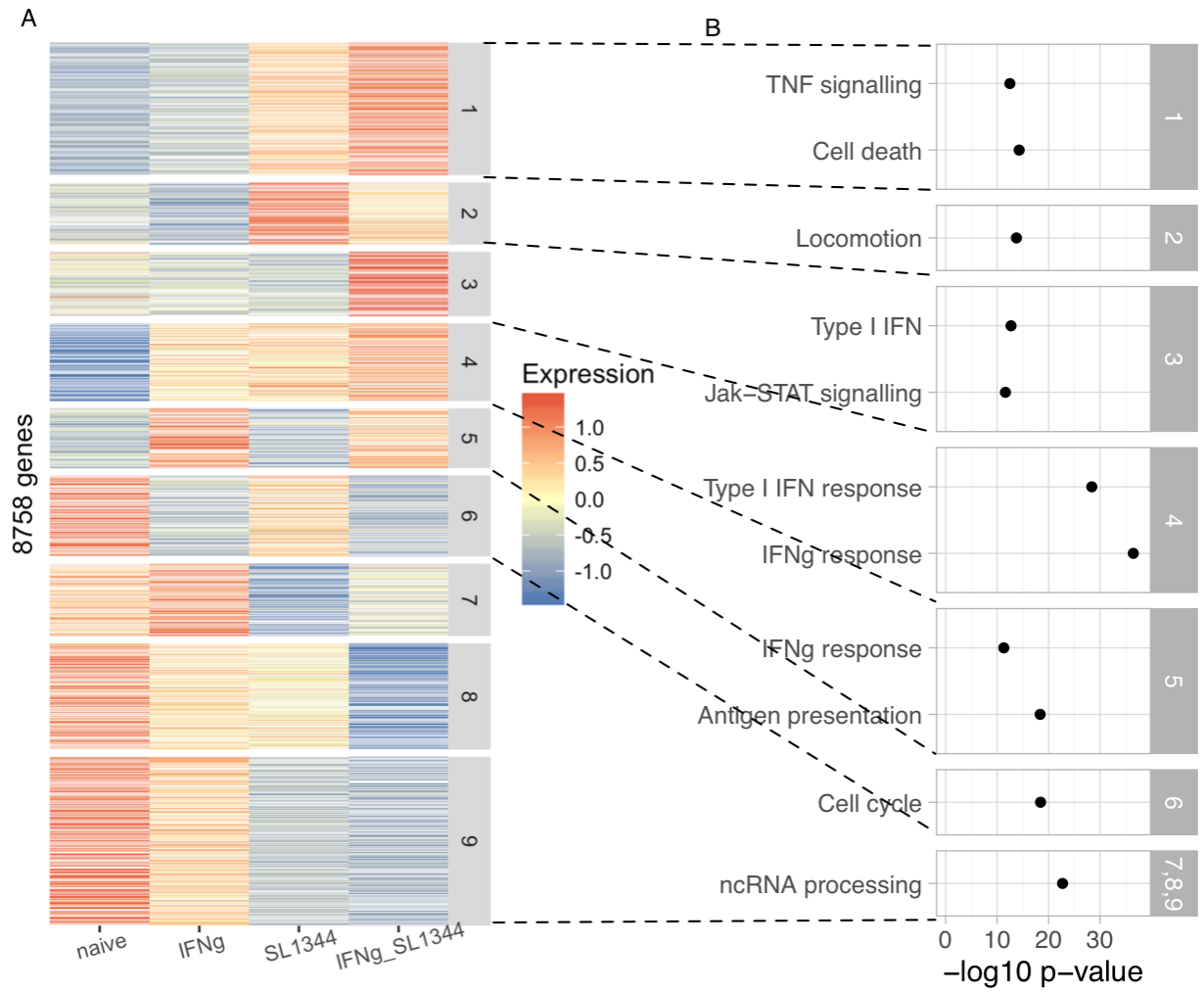

Figure 4.5: Differential gene expression between the four experimental conditions. (A) Heatmap of 8758 differentially expressed genes clustered into nine distinct patterns of expression. **(B)** A selection of Gene Ontology (GO) terms specifically enriched in each cluster. Only enrichments with $p < 1\times10^{-8}$ are shown in the figure. 'IFNɣ response' was the only GO term with enrichment p-value $< 1\times10^{-8}$ in more than one cluster.

I identified 8758 genes that were > 2-fold differentially expressed across all four conditions and clustered them into nine distinct expression patterns (Figure 4.5A). I then used g:Profiler (Reimand et al., 2016) to perform pathway and Gene Ontology enrichment analysis on these clusters. Cluster 1 (genes strongly upregulated by *Salmonella* or IFNɣ + *Salmonella*) was enriched for TNF and NF-κB signalling pathways (IL1B, TRAF1) as well as pathways involved in cell death and apoptosis (Figure 4.5B). This agrees with the observation that we recovered less total RNA from *Salmonella* and especially IFNɣ + *Salmonella* conditions (Figure 4.6), which

would also result from greater cell death following *Salmonella* infection. Cluster 2 (upregulated by Salmonella) was enriched for genes involved in locomotion. Cluster 3 consisted of genes that responded to Salmonella infection only after the cells had been pre-treated with IFNɣ. This cluster was enriched for type I interferon genes (IFNA1/8, IFNL2/3, IFNW1) and JAK-STAT signalling, but also contained other important inflammatory genes such as NOD2 and IL12A. Moreover, the synergistic activation of IL12A in response to IFNɣ and LPS is well established in monocyte-derived macrophages (Qiao et al., 2013). Cluster 4 contained genes that were upregulated similarly by IFNɣ and Salmonella and it was strongly enriched for type I interferon response and IRF1 target genes (CXCL8, IRF1, ATF3, STAT2, IDO1/2). This is consistent with the production of IFNβ and activation of IFNβ signalling downstream of TLR4 activation (Ivashkiv and Donlin, 2014). Genes in cluster 5 were only upregulated by IFNɣ and they were strongly enriched for antigen processing and presentation and MHC class II protein complex (CIITA). Again, the role of IFNɣ in activating antigen presentation genes is well established (Schroder et al., 2004).

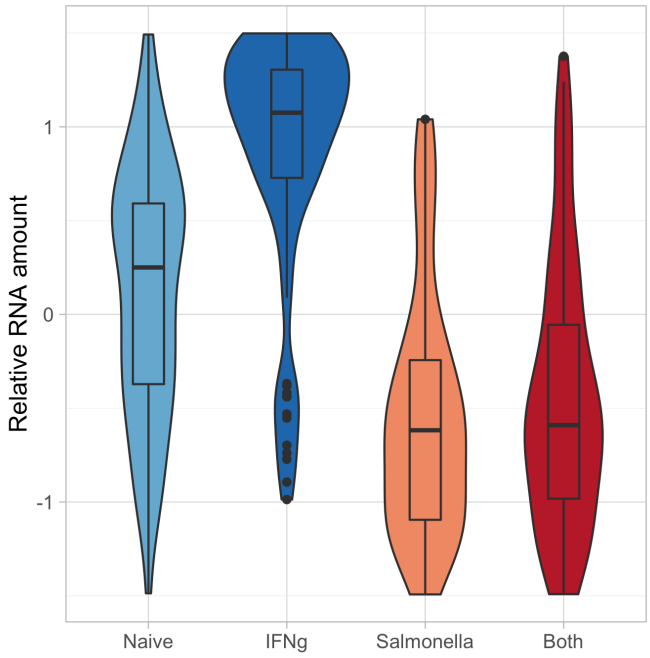

Figure 4.6: Relative amount of RNA obtained from each condition across 84 macrophage lines. I quantified the total amount of RNA obtained from each sample. For all four samples from a single line (corresponding to four conditions) I then subtracted the mean RNA amount across conditions and divided by standard deviation to obtain relative RNA amount.

Genes downregulated in the stimulated conditions also clustered into four distinct groups (Figure 4.5). Here, cluster 6 (downregulated by IFNɣ) were strongly enriched for cell cycle genes. This is consistent with multiple reports that stimulation with IFNɣ induces cell cycle arrest in macrophages (Schroder et al., 2004; Xaus et al., 1999). Finally, clusters 7,8 and 9 (all downregulated by *Salmonella*) was strongly enriched for ncRNA processing, ribosome biogenesis and tRNA processing, perhaps representing repression of translation as a general stress response.

4.4 Genetics of gene expression

4.4.1 Gene expression QTL mapping

Table 4.1: Number of eQTLs detected in +/-500kb window around each gene using either linear model (FastQTL) or allele-specific model (RASQUAL).

condition	FastQTL	RASQUAL	% difference
Naive	1932	2590	34
IFNɣ	1985	2478	25
Salmonella	1518	1882	24
Both	1449	1869	29

I used two alternative approaches to map eQTLs in each of the four conditions. First, I used standard linear model implemented in the FastQTL (Ongen et al., 2016) software. Secondly, I also used a novel RASQUAL (Kumasaka et al., 2016) method that combines both allele-specific and between-individual signal to increase the power of detecting eQTLs and also improves fine mapping causal variants. I decided to use both models for two reasons: (1) I wanted to take advantage of the allele-specific information to increase eQTL detection power (2) gene-level permutation p-values and summary statistics from the linear model can be directly used in eQTL replication and colocalisation analyses whereas this is not as straightforward for the RASQUAL output. I found that at the same 10% FDR level RASQUAL was able to detect on average 28% more genes with significant eQTLs (Table 4.1). The increase in power was also evident on the quantile-quantile (Q-Q) plot (Figure 4.7).

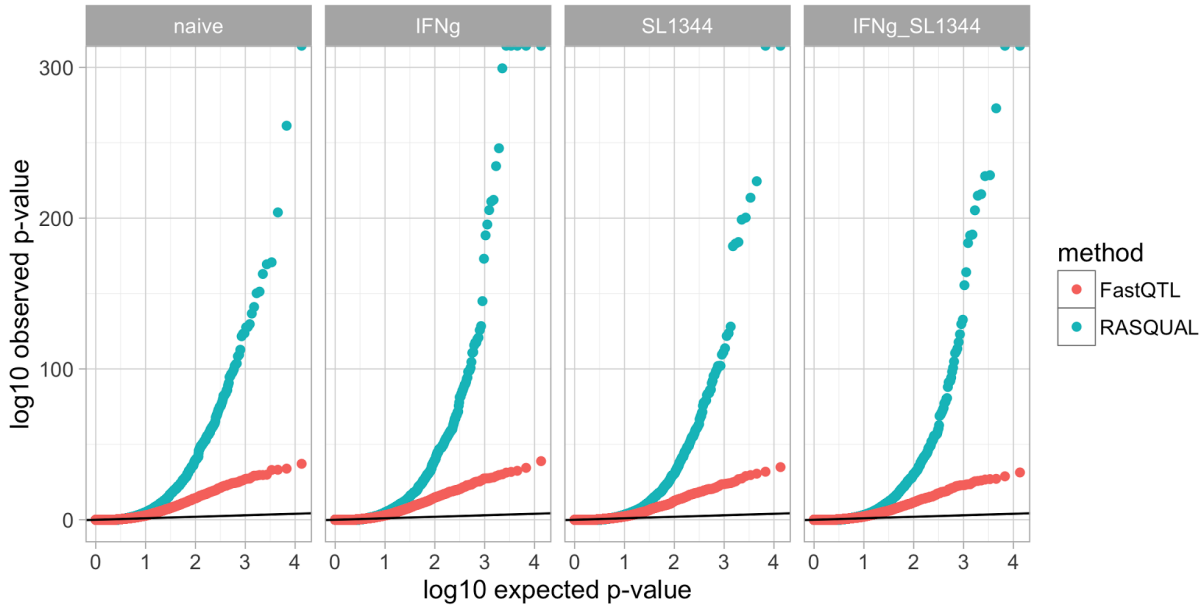

Figure 4.7: Quantile-quantile plots for the p-values of eQTLs detected either with RASQUAL or FastQTL. Solid lines represent the expected distribution of p-values under the null model.

4.4.2 Transcript ratio QTL mapping

I also used FastQTL in combination with the three quantification methods described above to map transcript ratio QTLs (trQTLs) in a +/-100 kb cis-window around the feature in all four conditions. I use smaller cis-window for trQTLs compared to eQTLs (+/-500kb), because trQTLs are known to be strongly enriched near the exon boundaries (Li et al., 2016c). Using either raw reference transcripts (Salmon + Ensembl 85) or transcription events constructed from them (Salmon + reviseAnnotatons), I detected between 1,500 and 2,500 trQTLs per condition (Table 4.2). Ensembl 85 results contained slightly more unique genes while reviseAnnotatons was able to identify multiple independent trQTLs for a subset of genes as illustrated by the IRF5 example below. Finally, LeafCutter detected approximately 45% less trQTLs than the annotation-based methods.

Table 4.2: Number of transcript ratio QTLs detected by different quantification methods at 10% FDR. Only variants within +/- 100kb of the transcript were included in the analysis.

Condition	LeafCutter	Salmon + Ensembl 85	Salmon + reviseAnnotations

Naive	1953	2201	2429
IFNɣ	1756	2095	2314
Salmonella	1496	1743	1858
Both	1304	1481	1547

4.4.3 Concordance of QTLs detected by different methods

Comparing different QTL mapping approaches just by the numbers of QTLs found is not very informative, because it completely ignores the identity of the QTLs detected. Looking at simple overlaps between lead QTL variants can also be misleading, because the lead SNPs can be randomly different between the methods and still tag the same causal variant in high LD. To overcome this limitation, I decided to test if the lead variants for the same sets of genes (or transcripts) were concordant with each other for two different QTL mapping approaches. Specifically, I took all lead variants at 1% FDR from one method and compared them to the lead variants of the same genes (or transcripts) from a different method (even if below the 1% threshold). I then calculated the fraction of lead variant pairs that were in high LD ($R^2 > 0.8$) with each other. Note that this approach is likely to underestimate the true extent of QTL sharing between methods in cases where there are multiple independent QTLs per gene.

First, I found that 60% of the eQTL lead variants detected by FastQTL were also found by RASQUAL whereas only 40% of the RASQUAL QTLs were detected by FastQTL (Figure 4.8). This is consistent with the smaller number of eQTLs detected by the linear model (Table 4.1). I found similar level of lead variant sharing (~60%) between trQTLs detected using reviseAnnotations and Ensembl 85 annotations whereas sharing between reviseAnnotations and LeafCutter trQTLs was considerably lower (30-40%). This suggests that LeafCutter might be more efficient in capturing unannotated alternative exons that are not present in reference annotations. Finally, there was only moderate (20-30%) lead variant sharing between FastQTL eQTLs and reviseAnnotations trQTLs and this decreased to 10-12% when comparing to LeafCutter. This suggests that eQTLs and trQTLs are to a large extent independent from each other.

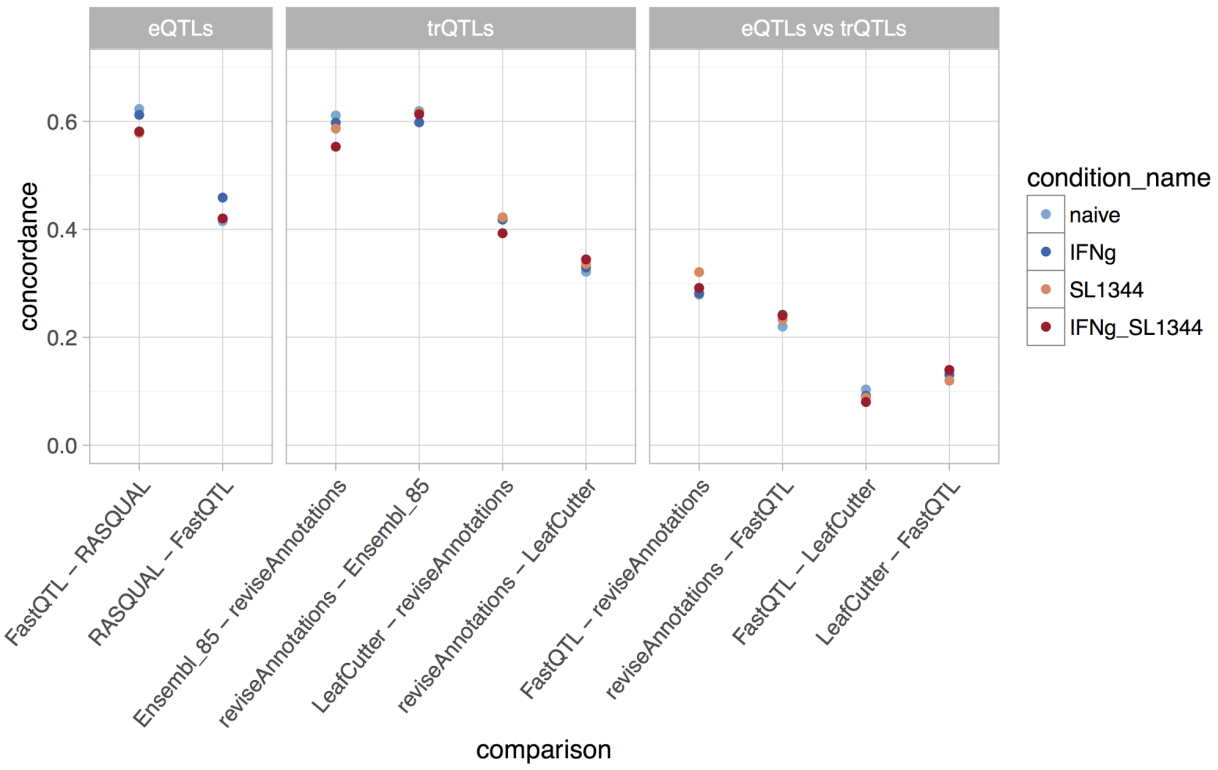

Figure 4.8: Concordance of lead QTL variants detected by different methods. In the gene expression (eQTL) comparison (left panel) I used FastQTL and RASQUAL lead variants from +/-500kb cis-window. For the eQTL and trQTL comparison (rightmost panel) I reran FastQTL eQTL mapping in a 100kb around the gene to ensure that the lead variants were comparable to the trQTLs.

4.4.4 Condition specificity of eQTLs and trQTLs

Next, I used two different approaches to estimate the proportion of condition specific eQTLs and caQTLs. First, I used Storey's π_1 statistic to estimate the sharing of QTLs between conditions. Briefly, I identified eGenes at 10% FDR in each condition and then looked their minimal p-values in the other three conditions and estimated the fraction of those that were true positives. I found that the fraction of shared eGenes varied between 0.75 and 0.90 with the lowest sharing observed between naive and IFNɣ + *Salmonella* conditions (Figure 4.9). This is somewhat higher than the 53-80% sharing observed between different tissues (Nica et al., 2011; The GTEx Consortium, 2015), but much lower than the sharing of eQTLs in the same tissue across twin pairs (Nica et al., 2011).

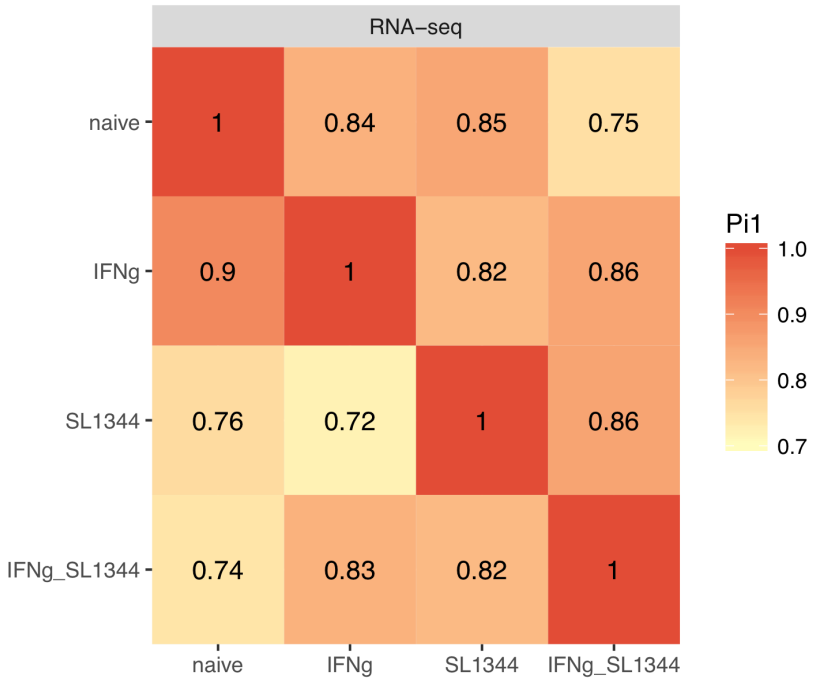

Figure 4.9: Replicability of eGenes between conditions. The heatmap shows the pairwise Storey's π_1 statistic for eQTLs detected between conditions.

However, this type of replicability analysis has several limitations. First, it considers only the p-value of one lead variant per gene and ignores patterns of linkage disequilibrium. Consequently, if the gene has two unlinked highly condition-specific eQTLs then this would be considered a successful replication even though both of the variants have condition-specific effects. Secondly, calculating the π_1 statistic requires that the null p-values are uniformly distributed. This assumption is not satisfied by the Bonferroni corrected p-values from RASQUAL or trQTL analyses where most p-values are strongly skewed towards 1. As a result, π_1 statistic cannot be used on those datasets.

To overcome these limitations, I decided to use the same lead variant concordance analysis described above to compare QTLs from different conditions. I found that ~55% of the eQTL lead variants and ~65 trQTL lead variants were shared between conditions, suggesting that trQTLs are slightly less likely to be condition specific than eQTLs (Figure 4.10).

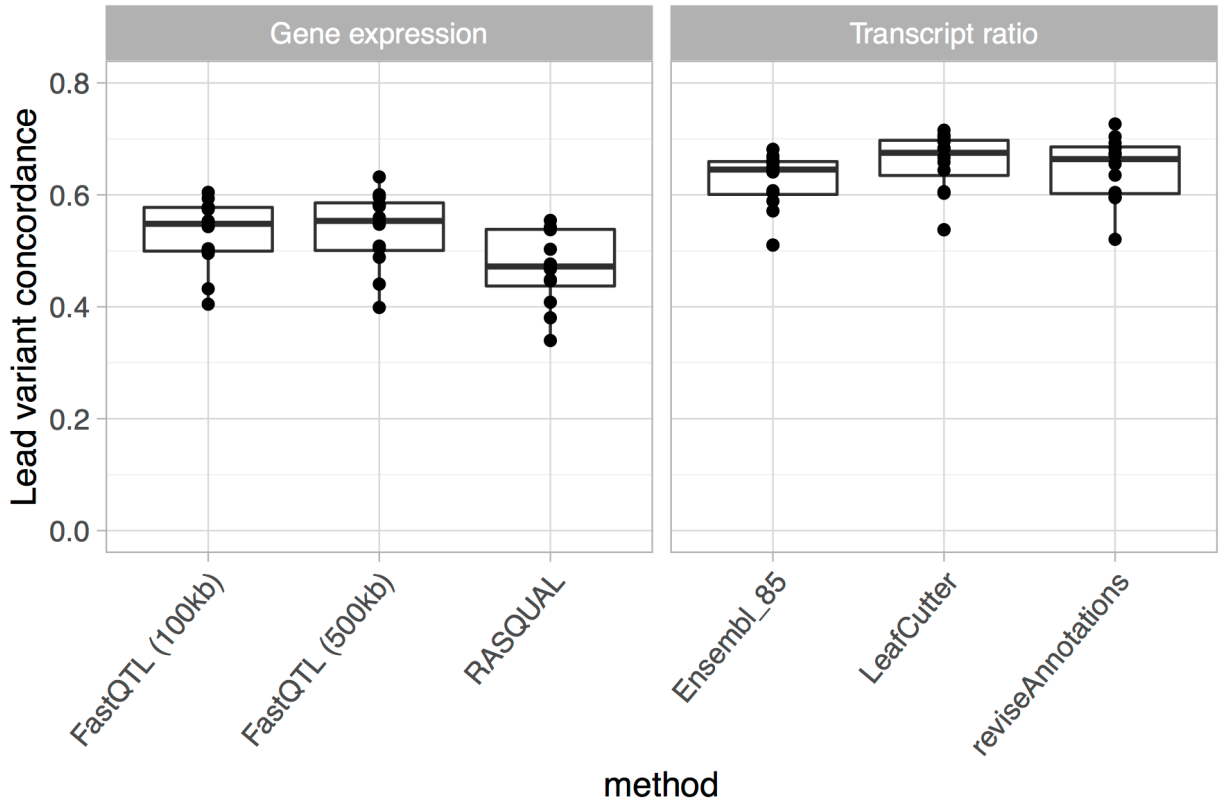

Figure 4.10: Concordance of QTL lead variants between pairs of conditions detected by different QTL mapping methods. Each dot represents one pairwise comparison between conditions (such as IFNɣ vs naive). I mapped eQTLs with FastQTL in both +/- 500kb and +/- 100kb cis-windows to match the 100 kb window used for transcript ratio QTLs.

Identifying condition-specific eQTLs

Although the π_1 and lead variant concordance analyses are useful to estimate the global level of eQTL replicability between conditions, they do not identify specific variants and analyse their effect sizes. To identify individual condition-specific eQTL and their target genes, I compiled all independent ($R^2 < 0.8$) lead SNP-gene pairs from RASQUAL across conditions and used standard ANOVA model to test for interactions between genotype and condition (See methods). A Q-Q plot revealed that the p-values of the interaction test were well calibrated (Figure 4.11A). I found that 1,172/5,782 (20%) lead eQTL variants corresponding to 996/3,905 (26%) eGenes had significantly different effect sizes between conditions.

Although statistically significant, sometimes the effect size differences were relatively small. As a measure of the effect size of an eQTL I used the $\log_2$ fold change ($\log_2$FC) between reference

and alternative alleles estimated by RASQUAL. For an eQTL to be considered condition specific I required the difference in log_2FC between naive and any one of the stimulated conditions to be greater than 0.32 (~1.25 fold). In our dataset, 741/996 condition-specific eQTLs passed this threshold out of which 496 appeared after stimulation (i.e. log_2FC was less than <= 0.59 (~1.5-fold) in the naive condition, Figure 4.11C) and 245 disappeared after stimulation (log_2FC was greater than 0.59 (~1.5-fold) in the naive condition, Figure 4.11B). Finally, I used k-means clustering of the relative effect sizes to assign eQTLs into different activity patterns (Figure 4.11B-C). I observed that slightly more eQTLs appeared after *Salmonella* infection (clusters 2,3 and 4, n = 260) than after IFNɣ stimulation (clusters 5,6, n = 156). Furthermore, 83 eQTLs only appeared after both of the stimuli were present (cluster 1), highlighting the importance of studying combinations of stimuli.

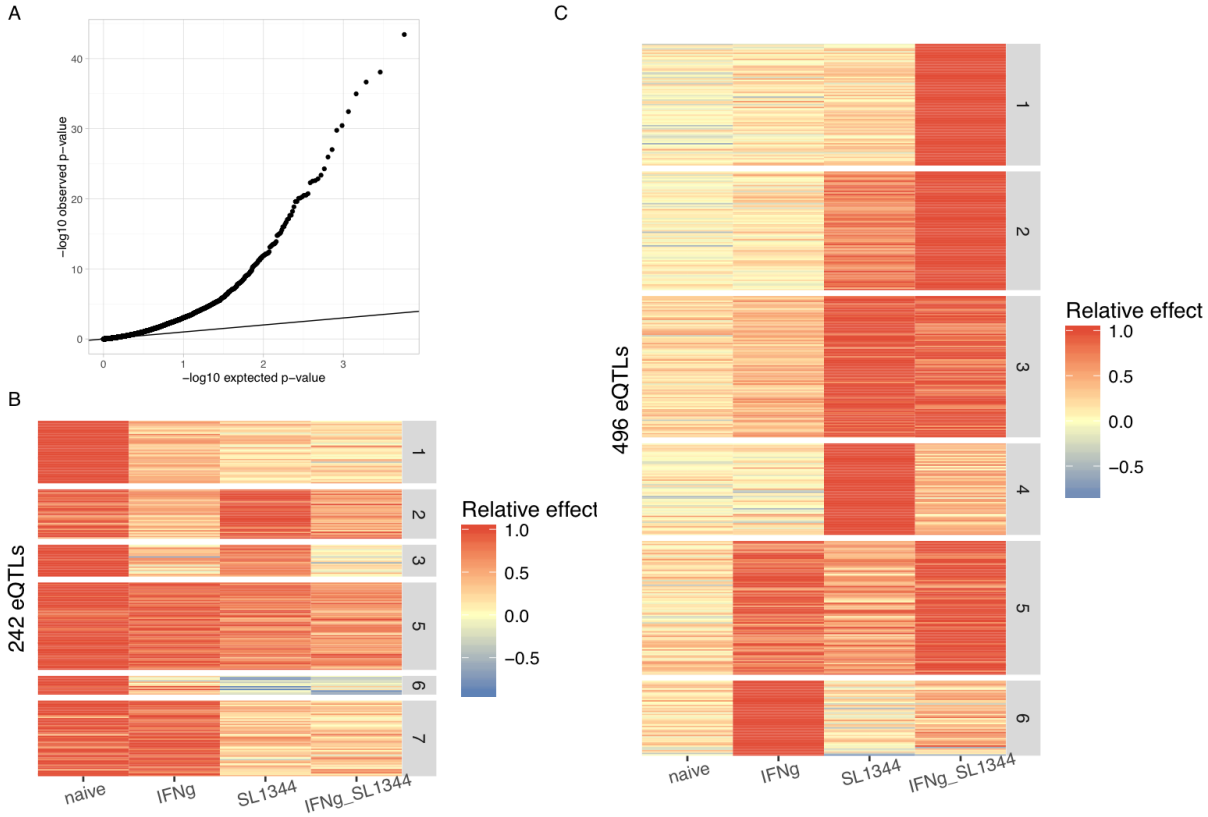

Figure 4.11: Condition-specific eQTLs clustered by their effect size. **(A)** Quantile-quantile plot of the expected and observed p-values for the interaction test **(B)** Effect size heatmap of the seven clusters of eQTLs that disappeared after stimulation. **(C)** Effect size heatmap of the six clusters of eQTLs that appeared after stimulation. For each gene, the relative effect size was calculated by dividing the eQTL effect size in each condition by the maximal absolute effect size

across conditions. This ensured that the eQTLs with different absolute effect sizes were visually comparable on the heatmap.

4.5 Case study: genetics of IRF5 transcription

To illustrate the power of using complementary approaches for gene expression and transcript ratio QTL mapping, I focussed on the IRF5 gene. Using total read counts and the standard linear model (FastQTL), I was not able to detect any significant eQTLs for this gene. Transcript level analysis with Ensembl 85 annotations, however, identified a very strong trQTL (rs10954213, $p < 2.9\times10^{-32}$, MAF = 0.46) that on a closer inspection turned out to regulate 3′ UTR usage (Figure 4.12). The association between the rs10954213 variant and 3′ UTR usage of the IRF5 gene has been previously reported by multiple studies (Cunninghame Graham et al., 2007; Yoon et al., 2012; Zhernakova et al., 2013) and the lead variant is likely to be the causal one because it changes the canonical polyadenylation signal from AATAAA to AATGAA.

Using alternative transcription events from reviseAnnotations not only detected the 3′ UTR QTL (Figure 4.12), but also identified an additional trQTL regulating alternative promoter usage (rs3778754, $p < 4.7\times10^{-16}$, MAF = 0.33) independently of the 3′ UTR usage (MAF = 0.43) (Figure 4.13). A key advantage of reviseAnnotations was that it was able to correctly identify that one of the trQTLs regulated 3′ UTR usage while the other one regulated alternative promoters, thus greatly improving the interpretability of the detected trQTLs. Although the promoter QTL was also detected by LeafCutter ($p < 3\times10^{-17}$) the 3′ UTR QTL was not, because alternative polyadenylation will not result in detectable changes in exon-exon junction reads. The lead promoter QTL variant (rs3778754) is also in high LD ($R^2 = 0.84$) with a GWAS lead SNP rs4728142 for Systemic lupus erythematosus and Ulcerative colitis. Moreover, a recent fine mapping analysis of the GWAS locus identified rs3757387 as the most likely causal variant which is in even higher LD with the promoter QTL ($R^2 = 0.93$) (Kottyan et al., 2015).

Finally, RASQUAL detected a third trQTL for the same gene (rs199508964, $p < 4.9\times10^{-33}$, MAF =0.48) that seems to influence the excision of an alternative intron in the fifth coding exon of the gene (Figure 4.14). Although the lead variant directly overlaps the splice site of the retained intron, it is a 33 bp deletion that is also in moderate LD with the 3′ UTR QTL variant ($R^2 = 0.58$). Therefore, some care is in order when interpreting this variant. This trQTL was missed by LeafCutter, because it does not detect intron retention events.

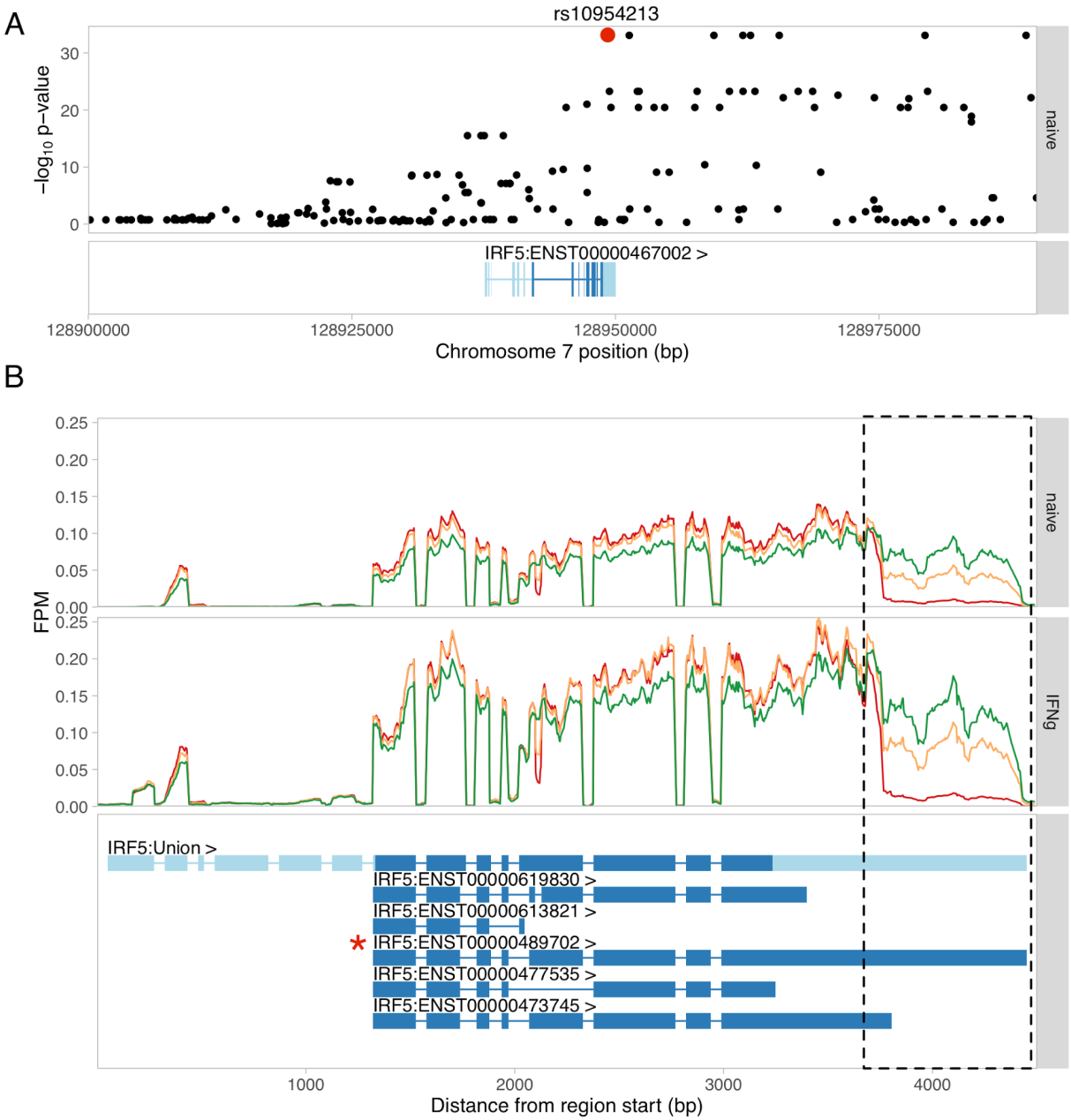

Figure 4.12: Example of a trQTL for the IRF5 gene that influences the proximal polyadenylation site usage. (A) Manhattan plot of the associated variants around the IRF5 gene in the naive condition. The lead variant rs10954213 disrupts the proximal polyadenylation site motif. **(B)** RNA-seq read coverage stratified by the lead variant genotype. The panel below the coverage plot shows the union of IRF5 exons (top row) together with transcription events constructed by reviseAnnotations (other rows). The alternative 3′ UTR is highlighted by the dashed box.

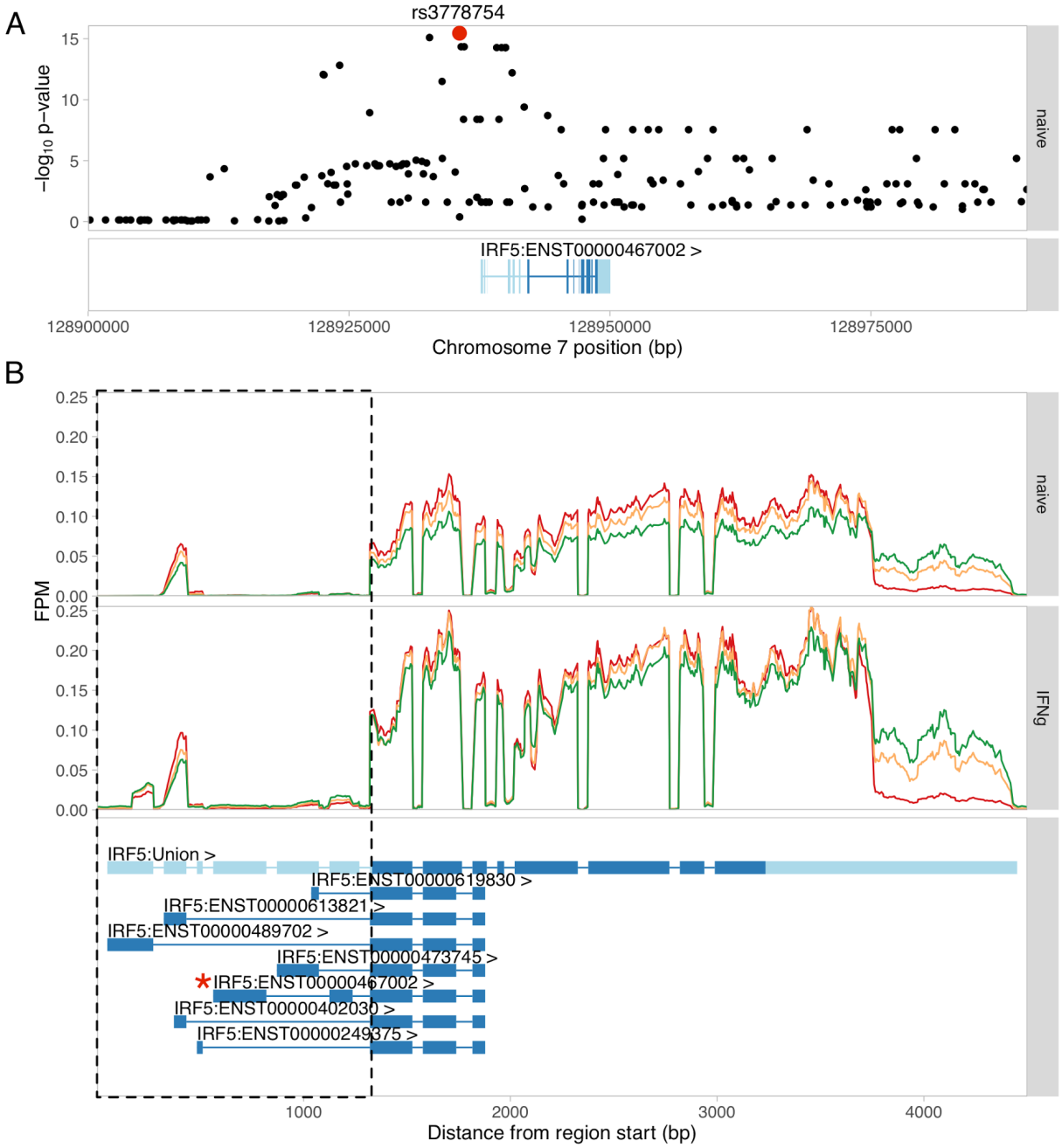

Figure 4.13: Alternative promoter QTL for the IRF5 gene. (A) Manhattan plot of the associated variants upstream of the IRF5 promoter in the naive condition. **(B)** RNA-seq read coverage across the IRF5 gene stratified by the genotype of the lead promoter QTL variant (rs3778754). The panel below the coverage plot shows the union of IRF5 exons (top row) followed by alternative promoter annotations constructed by reviseAnnotations.

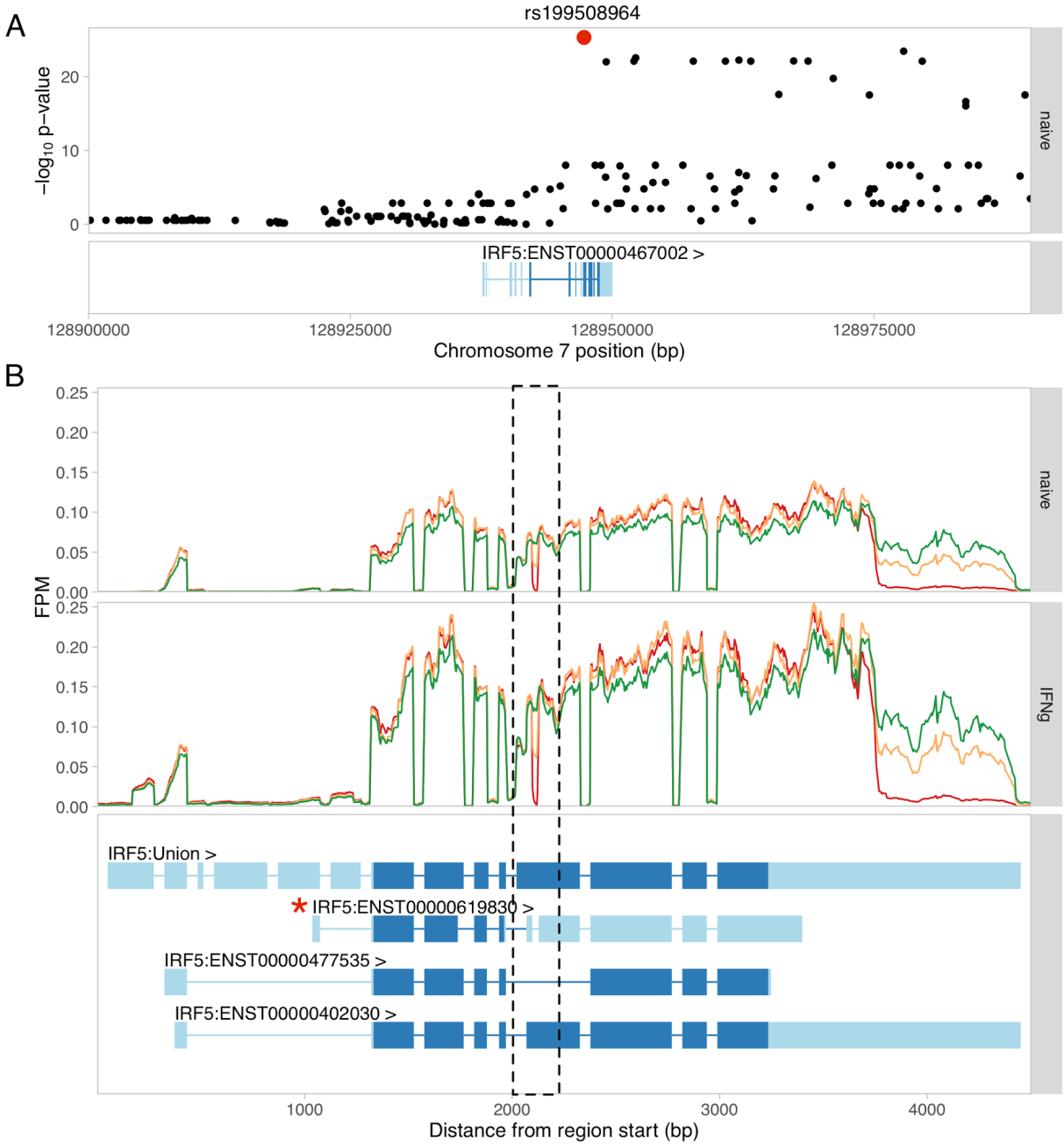

Figure 4.14: Intron excision QTL in the IRF5 gene. (A) Manhattan plot of eQTL p-values from RASQUAL in the naive condition. **(B)** Read coverage across the IRF5 gene stratified by the genotype of the lead QTL variant (rs199508964). The alternatively excised intron is highlighted by the dashed box.

4.6 Overlap with GWAS hits

An important motivation for studying the genetics of gene expression is to identify molecular QTLs that enable GWAS hits to be linked to their target genes and thereby provide a mechanistic hypothesis that could potentially explain the GWAS association. I have performed a naive overlap analysis ($R^2 > 0.8$) between all independent GWAS associations from the NHGRI-EBI GWAS catalogue and all eQTLs and trQTLs identified from the macrophage RNA-seq data. As a result, the probability that any individual overlap represents a shared causal mechanism is likely to be low. However, looking at the overlaps in aggregate can inform us about the traits and diseases for which iPSC-derived macrophages might be a relevant cell type.

First, I assessed how many potential GWAS overlaps are missed when looking at eQTLs and trQTLs only in the naive condition. I found using eQTLs and trQTLs from all four conditions as opposed to just from the unstimulated cells identified at least twice as many overlapping GWAS associations (Figure 4.15). Furthermore, the GWAS overlaps with eQTLs and trQTLs were largely independent from each other as illustrated by the fact that joint analysis with all QTLs identify 40% more overlaps. It is important to stress that most of these overlaps are likely to be spurious and careful colocalisation analyses are needed to dissect individual loci.

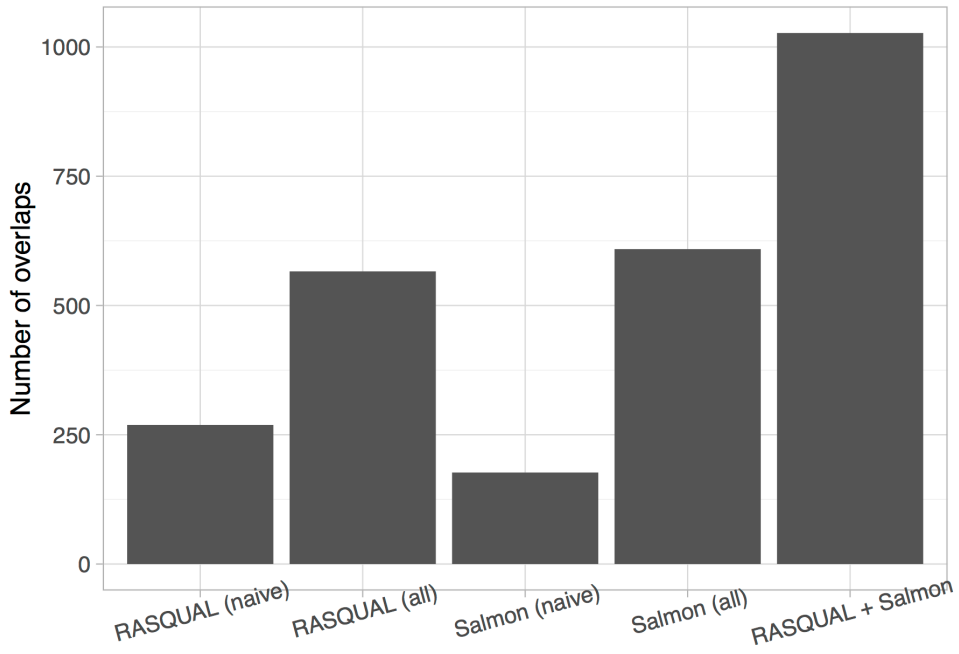

Figure 4.15: Number of RASQUAL eQTLs and Salmon trQTLs overlapping GWAS hits. 'Naive' represents QTLs from the unstimulated condition only while 'all' stands for all independent (R^2 < 0.8) QTLs across conditions. Lead QTL and GWAS variants were considered to be overlapping if the distance between the variants was less than 1 Mb and R^2 between the variants was > 0.8.

Secondly, I counted the number of overlaps for each trait in the GWAS catalogue and ranked the traits by fraction of associations that overlapped a macrophage QTL. I found that top 20 traits with the largest fraction of associations overlapping macrophage QTLs contained Alzheimer's disease, multiple autoimmune disorders and multiple lipid traits, suggesting that iPSC-derived macrophages might be a relevant cell type for studying the genetic mechanisms underlying these traits. As a negative control, height ranked 56th with only 10% of its associations overlapping macrophage eQTLs and trQTLs and most cancers had even smaller overlap.

Table 4.3: List of top 20 traits with largest overlap between GWAS hits and macrophage eQTLs/trQTLs. Only traits with more than 15 independent associations were included. Autoimmune traits are highlighted in red, lipid traits in green and blood traits in blue.

	Trait	Overlap size	Trait size	Fraction
1	Ankylosing spondylitis	5	17	0.29
2	Primary biliary cirrhosis	8	28	0.29
3	Testicular germ cell tumor	5	21	0.24
4	Alzheimer's disease (late onset)	8	36	0.22
5	Metabolic traits	8	36	0.22
6	Fibrinogen	5	25	0.2
7	White blood cell count	4	20	0.2
8	Inflammatory bowel disease	21	111	0.19
9	Menopause (age at onset)	6	32	0.19

10	Idiopathic membranous nephropathy	3	16	0.19
11	Platelet count	10	58	0.17
12	HDL cholesterol	15	90	0.17
13	C-reactive protein levels	3	18	0.17
14	Triglycerides	10	61	0.16
15	Liver enzyme levels (gamma-glutamyl transferase)	4	25	0.16
16	Homocysteine levels	3	19	0.16
17	Crohn's disease	17	109	0.16
18	LDL cholesterol	11	71	0.15
19	Multiple sclerosis	19	123	0.15
20	Cholesterol, total	12	78	0.15

4.7 Discussion

In this chapter I have shown that iPSC-derived macrophages are able to well recapitulate known aspects of macrophage biology in immune response. In particular, I have shown that their gene expression response to *Salmonella* infection and IFNɣ stimulation matches what is known from the literature. I have also shown iPSC-derived macrophages are a robust cell culture based system that can be used to map condition-specific genetic effects on both gene and transcript expression level.

We detected around 2,000 gene expression and transcript ratio QTLs in each experimental condition and found that ~25% of the QTLs were condition specific. This also included 495 eQTLs that were completely hidden in the unstimulated cells and only appeared after stimulation. Many potential overlaps with disease hits were also only detected in the condition-specific samples. Together these results highlight that the effect of some genetic variants on

gene expression manifests most clearly in specific environmental conditions. Hence, to construct a comprehensive catalogue of regulatory variation we need to profile gene expression in a large number of conditions. IPSC-derived cells provided an excellent opportunity for this, because they can be reliably obtained in large numbers from the same set of individuals.

The three independent transcript ratio QTLs regulating alternative promoter usage, alternative intron retention and alternative 3′ UTR usage of the IRF5 gene highlight that different parts of the same transcript can be regulated by independent genetic mechanisms. This can be a challenge for transcript ratio QTL mapping, because all possible combinations of promoters, exons and 3′ ends are usually not represented by the set of annotated transcripts. Furthermore, up to 30% of the human protein coding transcripts annotations are incomplete and miss either their 3′ or 5′ ends. As a result, methods that focus on individual alternative transcription events such as MISO (Katz et al., 2010), DEXSeq (Anders et al., 2012) and LeafCutter (Li et al., 2016b) have proven to be very successful. The first contribution of my reviseAnnotations approach is that it extends truncated transcripts with known exons of the gene. It then splits known transcripts into alternative 5′ ends, middle sections and 3′ ends. It is therefore a hybrid approach between full transcript and exon level analyses, that is still able to take advantage of the read coverage patterns over multiple exons (such as alternative promoters skipping multiple first exons) and at the same time identify independent effects on different parts of the gene.
I found that eQTLs and LeafCutter trQTLs were largely independent from each other, thus confirming an earlier observation in LCLs (Li et al., 2016c). I also mapped trQTLs on transcription event level (Salmon + reviseAnnotations) and found that these QTLs were also largely independent from eQTLs, although to a lesser degree. Although LeafCutter and Salmon detected similar numbers of trQTLs, I found that only 30-40% of the lead variants were shared. One reason for this discrepancy is that the two approaches capture different transcription events. LeafCutter is able to detect QTLs for alternative exons that have not been annotated. Salmon, on the other hand, is able to detect QTLs for annotated alternative 3’ and 5’ ends that do not involve splicing (i.e. alternative polyadenylation) and are therefore missed by LeafCutter. Salmon might also be more powerful for lowly expressed genes and weaker effects, because it is not limited to exon-exon junction reads and is able to correct for fragment length and GC-content bias during quantification.

5 Genetics of chromatin accessibility in macrophage immune response

Collaboration note

The work in this chapter was performed in collaboration with Julia Rodrigues who was a research assistant in Daniel Gaffney's lab at the time. I designed the experiments, performed Salmonella infection and IFNɣ stimulation assays, took care of sample logistics and performed all of the data analysis. Julia prepared the cells for experiments and performed the experimental side of the ATAC-seq protocol. We shared macrophage differentiation tissue culture responsibilities.

5.1 Introduction

A major limitation of gene expression quantitative trait loci (eQTL) mapping studies is that due to linkage disequilibrium we are usually unable to identify causal variant(s). Although genetic variation can influence a gene expression through a variety of transcriptional and post-transcriptional mechanisms, a large fraction of local eQTLs act by modulating the activity of regulatory elements (promoters and enhancers) and, subsequently, the rate of transcription of the gene. For example, an early study that measured chromatin accessibility and gene expression in the same population of lymphoblastoid cell lines (LCLs) estimated that as many as 55% of eQTLs were also chromatin accessibility QTLs (caQTLs) (Degner et al., 2012). Furthermore, caQTLs are strongly enriched in a relatively small accessible region, thus narrowing down the set of likely causal variants. However, no study thus far has mapped both eQTLs and caQTLs in multiple conditions to study how genetic effects on chromatin level propagate down to gene expression level in the context of stimulation.

Since the original caQTL experiment (Degner et al., 2012), other studies have followed looking at the genetics of histone modifications and transcription factor binding (Ding et al., 2014; Grubert et al., 2015; Waszak et al., 2015). However, due to the large cell numbers required by chromatin assays, all of these studies have been conducted in LCLs. Therefore, although the cell type and condition specificity of eQTLs is well established (Fairfax et al., 2012, 2014), how

these effects manifest on the chromatin level and how they propagate down to gene expression is mostly unknown. The development of ATAC-seq (assay for transpose accessible chromatin) has made it possible to measure chromatin accessibility in much smaller number of cells, thus greatly increasing the number of cell types and conditions that can be profiled

This chapter has two main aims. First, I wanted to estimate how well iPSC-derived macrophages (IPSDMs) recapitulate known aspects of macrophage immune response on the chromatin level. Secondly, I aimed to understand how condition-specific are genetic effects on the chromatin level and how these effects propagate to changes in gene expression. To study these two questions, we used ATAC-seq to measure chromatin accessibility of IPSDMs in the same four experimental conditions (naive, IFNɣ, *Salmonella* and IFNɣ + *Salmonella*) that were used for eQTL mapping Chapter 4 in 31-42 individuals.

As highlighted in Chapter 4, the signalling pathways and transcription factors (TFs) activated by IFNɣ and *Salmonella* have been well characterised. Briefly, the activated TFs together with the DNA motifs that they recognise are illustrated on Figure 5.1. ChIP-seq experiments in both human and mouse macrophages have shown that thousands of regulatory elements change their activity in response to these and other stimuli (Kaikkonen et al., 2013; Ostuni et al., 2013; Qiao et al., 2013; Schmidt et al., 2016). Furthermore, while most of the enhancers that became active after stimulation are already primed in the naive state, a subset of them are created *de novo* after the simulation (Kaikkonen et al., 2013; Ostuni et al., 2013).

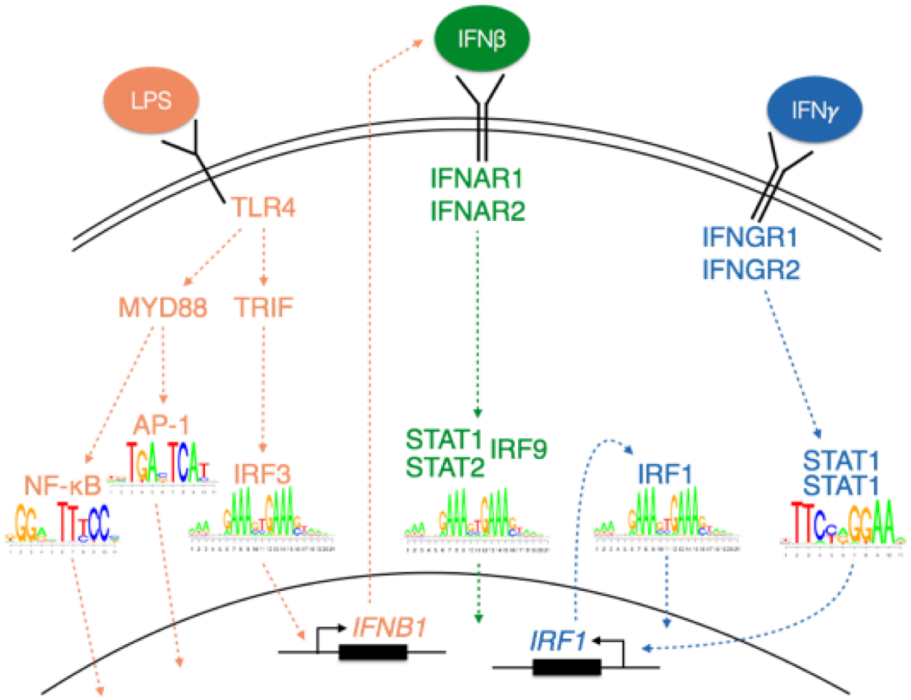

Figure 5.1: Main signalling pathways activated in macrophages after *Salmonella* infection and IFNɣ stimulation. Macrophages recognise LPS on the *Salmonella* cell wall via the TLR4 receptor that leads to the downstream activation of the nuclear factor kappa B (NF-κB) and activator protein 1 (AP-1) (Takeuchi and Akira, 2010) as well as the interferon response factor 3 (IRF3) (Doyle et al., 2002) TFs. IFNɣ, on the other hand, activates signal transducer and activator of transcription 1 (STAT1) and IRF1 TFs. Finally, IRF3 can also activate the IFNβ signalling pathway that culminates with the activation of STAT1-STAT2-IRF9 complex. While AP-1, NF-κB and STAT1 all recognise distinct DNA motifs (illustrated by the sequence logos under the TF names), IRF3, STAT1-STAT2-IRF9 and IRF1 recognise similar interferon-specific response element (ISRE) motif.

By using motif enrichment analysis and comparing IPSDM ATAC-seq signal to published ChIP-seq experiments, I was able to show that IPSDMs are able to recapitulate many known aspects of chromatin dynamics in macrophage immune response. Secondly, I identified caQTLs for 4,000-10,000 ATAC-seq peaks depending on the condition and showed that approximately 25% of the caQTLs were condition specific. I also identify a small number of 'multi-peak' caQTLs where a single putative causal variant influenced chromatin accessibility of multiple independent peaks. I showed that some single-peak caQTLs can become multi-peak caQTLs after stimulation, thus highlighting hierarchical relationships between regulatory elements. Finally, I showed that for approximately 50% of stimulation-specific eQTLs the corresponding caQTL was

visible already in the naive state, suggesting that a proportion of caQTLs correspond to primed enhancers that are waiting for an appropriate environmental signal before regulating gene expression.

5.2 Methods

The experimental protocols for cell culture and stimulation experiments are described in Chapter 3. This section focusses on methods that were specific to the chromatin accessibility part of the study.

5.2.1 ATAC-seq

Experimental procedures

Approximately 150,000 cells were seeded into 1 well of a 6-well plate and treated identically to the RNA-seq samples. After stimulation, cells were washed once with ice-cold D-PBS and incubated for 12 minutes on ice in 500 µl sucrose buffer (10 mM Tris-Cl pH 7.5, 3 mM $CaCl_2$, 2mM $MgCl_2$, 0.32 M sucrose). After 12 minutes, 25 µl of 10% Triton-X-100 (FC = 0.5%) was added and the cells were incubated for another 6 minutes to release the nuclei. Cells were centrifuged at 300 rpm for 8 minutes at 4°C and the supernatant was discarded. Tagmentation was performed with Illumina Nextera DNA Sample Preparation Kit as specified in the original ATAC-seq protocol (Buenrostro et al., 2013). Finally, size selection was performed using agarose gel and SPRI beads (Kumasaka et al., 2016). Five samples were pooled per lane and 75 bp paired end reads were sequenced on Illumina HiSeq 2000 using the V4 chemistry.

Read alignment

Illumina Nextera sequencing adapters were trimmed using skewer v0.1.127 (Jiang et al., 2014) in paired end mode. Trimmed reads were aligned to GRCh38 human reference genome using bwa mem v0.7.12 (Li, 2013) (Li, 2013). Reads mapping to the mitochondrial genome and alternative contigs were excluded from all downstream analysis. Picard 1.134 MarkDuplicates was used to remove duplicate fragments. I used verifyBamID (Jun et al., 2012) 1.1.2 to detect and correct potential sample swaps between individuals. Fragment coverage BigWig files were constructed using bedtools v2.17.0 (Quinlan and Hall, 2010).

Peak calling

I used MACS2 (Zhang et al., 2008b) v2.1.0 with '--nomodel --shift -25 --extsize 50 -q 0.01' to identify open chromatin regions (peaks) that were enriched for transposase integration sites compared to the background at 1% FDR level. With these parameters I detected between 31,658 and 208,330 peaks per sample. I constructed consensus peak sets in each condition separately by pooling all of the peak calls from all of the samples. For each peak, I counted the number samples in which that peak was identified and calculated the union of all peaks that were detected in at least 3 samples. Finally, I pooled the consensus peaks from all four conditions to obtain the final set of 296,220 unique peaks that were used for all downstream analyses. I used featureCounts (Liao et al., 2014) v.1.5.0 to count fragments overlapping consensus peak annotations and ASEReadCounter (Castel et al., 2015) from Genome Analysis Toolkit (GATK) to quantify allele-specific chromatin accessibility.

Sample quality control

I used the following criteria to assess the quality of ATAC-seq samples:

- *Assigned fragment count* - the total number of paired end fragments assigned to peaks by featureCounts.
- *Mitochondrial fraction* - fraction of total fragments aligned to the mitochondrial genome.
- *Assigned fraction* - fraction of non-mitochondrial reads assigned to consensus peaks. A measure of signal-to-noise ratio.
- *Duplicated fraction* - fraction of fragments that were marked as duplicates by Picard MarkDuplicates.
- *Peak count* - number of peaks called by MACS2.
- *Length ratio* - # of short fragments (< 150 nt) / # long fragments (>= 150 nt). This measures if the read length distribution has characteristic ATAC-seq profile with clearly visible mono-nucleosomal and di-nucleosomal peaks.

I used these criteria to exclude 5 samples prior to performing caQTL mapping. One sample was excluded because of very low assigned fraction (~10%) and peak count, two more were excluded because of extremely large length ratio (>7) and an uncharacteristic ATAC-seq profile. The final two samples were excluded because they appeared to be outliers in the principal component analysis.

Differentially accessible regions

I used limma voom v3.26.3 (Law et al., 2014) to identify 63,430 peaks that were more than 4-fold differentially accessible (FDR < 0.01) between naive and any one of the stimulated conditions. I noticed that limma voom was sensitive to lower quality samples. Therefore, I only used high quality samples from 16 donors (64 samples) for the differential accessibility analysis. Subsequently, I quantile-normalised the peak accessibility data using cqn (Hansen et al., 2012), calculated the mean accessibility of each peak in each condition and used Mfuzz v.2.28 (Kumar and E Futschik, 2007) to cluster the peaks into seven distinct activity patterns. For principal component analysis (PCA) I normalised the peak fragment counts data using transcripts per million (TPM) (Wagner et al., 2012) approach.

Motif enrichment

I downloaded the CIS-BP (Weirauch et al., 2014) human TF motif database from the MEME website and used FIMO (Grant et al., 2011) to identify the occurrences of all TF motifs within the ATAC consensus peaks with FIMO threshold p-value < 1e-5. I also performed the same motif scan for 2 kb promoter sequences upstream of 21,350 human genes (downloaded from the PWMEnrich (Stojnic and Diez, 2015) R package) and used this as the background set. I used Fisher's exact test to identify motifs that occurred significantly more often in macrophage open chromatin regions compared to the background promoter sequences. Because the CIS-BP database contains many redundant motifs, I manually selected 21 representative motifs for downstream analysis corresponding to the major TFs important in macrophage biology: AP-1, IRF-family, ETS-family (PU.1, ELF1, FLI1), NF-κB, CEBPα, CEBPβ, ATF4, CTCF, STAT1, MAFB, MEF2A and USF1. I also used Fisher's exact test to identify motifs that were specifically enriched in each cluster of differentially accessible peaks compared to the background of all macrophage ATAC peaks.

5.2.2 ChIP-seq data analysis

The public ChIP-seq datasets used in this study are summarised in section 'Summary of public ChIP-seq datasets used in the analyses'. Single-end datasets (Pham *et al* and Qiao *et al*) were aligned to the GRCh38 human reference genome using bwa aln v0.7.12 with default parameters. Paired-end datasets (Reschen *et al*, Schmidt *et al* and Wong *et al*) were aligned to the GRCh38 reference genome using bwa mem v0.7.12 with the -M flag set. Only properly paired reads were used for downstream analysis. Duplicate reads were removed with Picard

v1.134 MarkDuplicates with the 'REMOVE_DUPLICATES=true' parameter set. I used bedtools v2.17.0 (Quinlan and Hall, 2010) to construct genome wide read (single-end) or fragment (paired-end) coverage tracks in BigWig format. I called peaks using MACS2 v2.1.0 with '-q 0.01' option.

Summary of public ChIP-seq datasets used in the analyses

[Pham *et al*] (Pham et al., 2012, 2013)
Purification: Gradient centrifugation (85% pure monocytes)
Culture conditions: Purified monocytes were differentiated into macrophages in RPMI 1640 medium (Biochrom) supplemented with 2% human pooled AB-group serum on Teflon foils for up to 7 days. Macrophages usually > 95% pure.
Stimulations: Naive only
Accession: GSE31621, GSE43098
PMID: 22550342, 23658224
ChIP-seq antibodies: CTCF, PU.1, C/EBPβ, H3K4me1, H3K27ac, H2AZ.
Sequencing: 36 bp single-end reads on Illumina GA I/II.
Replicates: 1

[Qiao *et al*] (Qiao et al., 2013)
Purification: Gradient centrifugation followed by positive selection with anti-CD14 beads (Miltenyi Biotec) (>97% pure)
Culture conditions: Monocytes were cultured in RPMI 1640 (Invitrogen) supplemented with 10% defined FBS (HyClone) and 10 ng/mL M-CSF (Peprotech) (days unknown).
Stimulations: Cells were treated with or without IFN-g (100U/ml) for 24 hours, and then stimulated with LPS (50 ng/ml) for 3 hours (STAT1, H3K27Ac) or 6 hours (IRF1). (Naive, IFNɣ, LPS, IFNɣ + LPS)
Accession: GSE43036
PMID: 24012417
ChIP-seq antibodies: STAT1, H3K27ac, IRF1
Sequencing: 50 bp single-end reads on Illumina HiSeq 2000
Replicates: Up to 2 per condition

[Reschen *et al*] (Reschen et al., 2015)
Purification: Gradient centrifugation followed by positive selection with anti-CD14 beads (Miltenyi Biotec) (>95% pure)

Culture conditions: Cells were maintained in RPMI 1640 medium with 10% FCF, 4 mM L-glutamine, 50 units/ml penicillin and 50 µg/ml streptomycin (Sigma, St Louis, MO), supplemented with 50 ng/ml M-CSF (eBioscience, San Diego, CA) for 7 days.
Stimulations: Naive and oxLDL (50 µg/ml, 48h)
Accession: GSE54975
PMID: 25835000
ChIP-seq antibodies: C/EBPβ, H3K27ac, FAIRE-seq
Sequencing: 50 bp paired-end reads on HiSeq 2000/2500.
Replicates: 2-4

[Schmidt et al] (Schmidt et al., 2016)
Purification: Gradient centrifugation followed by positive selection with anti-CD14 beads (Miltenyi Biotec)
Culture conditions: Monocytes were cultured for 72h with GM-CSF (500 U/ml) in RPMI 1640 medium containing 10% FCS.
Stimulations: Naive, IFNɣ (200 U/ml, 72h), TPP (TNF (800 U/ml), PGE2 (1µg/ml) and Pam3CSK4 (1µg/ml), 72h), IL-4 (500 U/ml, 72h).
Accession: GSE66594
PMID: 26729620
ChIP-seq antibodies: PU.1, H3K27me3, H3K27ac, H3K4me1
Sequencing: 75 bp single-end on Illumina HiSeq 1000

[Wong *et al*] (Wong et al., 2014)
Purification: Gradient centrifugation followed by positive selection with anti-CD14 beads (Miltenyi Biotec)
Culture conditions: Experiments were done on monocytes.
Stimulations: Naive and IFNɣ (10 ng/mL, 24 h)
Accession: E-MTAB-2424
PMID: 25366989
ChIP-seq antibodies: CIITA, RFX5
Sequencing: 51 bp paired-end reads on HiSeq

Detecting regions with differential H3K27Ac signal

I performed differential histone acetylation analysis on the Qiao et al (Qiao et al., 2013) dataset to compare it to our ATAC-seq data. As H3K27Ac peaks are generally broader than ATAC-seq

peaks, I used MACS2 to call both broad and narrow peaks. Within each condition I only kept broad and narrow peaks that were detected at the 1% FDR threshold in both biological replicates. By visualising the data in a genome browser, I observed that at the 1% FDR threshold MACS2 called an excess of broad peaks compared to the narrow peaks so I further removed broad peaks that did not overlap any narrow peaks in the same condition. I then defined the union of broad peaks identified in each condition as the consensus set of peaks. I used featureCounts (Liao et al., 2014) to count the number of reads overlapping the consensus peaks in each sample. Finally, I used limma voom (Law et al., 2014) to identify peaks that showed at least 2-fold differential histone acetylation between naive and one of the stimulated states at 10% FDR. I used less stringent fold change and FDR thresholds for the histone acetylation data compared to the ATAC-seq data, because the broad histone peaks were less dynamic than the narrow ATAC peaks and because the histone dataset had only two biological replicates.

Peak overlap analysis

I used a permutation-based approach implemented in the Genomic Association Test (GAT) (Heger et al., 2013) software to test if the overlap between two sets of genomic annotations (such as ATAC-seq peaks and H3K27Ac peaks) was larger than expected by chance.

5.2.3 Chromatin accessibility QTL mapping

I used identical methodology to map eQTLs and caQTLs and assess their condition specificity. The full details of the pipeline are described in Chapter 4. Briefly, this involved mapping caQTLs using linear and allele-specific models, assessing replicability of caQTLs between conditions and using a linear model to identify peaks that show significant interactions between genotype and condition (condition-specific caQTLs). This section describes the areas where caQTL mapping differed from eQTL mapping. The size of the cis window for the caQTL mapping was +/- 50kb around the peak.

Filtering condition-specific caQTLs by effect size

I extracted the RASQUAL caQTL effect size estimates π for each peak-variant pair in each conditions and converted them into $\log_2$ fold changes between the two homozygotes using the formula $\log_2FC = -\log_2(\pi/(1-\pi))$. I then filtered the significant condition-specific caQTLs by requiring the maximal absolute $\log_2FC$ across conditions $|\log_2FC_{max}|$ to be > 0.59 (corresponding to 1.5-fold difference between the homozygotes), the minimal absolute $\log_2FC$ across conditions

$|\log_2 FC_{min}|$ to be < 0.59 and the absolute difference between the two $|\log_2 FC_{max} - \log_2 FC_{min}|$ to be >0.59.

QTL replicability between conditions

For the Storey's π_1 analysis (Nica et al., 2011), I identified caQTL peaks at 10% FDR in one condition, took their permutation-based lead variant p-values in the other condition and used the qvalue (Dabney et al., 2010) package to estimate the proportion of non-null p-values. For the lead variant concordance analysis, I identified caQTL peaks together with their lead variants at 1% FDR in one condition, extracted their lead variants in the other condition and counted how often R^2 between the two lead variants of the same caQTL peak was > 0.8.

Motif disruption analysis

I limited motif disruption analysis to caQTL peaks that did not contain associated indels and had <= 3 overlapping SNPs in them. For each SNP-peak pair I focussed on the sequence +/- 25 bp from the SNP. I constructed both reference and alternative versions of the sequence and used TFBSTools (Tan and Lenhard, 2016) to calculate the relative binding scores for both alleles (expressed as percentage from 0-100%). I considered the variant to be motif disrupting if the difference in relative binding score between the two alleles was > 3 percentage points. I also required the relative binding score for at least one of the alleles to be >= 85% of the theoretical maximum. This filter was necessary to exclude potential motif disruption events in very weak motif matches that are not likely to correspond to binding *in vivo* and is similar to the default recommended by TFBSTools. I used the hypergeometric test to identify motifs that were significantly more often disrupted in one of the six condition-specific caQTL clusters compared to all caQTLs.

Identifying condition-specific dependent peaks

To identify condition-specific dependent peaks, I tested if the effect size of the caQTL changed differently for master and dependent peaks between two pairs of conditions. This was equivalent to testing the significance of a three-way interactions between genotype, peak (master or dependent) and condition. I implemented this as the comparison of two standard linear models in R:

H0: y ~ peak + condition + peak*condition + genotype*peak + genotype*condition + covariates

H1: y ~ peak + condition + peak*condition + genotype*peak + genotype*condition + genotype*condition*peak + covariates

Similarly to condition-specific caQTL analysis, I used the first three principal components calculated separately for each condition as covariates in the model. I used the log_2FC from RASQUAL as the measure of caQTL effect size. To identify true condition-specific dependent peaks, I further filtered the results by requiring the absolute log_2FC of the master peak to be > 0.59 (1.5-fold) in the naive condition and the change in the log_2FC for the dependent peak between the naive and stimulated condition to be > 0.59.

5.3 Quantifying chromatin accessibility

First, I tested whether the chromatin accessibility profile in IPSDMs was similar to that of primary macrophages. After multiple pre-processing steps (see Methods for details), I identified a total of 296,220 consensus ATAC-seq peaks in IPSDMs across four experimental conditions and quantified their accessibility. Principal component analysis (PCA) of the data revealed four distinct clusters corresponding to the four experimental conditions (Figure 5.2A).

To identify the transcription factors (TFs) that drive chromatin accessibility at these macrophage peaks I compared them to 21,350 human promoter sequences. I found that accessible chromatin regions in macrophages were enriched for binding motifs of multiple TFs that play important roles in macrophage function. The two most enriched motifs belonged to the AP-1 and PU.1 TFs (Figure 5.2B) whose collaborative interactions are well known to establish macrophage specific enhancers (Heinz et al., 2010). Other motifs enriched in the ATAC-seq peaks belonged to multiple TFs recognising the interferon-specific response element (ISRE) motif (IRF2, STAT2, IRF8, IRF1) as well as the CEBPα and CEBPβ TFs.

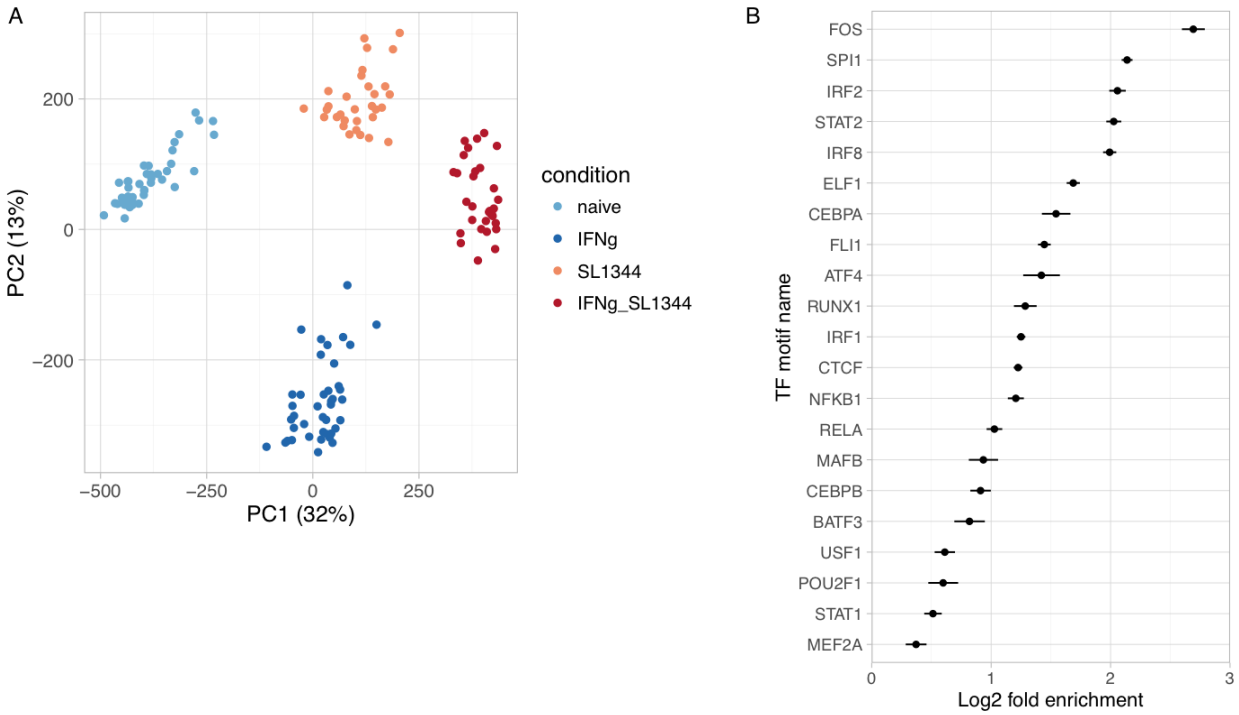

Figure 5.2: Summary of chromatin accessibility data. (A) PCA of macrophage chromatin accessibility data in four conditions. Axis labels indicate the percentage of variance explained by the first two principal components. **(B)** A selection of 21 representative TF motifs that are enriched in macrophage ATAC peaks relative to 21,350 human promoter sequences.

5.3.1 Differential chromatin accessibility between conditions

Many condition specific TFs are likely to regulate gene expression by altering chromatin accessibility. I next attempted to identify which TFs regulate chromatin accessibility in response to the three different stimuli in our study. I identified 63,430 peaks that were more than 4-fold differentially accessible (FDR < 0.01) between naive and any one of the stimulated conditions. I clustered the differential peaks into seven distinct activity patterns (Figure 5.3A) and to aid interpretation, I further grouped the seven clusters into four major groups. I used *post hoc* grouping of the clusters instead of clustering directly into four clusters because specifying a smaller number of clusters did not identify all of the four main patterns (See Figure 5.3A). I then used Fisher's exact test to identify TF motifs from the CIS-BP database that were enriched in each group of differentially accessible peaks relative to all macrophage ATAC peaks (Figure 5.3B).

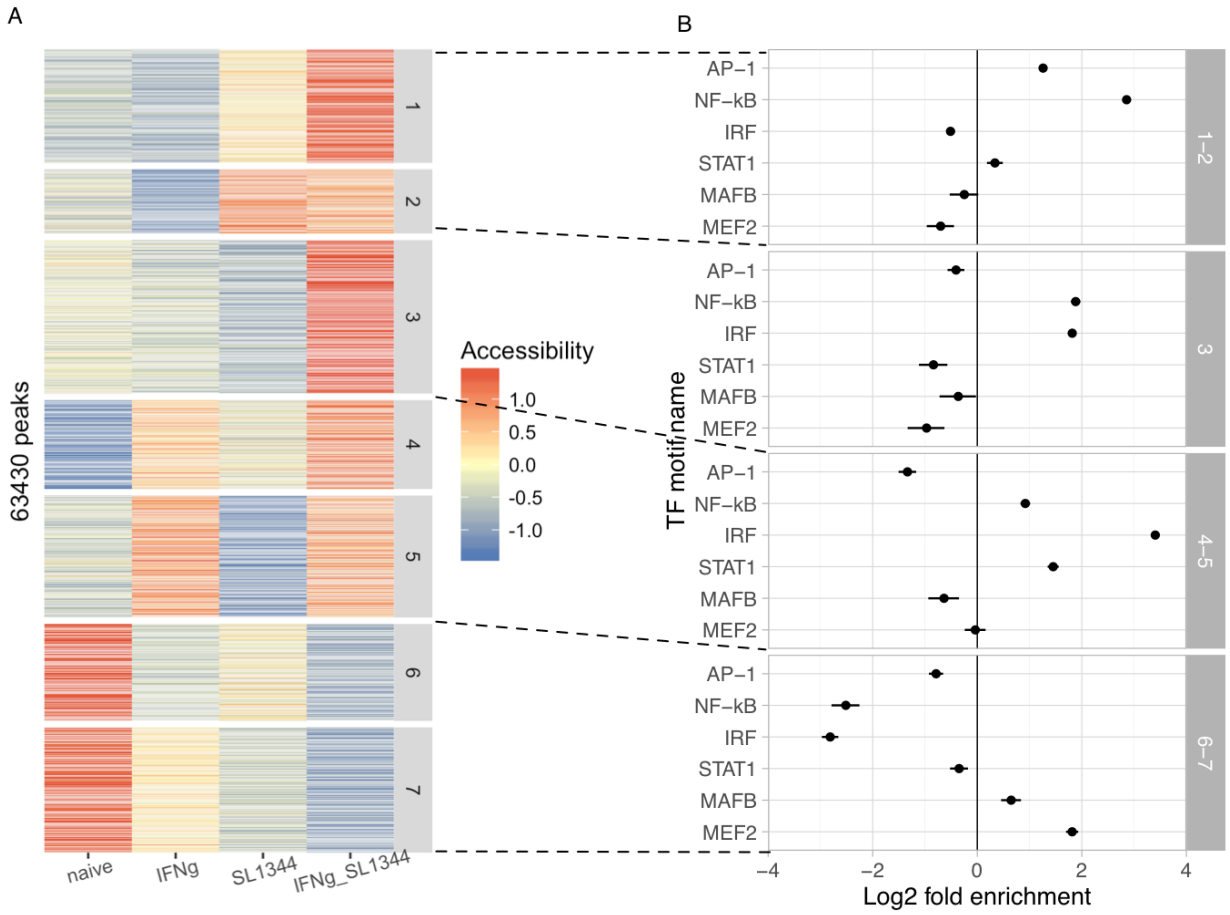

Figure 5.3: Dynamics of chromatin accessibility between conditions. (A) The 63,350 differentially accessible open chromatin regions were clustered into seven distinct patterns using c-means clustering implemented in the MFuzz packages. The clusters have been grouped into four groups according to whether their accessibility increased after *Salmonella* infection (clusters 1 and 2), IFNɣ stimulation (clusters 4 and 5), synergistically after both stimuli (cluster 3) or decreases after stimulation (clusters 6 and 7). **(B)** Enrichment of transcription factor motifs in each of the four groups.

Clusters 1 and 2, both of which became more accessible after *Salmonella* infection, were specifically enriched for NF-κB and AP-1 motifs, the two main TFs activated downstream of TLR4 signalling (Takeuchi and Akira, 2010). Cluster 3, which became accessible only after both of the stimuli were present, was enriched for the IRF (ISRE) and NF-κB motifs, suggesting possible collaborative interactions between IFNɣ-induced IRF1 and TLR4-activated NF-κB TFs that have been previously reported (Negishi et al., 2006). However, the motif analysis that I have performed does not distinguish between IRF1 and other IRF factors, because all IRF

factors have similar sequence preferences. In contrast, clusters 4 and 5 were activated by IFNɣ and were enriched for IRF and STAT1 motifs, consistent with the activation of STAT1 and IRF1 downstream of IFNɣ signalling (Schroder et al., 2004).

Finally, clusters 6-7, where accessibility decreased in response to all of the stimuli, were enriched for MEF2 and MAFB motifs. Interestingly, MafB binding has recently been shown to suppress self-renewal–associated macrophage enhancers in mouse and knocking out MafB together with c-Maf is sufficient to generate immortalised macrophages (Aziz et al., 2009; Soucie et al., 2016). This is further supported by our observation in Chapter 4 that genes downregulated by IFNɣ were strongly enriched for cell cycle and DNA replication pathways (Figure 4.5) and consistent with multiple reports that stimulation with IFNɣ induces cell cycle arrest in macrophages (Schroder et al., 2004; Xaus et al., 1999)

5.3.2 Overlap with ChIP-seq signals

Motif enrichment at differentially accessible peaks showed that iPSDMs activated the same set of TFs after stimulation that we would expect from primary monocyte-derived macrophages. However, it is not clear from motif enrichment alone if these TFs bind to the same genomic loci in both cell types. Unfortunately, there was no ATAC-seq data available from monocyte-derived macrophages (MDMs) from the same conditions to perform a direct comparison. Therefore, we resorted to comparing iPSDM ATAC peaks to multiple publicly available primary MDM ChIP-seq datasets.

First, I focussed on the (Qiao et al., 2013) study that had measured histone 3 lysine 27 acetylation (H3K27ac) with ChIP-seq in MDMs in very similar conditions to ours (naive, 3h LPS, 24h IFNɣ and 24h IFNɣ + 3h LPS). I identify 11,735 differentially acetylated ChIP-seq peaks (FDR < 0.1, fold-change > 2) and clustered them into six clusters using MFuzz (See Methods for details) (Figure 5.4A). Since H3K27Ac peaks are generally much longer than ATAC-seq peaks (median lengths 3369 and 231 bp, respectively), I used permutation-based approach implemented in the Genomic Association Tester (GAT) (Heger et al., 2013) software to test if the overlap between different clusters of peaks was larger than expected. I found strong overlap between respective groups of peaks in IPSDM ATAC-seq and MDM H3K27Ac data, suggesting that overlapping regulatory elements become active in both cell types after similar experimental treatments (Figure 5.4B).

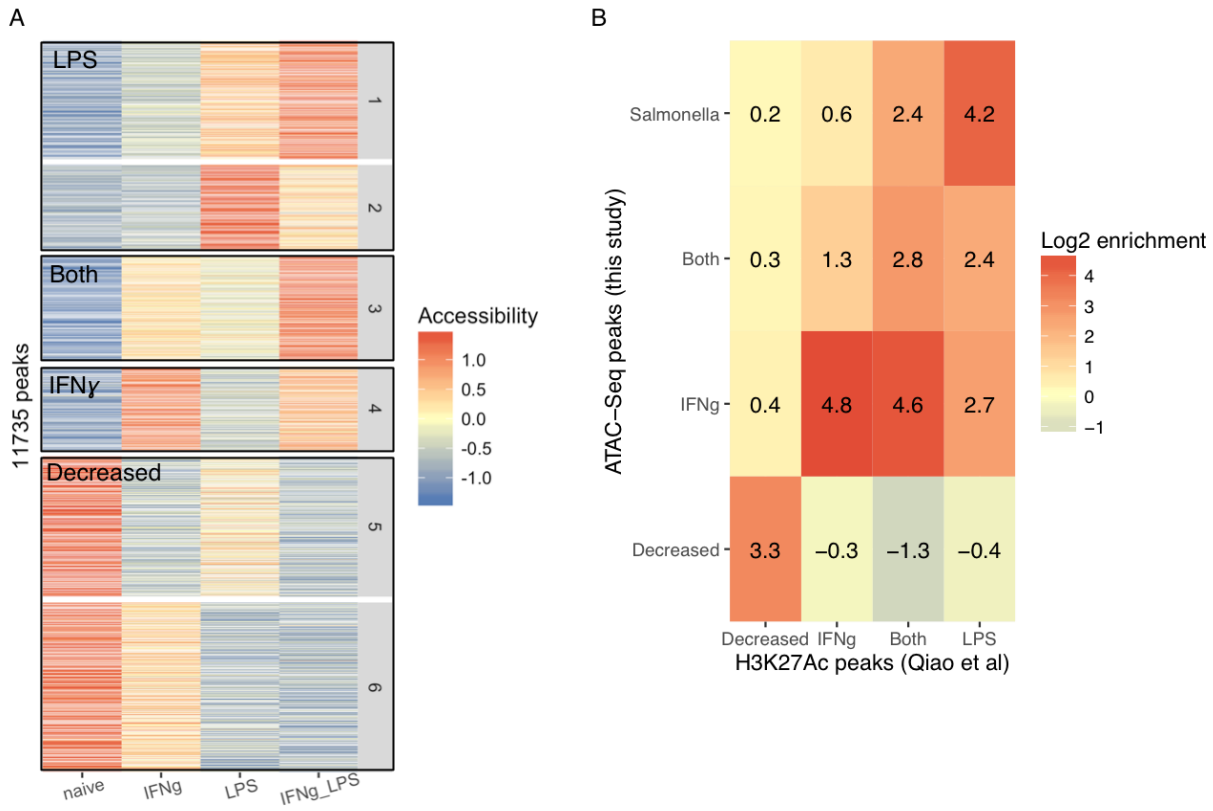

Figure 5.4: Concordance of chromatin changes between IPSDMs and MDMs. (**A)** Clustering of differential H3K27Ac peaks from (Qiao et al., 2013) study. The six clusters identified by MFuzz have been grouped into four groups based on whether the H3k27ac signal increases after LPS stimulation, IFNɣ stimulation, Both of the stimuli or decreases after stimulation. **(B)** Log_2 fold enrichment of overlap between differential peak groups identified in our IPSDM ATAC-seq data and MDM H3K27ac data. The log_2 fold enrichments of overlap were calculated using GAT (Heger et al., 2013).

I noticed that the gene expression level of the master regulator of MHC class II complex CIITA together with its downstream targets (MHC class II genes) was specifically upregulated after IFNɣ stimulation (Figure 5.5A, Figure 4.5). I therefore hypothesised that some of the ATAC peaks that appear after IFNɣ stimulation should correspond to CIITA binding events. Fortunately, (Wong et al., 2014) had performed ChIP-seq for CIITA and RFX5 TFs (two members of the same complex) in primary human monocytes before and after IFNɣ stimulation. After reanalysing their data, I identified peaks that were detected in both biological replicates and used GAT to test which ATAC peak clusters were enriched in the ChIP-seq peaks. I found that only ATAC peaks activated by IFNɣ were enriched for the CIITA and RFX5 ChIP-seq peaks

(Figure 5.5B), suggesting that IPSDMs use the same set of regulatory elements to upregulate MHC class II expression in response to IFNɣ as do primary monocytes.

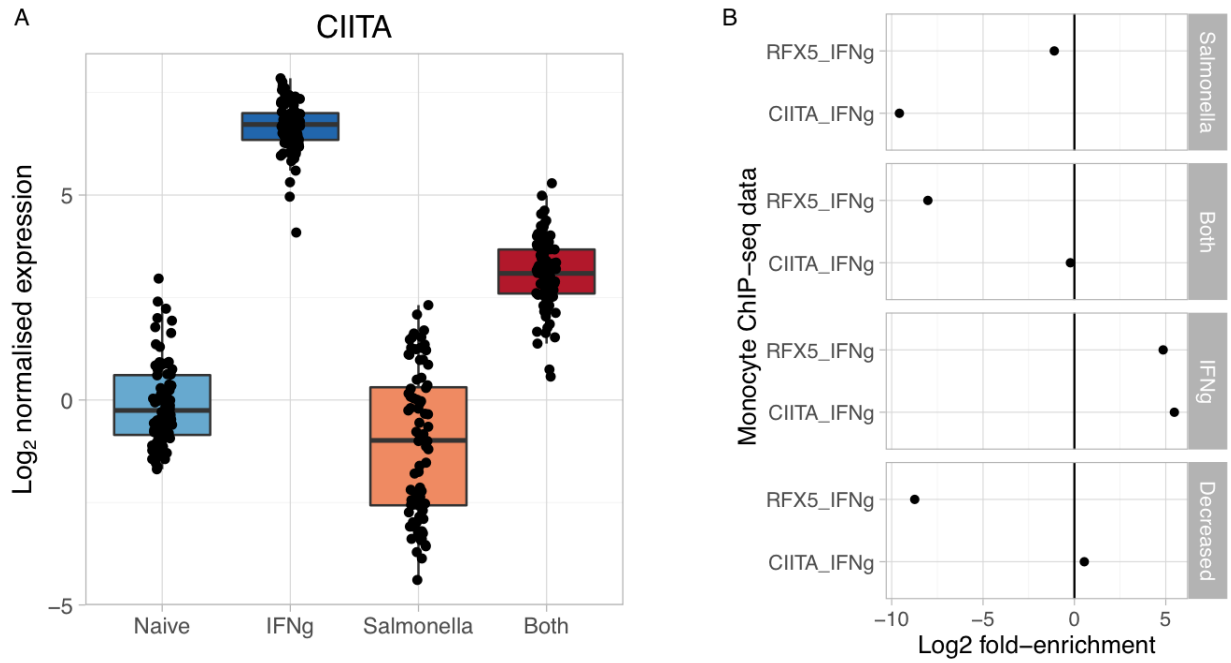

Figure 5.5: Regulation of MHC class II expression in IPSDMs. (A) Expression level of CIITA gene in IPSDMs in the four conditions. **(B)** Enrichment of monocyte RFX5 and CIITA ChIP-seq peaks (Wong et al., 2014) in IPSDM ATAC-seq peak clusters from Figure 5.3A.

5.4 Genetics of chromatin accessibility

Table 5.1: Number of caQTL peaks identified by the linear (FastQTL) and allele-specific (RASQUAL) models in a 50kb cis-window around the 296,220 peaks. Identical multiple testing correction approach was used for both FastQTL and RASQUAL results, i.e. for each peak, eigenMT (Davis et al., 2016) was used to correct for the number of independent tests performed in the cis-window and Benjamini-Hochberg FDR was used to correct for multiple independent peaks being tested.

condition	Sample size	FastQTL	RASQUAL
Naive	42	10735	10147
IFNɣ	41	10810	10192

Salmonella	31	5267	5493
Both	31	3782	4337

I used the same approaches to find chromatin accessibility QTLs (caQTLs) and assess their condition specificity that I used in Chapter 4 for eQTLs. Briefly, I used a standard linear model implemented in FastQTL (Ongen et al., 2016) software and the allele-specific model implemented in RASQUAL (Kumasaka et al., 2016) package to find caQTLs in a +/- 50kb window around each peak. I used both methods, because even though RASQUAL increases power to detect QTLs and fine map causal variants (Kumasaka et al., 2016), the summary statistics from the linear model can be directly used in replication and colocalisation analyses. Throughout this chapter, I will use *caQTL variants* to refer to the variants that are associated with chromatin accessibility at one or more open chromatin regions and I will use *caQTL peaks* to refer to the ATAC peaks that have at one or more independent significantly associated variants. Although RASQUAL and FastQTL identified similar number of caQTLs peaks at the 10% FDR level (Table 5.1), quantile-quantile (Q-Q) plots revealed that caQTLs from RASQUAL generally had much smaller p-values than caQTLs from the linear model (Figure 5.6), Consequently, using a stricter FDR threshold (such as 1%) resulted in more caQTLs detected with RASQUAL than with the linear model.

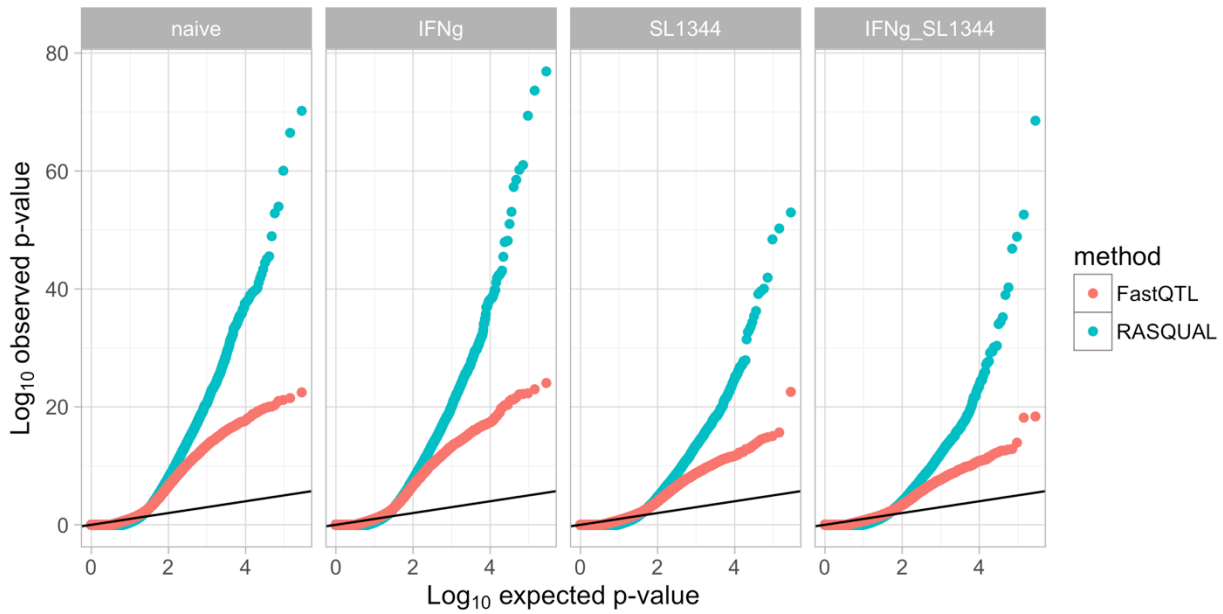

Figure 5.6: Q-Q plots of caQTLs identified by RASQUAL and FastQTL in each of the four conditions. On each plot, the solid line corresponds to the expected distribution of p-values under the null model of no association. The FastQTL and RASQUAL p-values have been corrected for the number of independent variants tested using eigenMT.

5.4.1 Fine mapping putative causal variants

Chromatin accessibility QTL variants have previously been observed to be strongly enriched either within the peak itself or within other nearby peaks (Degner et al., 2012; Kumasaka et al., 2016). This suggests that, unlike expression QTLs, the causal variants that underlie caQTLs are often likely to be found in a relatively small genomic region. Furthermore, recent evidence indicates that local caQTLs can influence chromatin accessibility by at least two conceptually distinct mechanisms (Deplancke et al., 2016). Most commonly, the causal variant is located within the accessible region and directly disrupts the binding of a sequence-specific factor. We refer to these caQTLs as ‘master’ caQTLs (Figure 5.7). However, sometimes a single causal variant in master caQTL peak can be associated with the accessibility of additional regions often many kilobases away from the master region forming so called ‘dependent’ caQTLs (Kumasaka et al., 2016) (Figure 5.7). The mechanisms that lead to the formation of dependent peaks have not yet been elucidated, but similar hierarchical relationships between regulatory elements have recently also been observed in the regulation of the WAP gene in mouse mammary tissue (Shin et al., 2016). Thus, discovering these associations between peaks can provide important insight into how multiple regulatory elements interact to regulate the expression of their target genes.

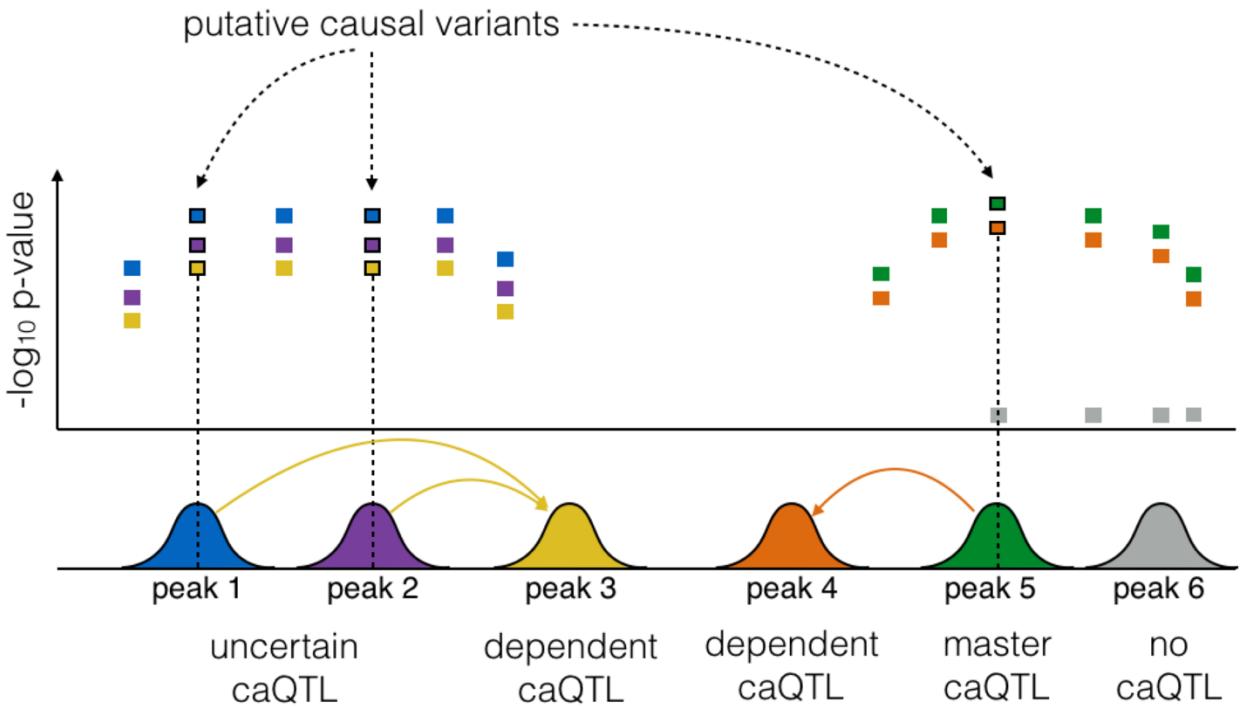

Figure 5.7: Heuristic approach to identify master and dependent caQTLs and their putative causal variants. Peak 5 is a *master caQTL peak*, because all of the variants in its credible set (green squares) overlap only peak 5 and no other caQTL region. Peak 1 and 2 are uncertain caQTLs, because the credible sets of peak 1 and peak 2 contain variants that overlap both peak. Peaks 3 and 4 are dependent caQTLs, because none of the variants in their credible set overlap the target peak, but they overlap some other peak (peaks 1 and 2 for peak 3 and peak 5 for peak 4).

I developed a heuristic approach to identify putative master and dependent caQTL peaks. Across the four conditions, I identified 13,872 caQTL peaks at 10% FDR. For each caQTL peak I first defined the credible set of causal variants as the set containing the lead SNP and all variants with $R^2 > 0.8$ with this SNP. In 88% of the cases (12,179 peaks) at least one variant in the credible set overlapped at least one consensus ATAC peak. The remaining 12% could be either false positive caQTLs or overlap open chromatin regions that were not detected by our peak calling approach. Furthermore, for 10,339/12,179 (85%) caQTL regions at least one variant in the credible set overlapped the region itself, confirming previous observations that caQTLs are highly local (Degner et al., 2012) (see regions 1, 2 and 5 on Figure 5.7).

However, observing that a variant in the credible set overlaps the corresponding caQTL peak does not necessarily mean that we have identified the true causal variant. In addition to many technical limitations (discussed below), an important biological limitation is that, because of high LD between variants, the same credible set can often overlap multiple caQTL peaks (see regions 1 and 2 on Figure 5.7 for illustration). In such cases it can be difficult to distinguish if there are two linked causal variants in two independent peaks or if there is only one causal variant in one of the peaks that influences the accessibility of both peaks. Thus, to identify putative master caQTLs I further required that the credible set variants overlapped strictly only one caQTL peak. As a result, I was able to identify 7,903 putative master caQTL peaks containing 11,854 putative causal variants. Furthermore, 69% of peaks contained only one putative causal variant and 95% of the regions contained <= 3 putative causal variants (Figure 5.8) highlighting the power of caQTLs in fine mapping causal variants.

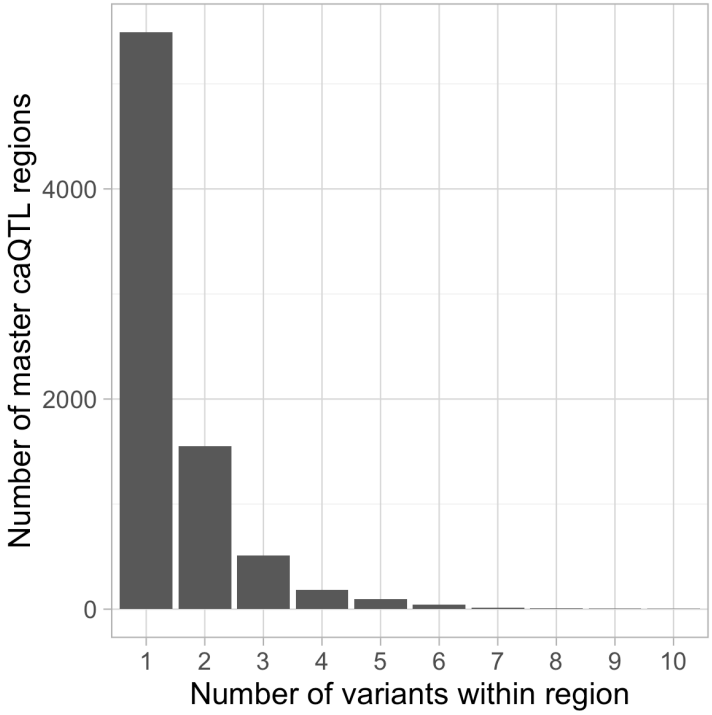

Figure 5.8. Histogram of the number of associated variants overlapping 7,903 putative master caQTL peaks.

Next, to identify dependent peaks, I focussed on the 1,840 caQTL peaks whose credible sets did not overlap the region itself. I found that for 753/1,840 peaks the credible set overlapped one of the putative master caQTL peaks identified above. This suggests that ~10% of the putative master caQTLs regions also have a dependent caQTL. However, this is likely to be an

underestimate, because dependent caQTLs generally have smaller effects than master caQTLs and we are less powered to detect them with our small sample size.

This approach has multiple limitations. First, it uses a fixed significance threshold (10% FDR) to identify open chromatin regions that do or do not have a caQTLs. This means that weaker dependent peaks will remain undetected. Secondly, some potential causal variants overlapping caQTL peaks might be missed, because region boundaries are defined by MACS2 peak calls that might themselves be inaccurate.

5.4.2 Assessing condition-specificity of caQTLs

I used two complementary approaches characterise the replicability of caQTLs between conditions. First, I used Storey's π_1 statistic (Nica et al., 2011) to estimate the fraction of caQTL peaks that were shared between each pair of conditions irrespective of their corresponding lead variants. I found that, similarly to eQTLs analysed in Chapter 3, the fraction of shared caQTL peaks varied between 0.75 and 0.90 with the lowest sharing observed between naive and IFNɣ + *Salmonella* conditions (Figure 5.9A). Secondly, I tested how often the lead caQTL variants were concordant ($R^2 > 0.8$) between two pairs of conditions (see Methods). I found that 75-80% of the lead caQTL variants were concordant between conditions which was considerably higher than 50-60% observed for eQTLs (Figure 5.9B). One possible reason for this discrepancy between the π_1 and lead variant concordance analyses could be that genes might have more independent QTLs between conditions than ATAC peaks.

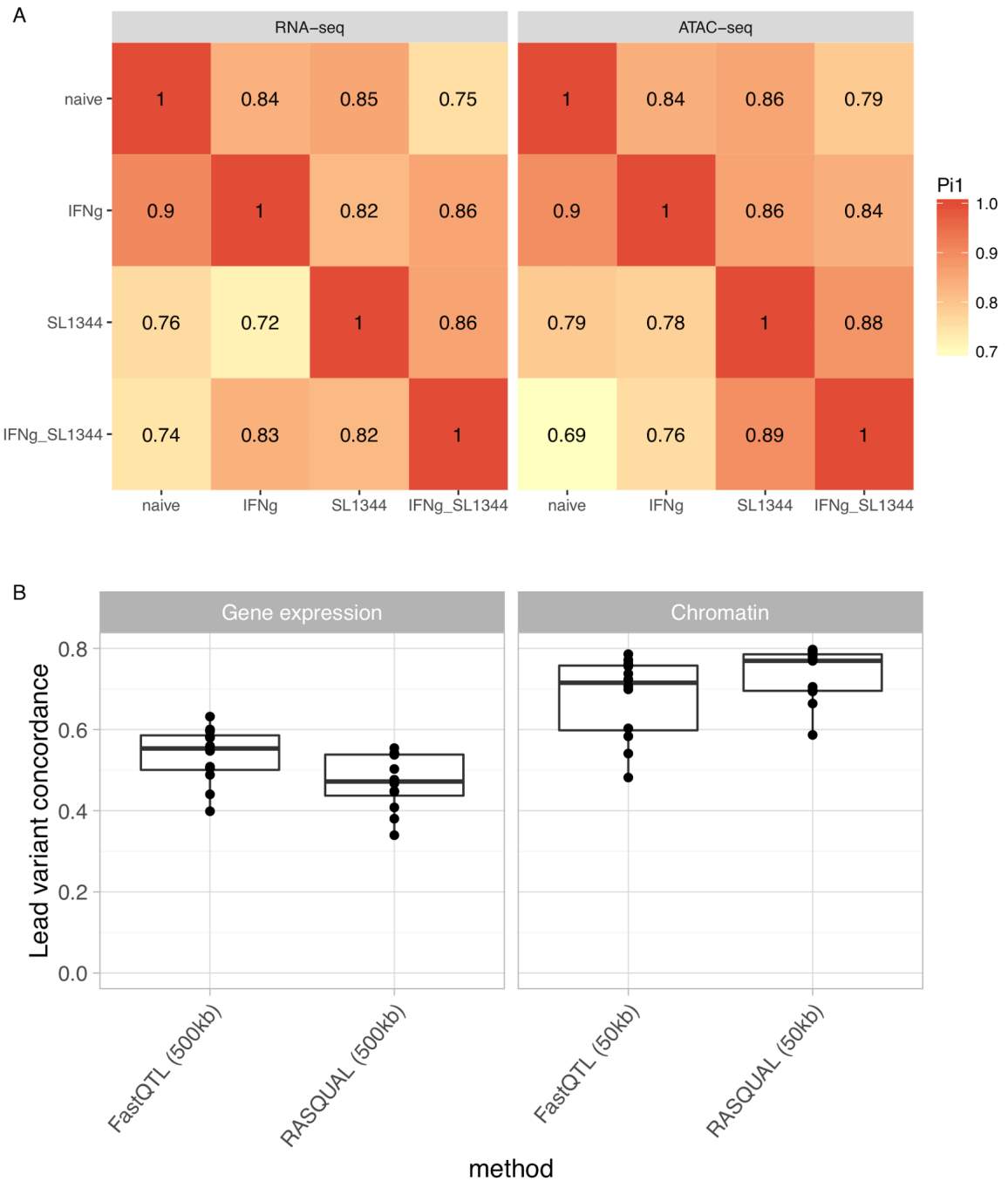

Figure 5.9: Replicability of eQTLs and caQTLs between conditions. (A) Feature-level replicability between conditions using the Storey's π_1 statistic. The π_1 statistic was calculated based on the FastQTL permutation p-values. **(B)** Pairwise concordance of the lead eQTL and caQTL variants for each feature. Each point corresponds to one pairwise comparison between two conditions. Concordance was calculated for both RASQUAL and FastQTL lead variants.

To identify individual peaks that have condition-specific caQTLs, I compiled all independent (R^2 < 0.8) variant-peak pairs across conditions and used two-way ANOVA to test for interactions between genotype and condition. Using sex and first three principal components of the dataset as covariates, I found that 4,947/16,924 (28%) caQTLs had significant interactions. After filtering out interactions with small effects, I identified 1,990 highly condition-specific caQTLs of which 1,113 appeared after stimulation (log_2FC_{naive} < 1) and 887 disappeared after stimulation (log_2FC_{naive} > 1).

I then clustered the condition-specific caQTLs based on their relative log_2FC across conditions. For the caQTLs that appeared after stimulation, I identified six distinct clusters of peaks (Figure 5.10A). I then tested if the likely causal variants for the condition-specific caQTLs were enriched for disrupting specific TF binding motifs compared to all caQTLs (Figure 5.10B). For this analysis I focussed only on the unique master peaks identified in Section 5.1 that had 1 to 3 likely causal variants overlapping the peak. I found that *Salmonella*-specific clusters 2 and 3 were enriched for disrupting NF-κB and AP-1 motifs whereas IFNɣ-specific clusters 5 and 6 were enriched for disrupting the ISRE motif. Furthermore, all condition-specific caQTLs were depleted for disrupting PU.1 binding motif (Figure 5.10B). This analysis suggests that condition-specific caQTLs are at least partly driven by variants that disrupt the binding sites of condition-specific TFs that are not active in the naive state. However, despite observing these motif enrichments, only ~15% condition specific caQTL could be explained by a motif disruption event at the thresholds that I used. Interestingly, I observed that almost all condition-specific caQTL peaks on Figure 5.10A were completely inaccessible in the naive condition and became most accessible in the condition with the largest caQTL effect size (Figure 5.11B). On the other hand, we observed no such relationship in the gene expression data where the genes with condition-specific eQTLs were on average equally highly expressed in all four conditions (Figure 5.11A).

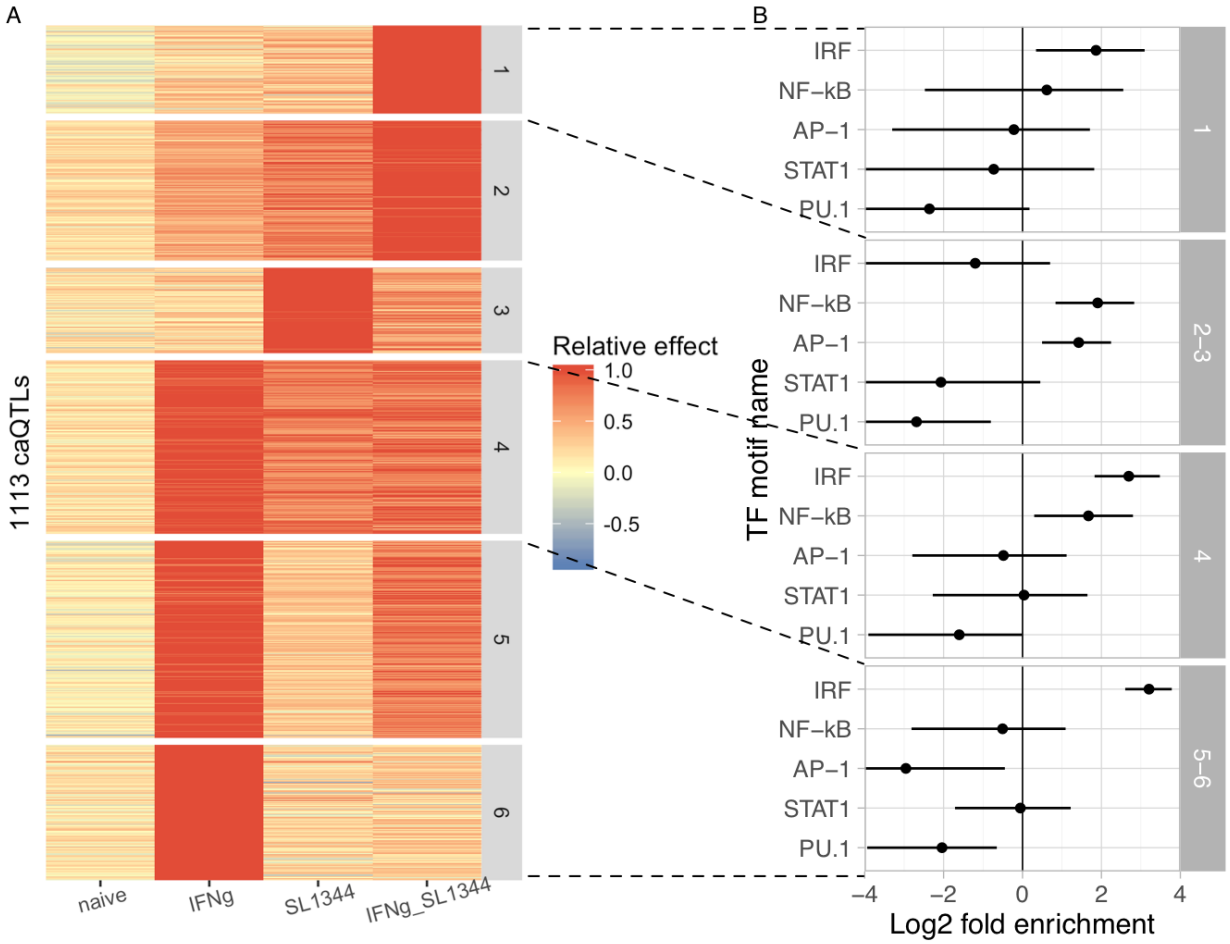

Figure 5.10: Identifying condition-specific caQTLs. (A) Condition-specific caQTLs clustered by their relative effect size. **(B)** Enrichment of TF motif disruptions in each cluster of caQTLs. The six cluster were grouped into four groups based on the caQTL activity pattern.

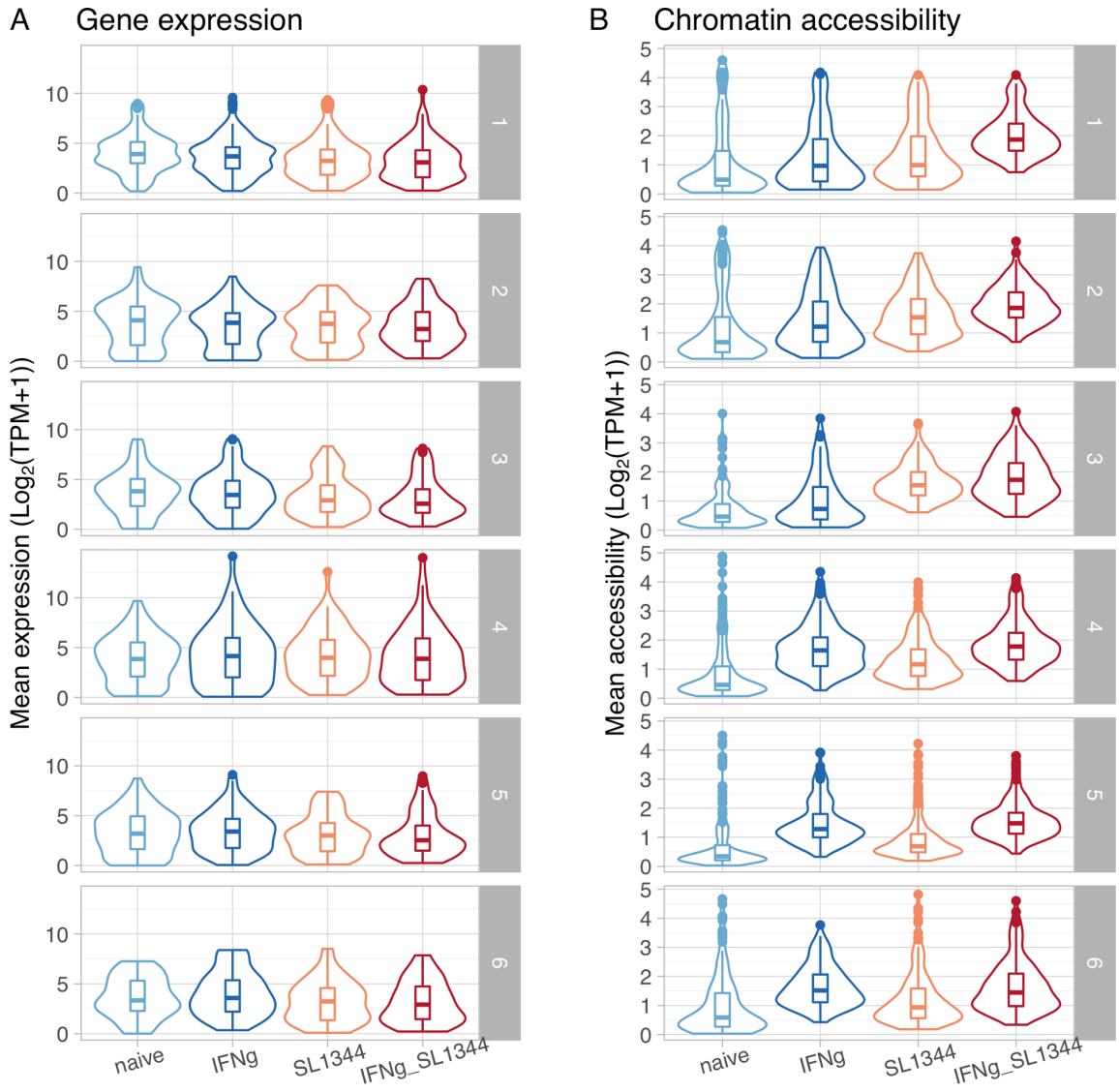

Figure 5.11: Relationship between QTL condition-specificity and mean gene expression or chromatin accessibility in each of the four conditions. (A) The distribution of mean gene expression values in each condition for the genes with conditions specific eQTLs from Figure 4.11C in Chapter 4. The numbered panels correspond to the same eQTL clusters that are shown on Figure 4.11C. **(B)** Mean chromatin accessibility of the ATAC-seq peaks from Figure 5.10A that had condition-specific caQTLs. The numbered panels correspond to the same caQTL clusters that are shown on Figure 5.10A.

5.4.3 Condition-specific dependent peaks

I noticed that some multi-peak caQTLs exhibited an interesting behaviour where the master caQTL peak was present in all conditions, but the dependent caQTL peak appeared or

disappeared in subset of the conditions (See Figure 5.12 for examples). To identify these cases systematically, I tested if the effect size of the caQTL changed differently for the master and dependent peak between conditions. This was equivalent to testing the significance of three-way interactions between genotype, peak (master or dependent) and condition (see Methods for details). After filtering by effect size, I identified 58 significant condition-specific dependent peaks. On the read coverage level 25/58 dependent peaks looked convincing, suggesting that the simple interaction test might have inflated false positive rate. The number of condition-specific dependent peaks that I identified is small, but with 31-42 samples we are clearly underpowered to detect most of these interactions.

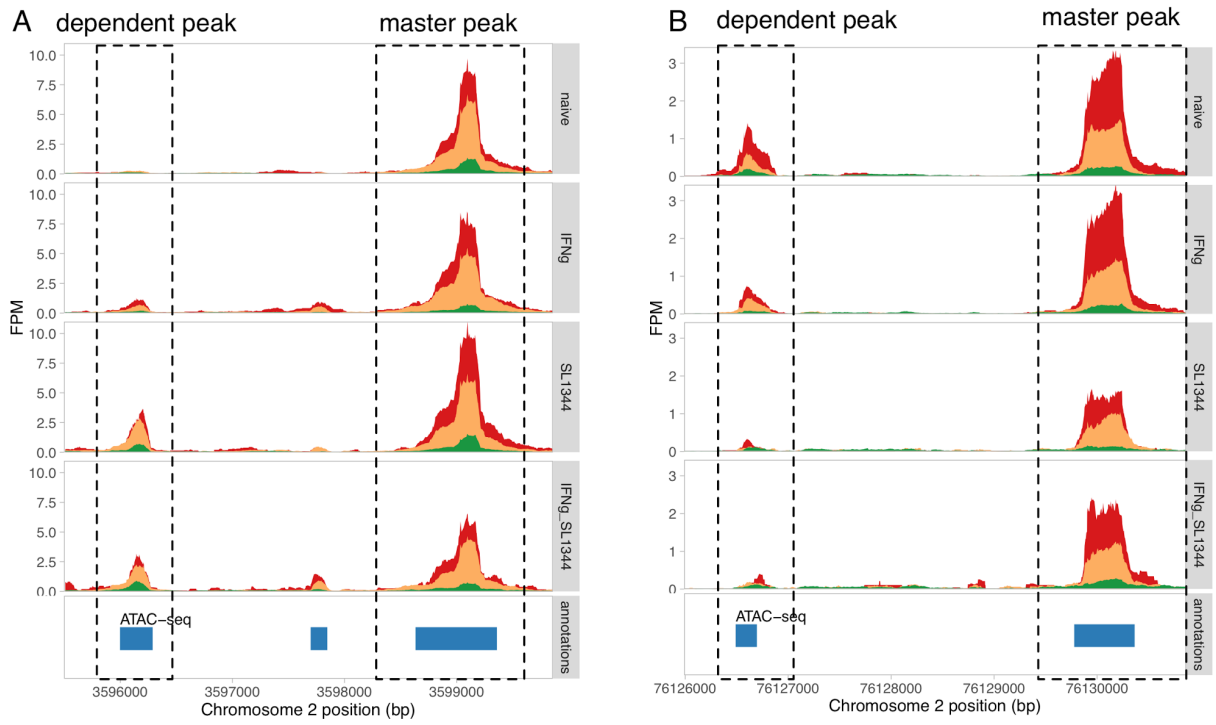

Figure 5.12: Two examples of condition-specific dependent peaks. (A) Dependent peak appears after *Salmonella* infection. **(B)** Dependent peak disappears after *Salmonella* infection.

5.5 Linking chromatin accessibility to the transcriptome

In addition to understanding the how sequence variation influences chromatin accessibility, combining caQTLs with eQTLs can also be used to link regulatory elements to their target genes.

5.5.1 Linking caQTLs to eQTLs

Knowing that a variant is an eQTL should increase our prior belief that the same variant might also be a chromatin accessibility QTL. However, modelling this formally can be challenging. I therefore decided to use two heuristic approaches with different levels of stringency. In the more stringent approach, I took lists of genome-wide significant eQTL genes and caQTL peaks (at 10% FDR) together with their lead variants and searched for instances where the two lead variants were in strong linkage disequilibrium ($R^2 > 0.8$). I did this either condition-by-condition or across conditions. I was able to find a corresponding caQTL for ~20% of the eQTLs. However, this approach strongly underestimated the true extent of overlap between eQTLs and caQTLs, because both our eQTL and caQTL mapping studies were underpowered. As an alternative approach, I focussed only on eQTL lead variants and tested in 100kb window around the lead variant for any associated ATAC peaks. I then used Bonferroni correction to account for multiple peaks tested per gene and used Benjamini-Hochberg FDR correction to account for multiple tested genes. With this approach I was able to identify corresponding caQTL for ~50% of the eQTLs.

Next, to understand how genetic effects propagate from chromatin to gene expression, I focussed on eQTLs that appeared after stimulation and that had a corresponding caQTL. One possible model is that chromatin accessibility largely mirrors gene expression and genetic effects become visible on both levels in the same condition. Alternatively, genetic effects on chromatin level might appear before they influence gene expression. To investigate these two hypotheses, I next examined the relative effect sizes of condition-specific eQTLs and corresponding caQTLs. I found that for approximately 50% of the eQTLs that appeared after IFNɣ stimulation or *Salmonella* infection the corresponding caQTL was already present before stimulation in naive cells (Figure 5.13). This is consistent with our previous observation that lead caQTL variants are more often concordant between conditions than lead eQTL variants (Figure 5.9B).

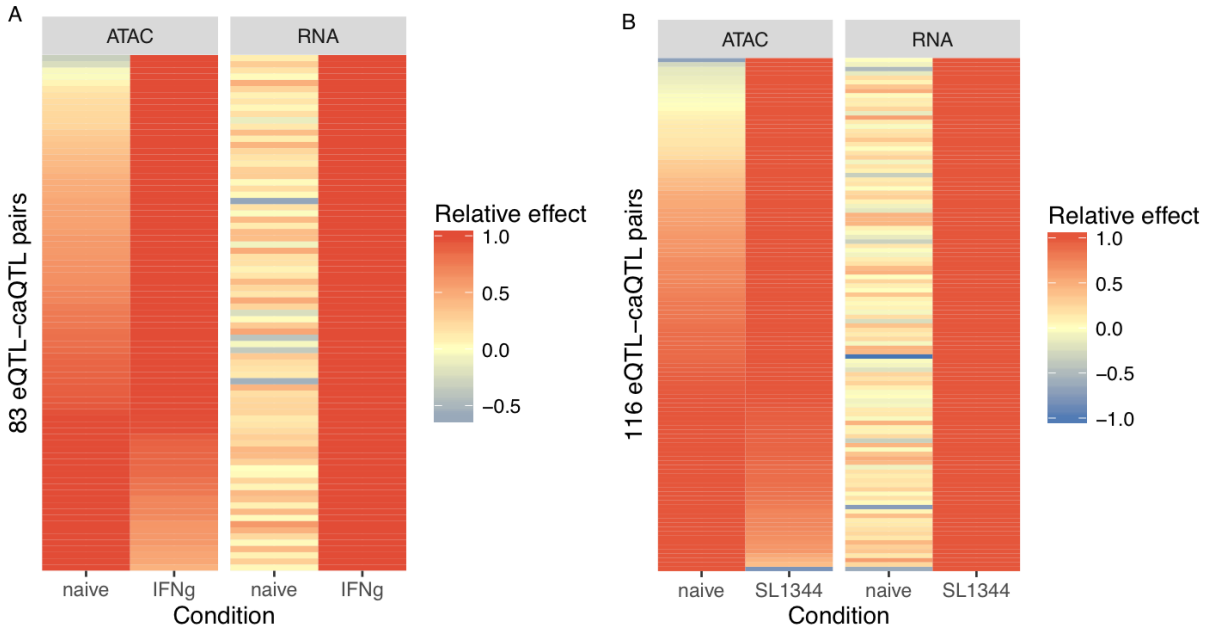

Figure 5.13: Comparison of effect sizes between condition-specific eQTLs and their corresponding caQTLs. (A) IFNɣ-specific eQTLs and their corresponding caQTLs. **(B)** *Salmonella*-specific eQTLs and their corresponding caQTLs.

A specific example is illustrated on Figure 5.14. The E1 peak is a master caQTL peak with a constitutive caQTL. The E1 peak has ten associated variants that are in almost perfect LD with each other (Figure 5.14A). However, only two of the ten variants overlap the E1 peak and only one of them (rs7594476) is located in the middle of a predicted PU.1 TF binding site (M6119_1.02 motif from in CIS-BP (Weirauch et al., 2014)). The alternative C allele has 9% lower relative binding affinity (87% vs 78%) that is consistent with reduced chromatin accessibility at the C allele. Furthermore, the same E1 peak has strong PU.1 ChIP-seq signal in a previously published macrophage dataset (Figure 5.14C) (Schmidt et al., 2016) suggesting that rs7594476 is the likely causal variant that alters chromatin accessibility at the E1 peak by disrupting a PU.1 binding site. The same variant is also associated with accessibility of 15 other ATAC peaks in the 200kb region, including the E2-E5 peaks shown Figure 5.14B. Interestingly, E2 is a condition specific dependent peak that appears after IFNɣ stimulation.

Finally, rs7594476 is also associated with the expression level of SPOPL and NXPH2 genes whose promoters are 200kb upstream and 90kb downstream from the peak, respectively. Colocalisation analysis revealed that the two eQTLs and the E1 caQTL are strongly colocalised (posterior probability = 0.98), suggesting that they are driven by the same causal variant.

Intriguingly, similarly to the E2 dependent peak, the eQTLs for SPOPL and NXPH2 genes become visible only after IFNɣ stimulation.

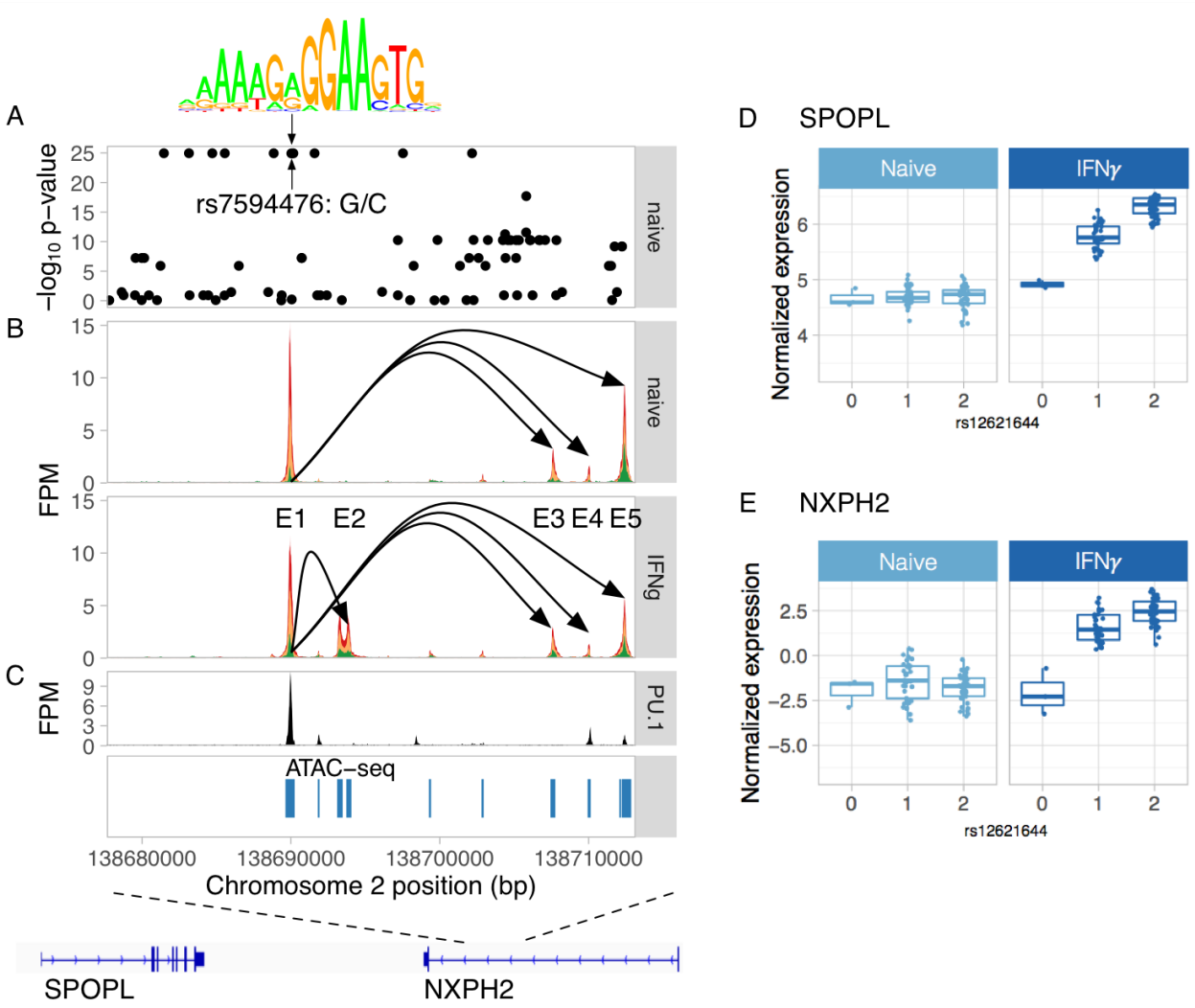

Figure 5.14: Example of a single QTL that influences chromatin accessibility at multiple peaks and the expression of two genes. **(A)** Manhattan plot of variants associated with the accessibility of the master caQTL peak E1. Only two of the associated variants overlap the E1 peak, and only rs7594476 is predicted to disrupt a PU.1 TF binding motif (M6119_1.02 in CIS-BP (Weirauch et al., 2014)). **(B)** Normalised ATAC-seq fragment coverage before and after IFNɣ stimulation stratified by the genotype at the rs7594476 SNP. Arrows correspond to links between the master peak E1 and dependent peaks E2-E5. **(C)** PU.1 ChIP-seq read coverage from (Schmidt et al., 2016). **(D)** Box plots of normalised SPOPL gene expression before and after IFNɣ stimulation. The boxplots are stratified by the genotype at rs7594476 SNP. **(E)** Box

plots of normalised SPOPL gene expression before and after IFNɣ stimulation. The boxplots are stratified by the genotype at rs7594476 SNP.

5.5.2 Using caQTLs to fine map causal variants for GWAS hits

In the previous section I showed that for ~70% of caQTLs at least one of the variants in the credible set overlapped the peak itself. This suggests that if there is an eQTL that is colocalised with a caQTL then the caQTL signal can be used to fine map causal variants for the eQTL.

PTK2B eQTL colocalises with a GWAS hit for Alzheimer's disease

Preliminary analysis with the NHGRI-EBI GWAS catalogue highlighted that lead eQTL SNP rs2322599 for PTK2B gene in the naive condition was in high LD (R^2 = 0.98) with rs28834970, a GWAS hit for Alzheimer's disease (Lambert et al., 2013). To see if both of these associations could be driven by the same causal variant, I downloaded Alzheimer's disease GWAS summary statistics from the International Genomics of Alzheimer's Project (IGAP) website (Lambert et al., 2013). I then used the coloc (Giambartolomei et al., 2014) software on a 250kb window around the GWAS lead SNP and found strong evidence of statistical colocalisation (posterior probability > 0.98). I also found that there was a caQTL in the same region that colocalised both with the GWAS hit as well as the eQTL (Figure 5.15A). Furthermore, the lead caQTL SNP rs28834970 was the only associated variant lying within the caQTL peak (Figure 5.15B), suggesting this is the most likely causal variant. The lead variant rs28834970 is T/C polymorphism and the alternative C allele is predicted to increase the relative binding score of the CEBPβ TF motif (M2268_1.02 in CIS-BP (Weirauch et al., 2014)) from 0.86 to 0.97 (Figure 5.15B). This is consistent with the increased chromatin accessibility at the C allele as well as increased expression of the PTK2B gene (Figure 5.15C). Furthermore, the variant also overlaps experimental CEBPβ ChIP-seq peak in primary human macrophages (Reschen et al., 2015) (Figure 5.15B). Together, this evidence suggests that rs28834970 is the likely causal variant for Alzheimer's disease risk that influences PTK2B expression by disrupting CEBPβ motif in an enhancer in the first intron of the gene. While the possible link between the rs28834970 Alzheimer's GWAS hit and PTK2B eQTL in monocytes has been highlighted before (Chan et al., 2015; Karch et al., 2016), we have been able to use statistical colocalisation together with caQTL data to pinpoint a single most likely causal variant and provide a plausible mechanism.

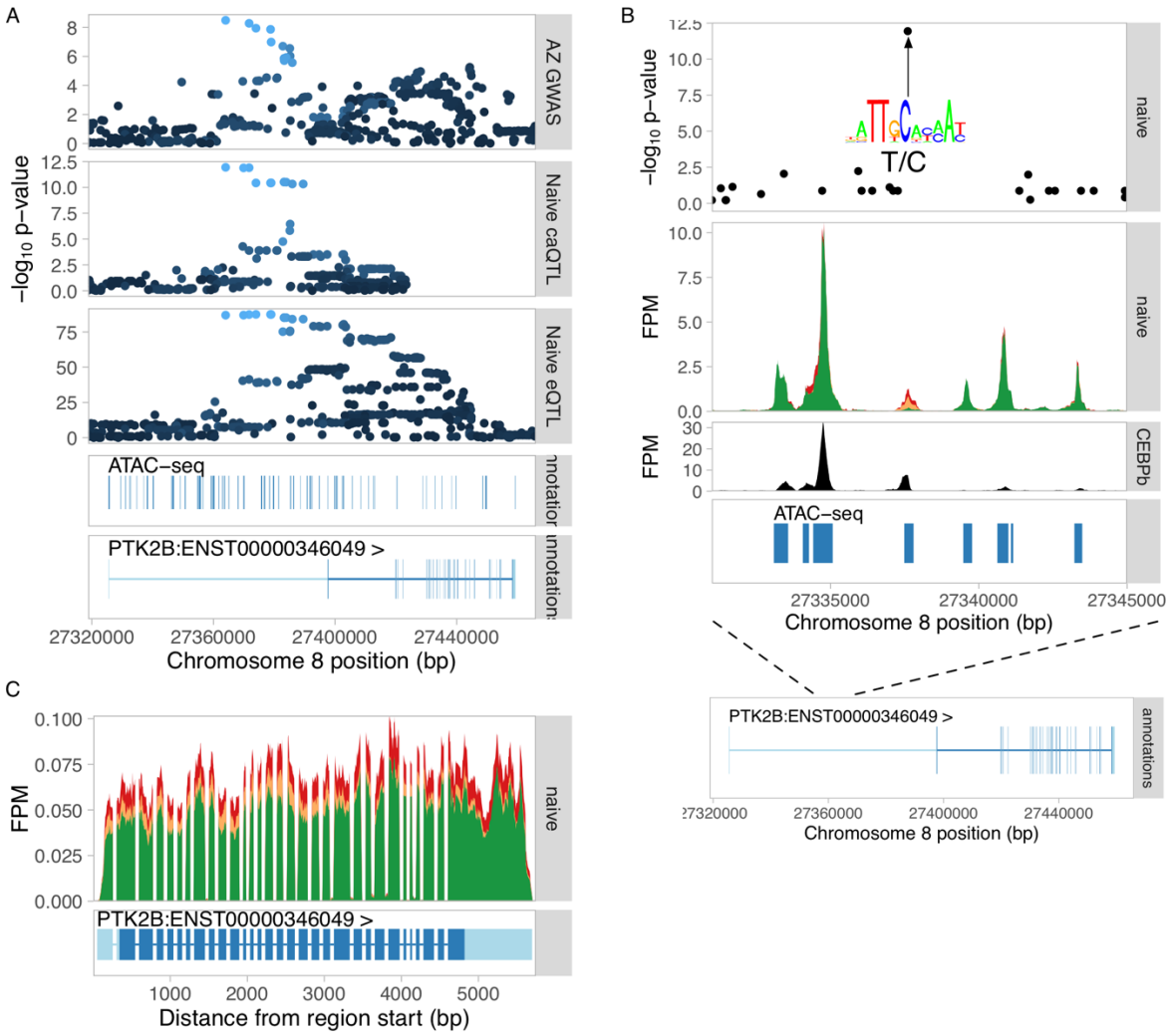

Figure 5.15: Dissecting the Alzheimer's disease causal variant at the PTK2B locus. (A) Manhattan plots for the Alzheimer's GWAS hit (top panel), colocalised caQTL (second panel) and colocalised eQTL for PTK2B gene (third panel). The bottom two tracks show all ATAC-seq peaks in the region as well was exons of the PTK2B gene. **(B)** ATAC-seq fragment coverage plot stratified by the rs28834970 genotype. **(C)** RNA-seq read coverage plot at the PTK2B gene stratified by the rs28834970 genotype.

5.6 Discussion

We have shown that, similarly to gene expression, (Chapters 2 and 4), the chromatin accessibility dynamics of IPSDMs also closely resemble that of primary macrophages. Evidence

for this comes from the motif enrichment analysis where constitutive and condition-specific macrophage ATAC peaks were enriched for expected macrophage-specific TF motifs such as PU.1, AP-1, NF-κB, STAT1 and ISRE motif representing multiple IRF factors. Secondly, overlap analysis with multiple public ChIP-seq datasets confirmed that overlapping regions changed their activity in IFNɣ and LPS response. Future studies where IPSDMs and MDMs are measured in the same experiment are needed to reliable detect any differential chromatin accessibility between the two cell types and identify TFs responsible for those differences.

Despite our modest sample size of 31-42 individuals, we identified thousands of caQTLs in each of the four conditions. We found that caQTL lead variants were 20% more likely to be shared between conditions than eQTL lead variants. This observation was further supported by the fact that for approximately 50% of the eQTLs that appeared after stimulation, the corresponding caQTL was already present in the naive state. Altogether, these observations suggest that a large fraction of genetic variation influences “primed” regulatory elements that wait for an appropriate environmental signal before regulating gene expression. Importantly, observing that a caQTL appears before eQTL allows us to infer that the caQTL is likely to be causal for the eQTL and not vice versa.

Multiple studies have shown that GWAS hits are enriched in gene regulatory regions that are often cell type specific (Maurano et al., 2012). Despite this observation, attempts to colocalise GWAS hits with specific eQTLs have had only limited success (Chun et al., 2016; Guo et al., 2015; Zhu et al., 2016). Chun *et al* (Chun et al., 2016) propose that regulatory regions might be accessible in multiple cell types and conditions (because they are bound by lineage determining pioneer TFs), but they might regulate gene expression in a few specific conditions. Importantly, this is consistent with our observation that caQTLs are less condition specific than eQTLs and for ~50% of condition-specific eQTLs their effect can be seen on chromatin level already before stimulation. Some evidence for the importance of cell-type specific pioneer TFs in disease comes from type 2 diabetes (T2D), where liver-specific pioneer TF FoxA2 (Iwafuchi-Doi et al., 2016) binding sites are enriched among fine-mapped T2D GWAS loci (Gaulton et al., 2015).

Similarly to previous studies (Grubert et al., 2015; Kumasaka et al., 2016; Waszak et al., 2015), we also found widespread evidence of single caQTL variants regulating the accessibility of multiple dependent caQTL peaks, often multiple kb away from the master peak. In total, we were able to detect at least one dependent caQTL peak for ~10% of the master caQTL peaks,

although this number is likely to increase with larger sample sizes. Importantly, measuring chromatin accessibility in multiple conditions allowed us to also identified a small number of dependent peaks that appeared or disappeared with stimulation. A number of those occurred in the SPOPL-NXPH2 locus (Figure 5.14), where the appearance of dependent caQTL peaks correlated with lead variant also becoming an eQTL for the two genes. This is consistent with a recently established model of hierarchical enhancer activation where signal-dependent transcription factors bind at or near primed enhancers to activate gene expression (Heinz et al., 2013; Romanoski et al., 2015).

Finally, the fact the caQTL variants are enriched within the peak whose accessibility they regulate allowed us to identify a small set of likely causal variants for thousands of caQTL peaks. By combining caQTLs with colocalised eQTLs and GWAS hits this can also facilitate fine mapping causal variants for those associations as illustrated by the SPOPL-NXPH2 (Figure 5.14) and PTK2B Alzheimer's GWAS hit (Figure 5.15) examples.

In summary, we have shown that mapping caQTLs in multiple conditions can provide insights into the principles of gene regulation and identify causal variants for eQTLs and GWAS hits. Larger sample sizes in multiple tissues and conditions together with methodological developments can undercover the true extent of dynamics between master and dependent peaks within multi-peak caQTLs.

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.

7 References

Adati, N., Huang, M.-C., Suzuki, T., Suzuki, H., and Kojima, T. (2009). High-resolution analysis of aberrant regions in autosomal chromosomes in human leukemia THP-1 cell line. BMC Res. Notes *2*, 153.

Alasoo, K., Martinez, F.O., Hale, C., Gordon, S., Powrie, F., Dougan, G., Mukhopadhyay, S., and Gaffney, D.J. (2015). Transcriptional profiling of macrophages derived from monocytes and iPS cells identifies a conserved response to LPS and novel alternative transcription. Sci. Rep. *5*, 12524.

Alipanahi, B., Delong, A., Weirauch, M.T., and Frey, B.J. (2015). Predicting the sequence specificities of DNA- and RNA-binding proteins by deep learning. Nat. Biotechnol. *33*, 831–838.

Anders, S., Reyes, A., and Huber, W. (2012). Detecting differential usage of exons from RNA-seq data. Genome Res. *22*, 2008–2017.

Anders, S., Pyl, P.T., and Huber, W. (2015). HTSeq--a Python framework to work with high-throughput sequencing data. Bioinformatics *31*, 166–169.

Aziz, A., Soucie, E., Sarrazin, S., and Sieweke, M.H. (2009). MafB/c-Maf deficiency enables self-renewal of differentiated functional macrophages. Science *326*, 867–871.

Barash, Y., Calarco, J.A., Gao, W., Pan, Q., Wang, X., Shai, O., Blencowe, B.J., and Frey, B.J. (2010). Deciphering the splicing code. Nature *465*, 53–59.

Barreiro, L.B., Tailleux, L., Pai, A.A., Gicquel, B., Marioni, J.C., and Gilad, Y. (2012). Deciphering the genetic architecture of variation in the immune response to Mycobacterium tuberculosis infection. Proc. Natl. Acad. Sci. U. S. A. *109*, 1204–1209.

Bates, D., Mächler, M., Bolker, B., and Walker, S. (2015). Fitting Linear Mixed-Effects Models Using lme4. J. Stat. Softw. *67*, 1–48.

Battle, A., Mostafavi, S., Zhu, X., Potash, J.B., Weissman, M.M., McCormick, C., Haudenschild, C.D., Beckman, K.B., Shi, J., Mei, R., et al. (2014). Characterizing the genetic basis of transcriptome diversity through RNA-sequencing of 922 individuals. Genome Res. 24, 14–24.

Battle, A., Khan, Z., Wang, S.H., Mitrano, A., Ford, M.J., Pritchard, J.K., and Gilad, Y. (2015). Impact of regulatory variation from RNA to protein. Science 347, 664–667.

Bell, O., Tiwari, V.K., Thomä, N.H., and Schübeler, D. (2011). Determinants and dynamics of genome accessibility. Nat. Rev. Genet. *12*, 554–564.

Benjamini, Y., and Speed, T.P. (2012). Summarizing and correcting the GC content bias in high-throughput sequencing. Nucleic Acids Res. *40*, e72.

Bock, C., Kiskinis, E., Verstappen, G., Gu, H., Boulting, G., Smith, Z.D., Ziller, M., Croft, G.F., Amoroso, M.W., Oakley, D.H., et al. (2011). Reference Maps of human ES and iPS cell variation enable high-throughput characterization of pluripotent cell lines. Cell *144*, 439–452.

Bray, N.L., Pimentel, H., Melsted, P., and Pachter, L. (2016). Near-optimal probabilistic RNA-seq quantification. Nat. Biotechnol. *34*, 525–527.

Buenrostro, J.D., Giresi, P.G., Zaba, L.C., Chang, H.Y., and Greenleaf, W.J. (2013). Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position. Nat. Methods *10*, 1213–1218.

Buenrostro, J.D., Wu, B., Litzenburger, U.M., Ruff, D., Gonzales, M.L., Snyder, M.P., Chang, H.Y., and Greenleaf, W.J. (2015). Single-cell chromatin accessibility reveals principles of regulatory variation. Nature 523, 486–490.

Cailhier, J.F., Partolina, M., Vuthoori, S., Wu, S., Ko, K., Watson, S., Savill, J., Hughes, J., and Lang, R.A. (2005). Conditional macrophage ablation demonstrates that resident macrophages initiate acute peritoneal inflammation. The Journal of Immunology *174*, 2336–2342.

Çalışkan, M., Baker, S.W., Gilad, Y., and Ober, C. (2015). Host genetic variation influences gene expression response to rhinovirus infection. PLoS Genet. *11*, e1005111.

Carpenter, S., Ricci, E.P., Mercier, B.C., Moore, M.J., and Fitzgerald, K.A. (2014). Post-transcriptional regulation of gene expression in innate immunity. Nat. Rev. Immunol. *14*, 361–376.

Castel, S., Levy-Moonshine, A., Mohammadi, P., Banks, E., and Lappalainen, T. (2015). Tools and best practices for data processing in allelic expression analysis. Genome Biol. *16*, 195.

Chan, G., White, C.C., Winn, P.A., Cimpean, M., Replogle, J.M., Glick, L.R., Cuerdon, N.E.,

Ryan, K.J., Johnson, K.A., Schneider, J.A., et al. (2015). CD33 modulates TREM2: convergence of Alzheimer loci. Nat. Neurosci. *18*, 1556–1558.

Chick, J.M., Munger, S.C., Simecek, P., Huttlin, E.L., Choi, K., Gatti, D.M., Raghupathy, N., Svenson, K.L., Churchill, G.A., and Gygi, S.P. (2016). Defining the consequences of genetic variation on a proteome-wide scale. Nature 534, 500–505.

Chun, S., Casparino, A., Patsopoulos, N., Croteau-Chonka, D., Raby, B., De Jager, P., Sunyaev, S., and Cotsapas, C. (2016). Shared effect modeling reveals that a fraction of autoimmune disease associations are consistent with eQTLs in three immune cell types. bioRxiv 053165.

Claussnitzer, M., Dankel, S.N., Kim, K.-H., Quon, G., Meuleman, W., Haugen, C., Glunk, V., Sousa, I.S., Beaudry, J.L., Puviindran, V., et al. (2015). FTO Obesity Variant Circuitry and Adipocyte Browning in Humans. N. Engl. J. Med. 373, 895–907.

Corces, M.R., Buenrostro, J.D., Wu, B., Greenside, P.G., Chan, S.M., Koenig, J.L., Snyder, M.P., Pritchard, J.K., Kundaje, A., Greenleaf, W.J., et al. (2016). Lineage-specific and single-cell chromatin accessibility charts human hematopoiesis and leukemia evolution. Nat. Genet. *48*, 1193–1203.

Cummings, B.B., Marshall, J.L., Tukiainen, T., Lek, M., Donkervoort, S., Foley, A.R., Bolduc, V., Waddell, L.B., Sandaradura, S.A., O'Grady, G.L., et al. (2016). Improving genetic diagnosis in Mendelian disease with transcriptome sequencing. bioRxiv 074153.

Cunninghame Graham, D.S., Manku, H., Wagner, S., Reid, J., Timms, K., Gutin, A., Lanchbury, J.S., and Vyse, T.J. (2007). Association of IRF5 in UK SLE families identifies a variant involved in polyadenylation. Hum. Mol. Genet. *16*, 579–591.

Dabney, A., Storey, J.D., and Warnes, G.R. (2010). qvalue: Q-value estimation for false discovery rate control. R package version 2.6.0, http://github.com/jdstorey/qvalue.

Davis, J.R., Fresard, L., Knowles, D.A., Pala, M., Bustamante, C.D., Battle, A., and Montgomery, S.B. (2016). An Efficient Multiple-Testing Adjustment for eQTL Studies that Accounts for Linkage Disequilibrium between Variants. Am. J. Hum. Genet. *98*, 216–224.

Degner, J.F., Pai, A.A., Pique-Regi, R., Veyrieras, J.-B., Gaffney, D.J., Pickrell, J.K., De Leon, S., Michelini, K., Lewellen, N., Crawford, G.E., et al. (2012). DNase I sensitivity QTLs are a

major determinant of human expression variation. Nature *482*, 390–394.

Deplancke, B., Alpern, D., and Gardeux, V. (2016). The Genetics of Transcription Factor DNA Binding Variation. Cell *166*, 538–554.

Ding, Z., Ni, Y., Timmer, S.W., Lee, B.-K., Battenhouse, A., Louzada, S., Yang, F., Dunham, I., Crawford, G.E., Lieb, J.D., et al. (2014). Quantitative Genetics of CTCF Binding Reveal Local Sequence Effects and Different Modes of X-Chromosome Association. PLoS Genet. *10*, e1004798.

Dobin, A., Davis, C.A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., Batut, P., Chaisson, M., and Gingeras, T.R. (2013). STAR: ultrafast universal RNA-seq aligner. Bioinformatics *29*, 15–21.

Doyle, S., Vaidya, S., O'Connell, R., Dadgostar, H., Dempsey, P., Wu, T., Rao, G., Sun, R., Haberland, M., Modlin, R., et al. (2002). IRF3 mediates a TLR3/TLR4-specific antiviral gene program. Immunity *17*, 251–263.

Durinck, S., Moreau, Y., Kasprzyk, A., Davis, S., De Moor, B., Brazma, A., and Huber, W. (2005). BioMart and Bioconductor: a powerful link between biological databases and microarray data analysis. Bioinformatics *21*, 3439–3440.

Ellis, S.E., Gupta, S., Ashar, F.N., Bader, J.S., West, A.B., and Arking, D.E. (2013). RNA-Seq optimization with eQTL gold standards. BMC Genomics *14*, 892.

Fairfax, B.P., Makino, S., Radhakrishnan, J., Plant, K., Leslie, S., Dilthey, A., Ellis, P., Langford, C., Vannberg, F.O., and Knight, J.C. (2012). Genetics of gene expression in primary immune cells identifies cell type-specific master regulators and roles of HLA alleles. Nat. Genet. *44*, 502–510.

Fairfax, B.P., Humburg, P., Makino, S., Naranbhai, V., Wong, D., Lau, E., Jostins, L., Plant, K., Andrews, R., McGee, C., et al. (2014). Innate immune activity conditions the effect of regulatory variants upon monocyte gene expression. Science *343*, 1246949.

Farh, K.K.-H., Marson, A., Zhu, J., Kleinewietfeld, M., Housley, W.J., Beik, S., Shoresh, N., Whitton, H., Ryan, R.J.H., Shishkin, A.A., et al. (2015). Genetic and epigenetic fine mapping of causal autoimmune disease variants. Nature 518, 337–343.

Ference, B.A., Robinson, J.G., Brook, R.D., Catapano, A.L., Chapman, M.J., Neff, D.R., Voros, S., Giugliano, R.P., Davey Smith, G., Fazio, S., et al. (2016). Variation in PCSK9 and HMGCR and Risk of Cardiovascular Disease and Diabetes. N. Engl. J. Med. 375, 2144–2153.

Finak, G., Frelinger, J., Jiang, W., Newell, E.W., Ramey, J., Davis, M.M., Kalams, S.A., De Rosa, S.C., and Gottardo, R. (2014). OpenCyto: an open source infrastructure for scalable, robust, reproducible, and automated, end-to-end flow cytometry data analysis. PLoS Comput. Biol. *10*, e1003806.

Fraley, C., and Raftery, A.E. (1999). MCLUST: Software for model-based cluster analysis. J. Classification *16*, 297–306.

Friedman, A.D. (2007). Transcriptional control of granulocyte and monocyte development. Oncogene *26*, 6816–6828.

Fu, J., Keurentjes, J.J.B., Bouwmeester, H., America, T., Verstappen, F.W.A., Ward, J.L., Beale, M.H., de Vos, R.C.H., Dijkstra, M., Scheltema, R.A., et al. (2009). System-wide molecular evidence for phenotypic buffering in Arabidopsis. Nat. Genet. *41*, 166–167.

Furey, T.S. (2012). ChIP-seq and beyond: new and improved methodologies to detect and characterize protein-DNA interactions. Nat. Rev. Genet. *13*, 840–852.

Garalde, D.R., Snell, E.A., Jachimowicz, D., Heron, A.J., Bruce, M., Lloyd, J., Warland, A., Pantic, N., Admassu, T., Ciccone, J., et al. (2016). Highly parallel direct RNA sequencing on an array of nanopores. bioRxiv 068809.

Gaulton, K.J., Ferreira, T., Lee, Y., Raimondo, A., Mägi, R., Reschen, M.E., Mahajan, A., Locke, A., William Rayner, N., Robertson, N., et al. (2015). Genetic fine mapping and genomic annotation defines causal mechanisms at type 2 diabetes susceptibility loci. Nat. Genet. *47*, 1415–1425.

Gautier, E.L., Shay, T., Miller, J., Greter, M., Jakubzick, C., Ivanov, S., Helft, J., Chow, A., Elpek, K.G., Gordonov, S., et al. (2012). Gene-expression profiles and transcriptional regulatory pathways that underlie the identity and diversity of mouse tissue macrophages. Nat. Immunol. *13*, 1118–1128.

van de Geijn, B., McVicker, G., Gilad, Y., and Pritchard, J.K. (2015). WASP: allele-specific software for robust molecular quantitative trait locus discovery. Nat. Methods 12, 1061–1063.

Ghazalpour, A., Bennett, B., Petyuk, V.A., Orozco, L., Hagopian, R., Mungrue, I.N., Farber, C.R., Sinsheimer, J., Kang, H.M., Furlotte, N., et al. (2011). Comparative analysis of proteome and transcriptome variation in mouse. PLoS Genet. *7*, e1001393.

Giambartolomei, C., Vukcevic, D., Schadt, E.E., Franke, L., Hingorani, A.D., Wallace, C., and Plagnol, V. (2014). Bayesian Test for Colocalisation between Pairs of Genetic Association Studies Using Summary Statistics. PLoS Genet. *10*, e1004383.

Ginhoux, F., Greter, M., Leboeuf, M., Nandi, S., See, P., Gokhan, S., Mehler, M.F., Conway, S.J., Ng, L.G., Stanley, E.R., et al. (2010). Fate mapping analysis reveals that adult microglia derive from primitive macrophages. Science *330*, 841–845.

Glaus, P., Honkela, A., and Rattray, M. (2012). Identifying differentially expressed transcripts from RNA-seq data with biological variation. Bioinformatics *28*, 1721–1728.

Gonzàlez-Porta, M., Frankish, A., Rung, J., Harrow, J., and Brazma, A. (2013). Transcriptome analysis of human tissues and cell lines reveals one dominant transcript per gene. Genome Biol. *14*, R70.

Gosselin, D., Link, V.M., Romanoski, C.E., Fonseca, G.J., Eichenfield, D.Z., Spann, N.J., Stender, J.D., Chun, H.B., Garner, H., Geissmann, F., et al. (2014). Environment drives selection and function of enhancers controlling tissue-specific macrophage identities. Cell *159*, 1327–1340.

Grant, C.E., Bailey, T.L., and Noble, W.S. (2011). FIMO: scanning for occurrences of a given motif. Bioinformatics *27*, 1017–1018.

Grubert, F., Zaugg, J.B., Kasowski, M., Ursu, O., Spacek, D.V., Martin, A.R., Greenside, P., Srivas, R., Phanstiel, D.H., Pekowska, A., et al. (2015). Genetic Control of Chromatin States in Humans Involves Local and Distal Chromosomal Interactions. Cell *162*, 1051–1065.

Gundra, U.M., Girgis, N.M., Ruckerl, D., Jenkins, S., Ward, L.N., Kurtz, Z.D., Wiens, K.E., Tang, M.S., Basu-Roy, U., Mansukhani, A., et al. (2014). Alternatively activated macrophages derived from monocytes and tissue macrophages are phenotypically and functionally distinct. Blood *123*, e110–e122.

Guo, H., Fortune, M.D., Burren, O.S., Schofield, E., Todd, J.A., and Wallace, C. (2015). Integration of disease association and eQTL data using a Bayesian colocalisation approach

highlights six candidate causal genes in immune-mediated diseases. Hum. Mol. Genet. *24*, 3305–3313.

Gupta, I., Clauder-Münster, S., Klaus, B., Järvelin, A.I., Aiyar, R.S., Benes, V., Wilkening, S., Huber, W., Pelechano, V., and Steinmetz, L.M. (2014). Alternative polyadenylation diversifies post-transcriptional regulation by selective RNA–protein interactions. Mol. Syst. Biol. *10*.

Hansen, K.D., Irizarry, R.A., and Wu, Z. (2012). Removing technical variability in RNA-seq data using conditional quantile normalization. Biostatistics *13*, 204–216.

Haraga, A., Ohlson, M.B., and Miller, S.I. (2008). Salmonellae interplay with host cells. Nat. Rev. Microbiol. *6*, 53–66.

Harrow, J., Frankish, A., Gonzalez, J.M., Tapanari, E., Diekhans, M., Kokocinski, F., Aken, B.L., Barrell, D., Zadissa, A., Searle, S., et al. (2012). GENCODE: the reference human genome annotation for The ENCODE Project. Genome Res. *22*, 1760–1774.

Heger, A., Webber, C., Goodson, M., Ponting, C.P., and Lunter, G. (2013). GAT: a simulation framework for testing the association of genomic intervals. Bioinformatics *29*, 2046–2048.

Heinz, S., Benner, C., Spann, N., Bertolino, E., Lin, Y.C., Laslo, P., Cheng, J.X., Murre, C., Singh, H., and Glass, C.K. (2010). Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities. Mol. Cell *38*, 576–589.

Heinz, S., Romanoski, C.E., Benner, C., Allison, K.A., Kaikkonen, M.U., Orozco, L.D., and Glass, C.K. (2013). Effect of natural genetic variation on enhancer selection and function. Nature *503*, 487–492.

Henikoff, S., and Shilatifard, A. (2011). Histone modification: cause or cog? Trends Genet. *27*, 389–396.

Herzenberg, L.A., Tung, J., Moore, W.A., Herzenberg, L.A., and Parks, D.R. (2006). Interpreting flow cytometry data: a guide for the perplexed. Nat. Immunol. *7*, 681–685.

't Hoen, P.A.C., Friedländer, M.R., Almlöf, J., Sammeth, M., Pulyakhina, I., Anvar, S.Y., Laros, J.F.J., Buermans, H.P.J., Karlberg, O., Brännvall, M., et al. (2013). Reproducibility of high-throughput mRNA and small RNA sequencing across laboratories. Nat. Biotechnol. 31, 1015–

1022.

Hormozdiari, F., Segre, A.V., van de Bunt, M., Li, X., Joo, J.W.J., Bilow, M., Sul, J.H., Sankararaman, S., Pasaniuc, B., and Eskin, E. (2016). Joint Fine Mapping of GWAS and eQTL Detects Target Gene and Relevant Tissue. bioRxiv 065037.

Hu, B.-Y., Weick, J.P., Yu, J., Ma, L.-X., Zhang, X.-Q., Thomson, J.A., and Zhang, S.-C. (2010). Neural differentiation of human induced pluripotent stem cells follows developmental principles but with variable potency. Proc. Natl. Acad. Sci. U. S. A. *107*, 4335–4340.

Hu, X., and Ivashkiv, L.B. (2009). Cross-regulation of signaling pathways by interferon-gamma: implications for immune responses and autoimmune diseases. Immunity 31, 539–550.

Ivashkiv, L.B., and Donlin, L.T. (2014). Regulation of type I interferon responses. Nat. Rev. Immunol. *14*, 36–49.

Iwafuchi-Doi, M., Donahue, G., Kakumanu, A., Watts, J.A., Mahony, S., Pugh, B.F., Lee, D., Kaestner, K.H., and Zaret, K.S. (2016). The Pioneer Transcription Factor FoxA Maintains an Accessible Nucleosome Configuration at Enhancers for Tissue-Specific Gene Activation. Mol. Cell *62*, 79–91.

Jacob, F., and Monod, J. (1961). Genetic regulatory mechanisms in the synthesis of proteins. J. Mol. Biol. *3*, 318–356.

Jakubzick, C., Gautier, E.L., Gibbings, S.L., Sojka, D.K., Schlitzer, A., Johnson, T.E., Ivanov, S., Duan, Q., Bala, S., Condon, T., et al. (2013). Minimal differentiation of classical monocytes as they survey steady-state tissues and transport antigen to lymph nodes. Immunity *39*, 599–610.

Jiang, H., Lei, R., Ding, S.-W., and Zhu, S. (2014). Skewer: a fast and accurate adapter trimmer for next-generation sequencing paired-end reads. BMC Bioinformatics *15*, 182.

Jiang, Y., Cowley, S.A., Siler, U., Melguizo, D., Tilgner, K., Browne, C., Dewilton, A., Przyborski, S., Saretzki, G., James, W.S., et al. (2012). Derivation and functional analysis of patient-specific induced pluripotent stem cells as an in vitro model of chronic granulomatous disease. Stem Cells *30*, 599–611.

Johansson, M., Bocher, V., Lehto, M., Chinetti, G., Kuismanen, E., Ehnholm, C., Staels, B., and Olkkonen, V.M. (2003). The two variants of oxysterol binding protein-related protein-1 display

different tissue expression patterns, have different intracellular localization, and are functionally distinct. Mol. Biol. Cell *14*, 903–915.

de Jong, H.K., Parry, C.M., van der Poll, T., and Wiersinga, W.J. (2012). Host-pathogen interaction in invasive Salmonellosis. PLoS Pathog. *8*, e1002933.

Jun, G., Flickinger, M., Hetrick, K.N., Romm, J.M., Doheny, K.F., Abecasis, G.R., Boehnke, M., and Kang, H.M. (2012). Detecting and estimating contamination of human DNA samples in sequencing and array-based genotype data. Am. J. Hum. Genet. *91*, 839–848.

Kaikkonen, M.U., Spann, N.J., Heinz, S., Romanoski, C.E., Allison, K.A., Stender, J.D., Chun, H.B., Tough, D.F., Prinjha, R.K., Benner, C., et al. (2013). Remodeling of the Enhancer Landscape during Macrophage Activation Is Coupled to Enhancer Transcription. Mol. Cell *51*, 310–325.

Kajiwara, M., Aoi, T., Okita, K., Takahashi, R., Inoue, H., Takayama, N., Endo, H., Eto, K., Toguchida, J., Uemoto, S., et al. (2012). Donor-dependent variations in hepatic differentiation from human-induced pluripotent stem cells. Proc. Natl. Acad. Sci. U. S. A. *109*, 12538–12543.

Kanitz, A., Gypas, F., Gruber, A.J., Gruber, A.R., Martin, G., and Zavolan, M. (2015). Comparative assessment of methods for the computational inference of transcript isoform abundance from RNA-seq data. Genome Biol. *16*, 150.

Karch, C.M., Ezerskiy, L.A., Bertelsen, S., Alzheimer's Disease Genetics Consortium (ADGC), and Goate, A.M. (2016). Alzheimer's Disease Risk Polymorphisms Regulate Gene Expression in the ZCWPW1 and the CELF1 Loci. PLoS One *11*, e0148717.

Karlsson, K.R., Cowley, S., Martinez, F.O., Shaw, M., Minger, S.L., and James, W. (2008). Homogeneous monocytes and macrophages from human embryonic stem cells following coculture-free differentiation in M-CSF and IL-3. Exp. Hematol. *36*, 1167–1175.

Kasowski, M., Grubert, F., Heffelfinger, C., Hariharan, M., Asabere, A., Waszak, S.M., Habegger, L., Rozowsky, J., Shi, M., Urban, A.E., et al. (2010). Variation in transcription factor binding among humans. Science *328*, 232–235.

Kasowski, M., Kyriazopoulou-Panagiotopoulou, S., Grubert, F., Zaugg, J.B., Kundaje, A., Liu, Y., Boyle, A.P., Zhang, Q.C., Zakharia, F., Spacek, D.V., et al. (2013). Extensive variation in chromatin states across humans. Science *342*, 750–752.

Katz, Y., Wang, E.T., Airoldi, E.M., and Burge, C.B. (2010). Analysis and design of RNA sequencing experiments for identifying isoform regulation. Nat. Methods *7*, 1009–1015.

Kelley, D.R., Snoek, J., and Rinn, J.L. (2016). Basset: learning the regulatory code of the accessible genome with deep convolutional neural networks. Genome Res. *26*, 990–999.

Kempf, H., Kropp, C., Olmer, R., Martin, U., and Zweigerdt, R. (2015). Cardiac differentiation of human pluripotent stem cells in scalable suspension culture. Nat. Protoc. *10*, 1345–1361.

Keren, L., Hausser, J., Lotan-Pompan, M., Vainberg Slutskin, I., Alisar, H., Kaminski, S., Weinberger, A., Alon, U., Milo, R., and Segal, E. (2016). Massively Parallel Interrogation of the Effects of Gene Expression Levels on Fitness. Cell 166, 1282–1294.e18.

Kilpinen, H., Waszak, S.M., Gschwind, A.R., Raghav, S.K., Witwicki, R.M., Orioli, A., Migliavacca, E., Wiederkehr, M., Gutierrez-Arcelus, M., Panousis, N.I., et al. (2013). Coordinated effects of sequence variation on DNA binding, chromatin structure, and transcription. Science *342*, 744–747.

Kilpinen, H., Goncalves, A., Leha, A., Afzal, V., Ashford, S., Bala, S., Bensaddek, D., Casale, F.P., Culley, O., Danacek, P., et al. (2016). Common genetic variation drives molecular heterogeneity in human iPSCs. bioRxiv 055160.

Kim, D., Pertea, G., Trapnell, C., Pimentel, H., Kelley, R., and Salzberg, S.L. (2013). TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. Genome Biol. *14*, R36.

Kim, D., Langmead, B., and Salzberg, S.L. (2015). HISAT: a fast spliced aligner with low memory requirements. Nat. Methods 12, 357–360.

Kim, S., Becker, J., Bechheim, M., Kaiser, V., Noursadeghi, M., Fricker, N., Beier, E., Klaschik, S., Boor, P., Hess, T., et al. (2014). Characterizing the genetic basis of innate immune response in TLR4-activated human monocytes. Nat. Commun. *5*, 5236.

Klimchenko, O., Di Stefano, A., Geoerger, B., Hamidi, S., Opolon, P., Robert, T., Routhier, M., El-Benna, J., Delezoide, A.-L., Boukour, S., et al. (2011). Monocytic cells derived from human embryonic stem cells and fetal liver share common differentiation pathways and homeostatic functions. Blood *117*, 3065–3075.

Kottyan, L.C., Zoller, E.E., Bene, J., Lu, X., Kelly, J.A., Rupert, A.M., Lessard, C.J., Vaughn, S.E., Marion, M., Weirauch, M.T., et al. (2015). The IRF5-TNPO3 association with systemic lupus erythematosus has two components that other autoimmune disorders variably share. Hum. Mol. Genet. *24*, 582–596.

Koyanagi-Aoi, M., Ohnuki, M., Takahashi, K., Okita, K., Noma, H., Sawamura, Y., Teramoto, I., Narita, M., Sato, Y., Ichisaka, T., et al. (2013). Differentiation-defective phenotypes revealed by large-scale analyses of human pluripotent stem cells. Proc. Natl. Acad. Sci. U. S. A. *110*, 20569–20574.

Krause, P., Morris, V., Greenbaum, J.A., Park, Y., Bjoerheden, U., Mikulski, Z., Muffley, T., Shui, J.-W., Kim, G., Cheroutre, H., et al. (2015). IL-10-producing intestinal macrophages prevent excessive antibacterial innate immunity by limiting IL-23 synthesis. Nat. Commun. *6*, 7055.

Kumar, L., and E Futschik, M. (2007). Mfuzz: a software package for soft clustering of microarray data. Bioinformation *2*, 5–7.

Kumasaka, N., Knights, A.J., and Gaffney, D.J. (2016). Fine-mapping cellular QTLs with RASQUAL and ATAC-seq. Nat. Genet. 48, 206–213.

Lachmann, N., Ackermann, M., Frenzel, E., Liebhaber, S., Brennig, S., Happle, C., Hoffmann, D., Klimenkova, O., Lüttge, D., Buchegger, T., et al. (2015). Large-Scale Hematopoietic Differentiation of Human Induced Pluripotent Stem Cells Provides Granulocytes or Macrophages for Cell Replacement Therapies. Nat. Rep. Stem Cells *4*, 282–296.

Lahens, N.F., Kavakli, I.H., Zhang, R., Hayer, K., Black, M.B., Dueck, H., Pizarro, A., Kim, J., Irizarry, R., Thomas, R.S., et al. (2014). IVT-seq reveals extreme bias in RNA sequencing. Genome Biol. *15*, R86.

Lambert, J.C., Ibrahim-Verbaas, C.A., Harold, D., Naj, A.C., Sims, R., Bellenguez, C., DeStafano, A.L., Bis, J.C., Beecham, G.W., Grenier-Boley, B., et al. (2013). Meta-analysis of 74,046 individuals identifies 11 new susceptibility loci for Alzheimer's disease. Nat. Genet. *45*, 1452–1458.

Lambert, N., Robertson, A., Jangi, M., McGeary, S., Sharp, P.A., and Burge, C.B. (2014). RNA Bind-n-Seq: quantitative assessment of the sequence and structural binding specificity of RNA

binding proteins. Mol. Cell *54*, 887–900.

Langmead, B., Trapnell, C., Pop, M., and Salzberg, S.L. (2009). Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biol. *10*, R25.

Lappalainen, T., Sammeth, M., Friedländer, M.R., ’t Hoen, P.A.C., Monlong, J., Rivas, M.A., Gonzàlez-Porta, M., Kurbatova, N., Griebel, T., Ferreira, P.G., et al. (2013). Transcriptome and genome sequencing uncovers functional variation in humans. Nature *501*, 506–511.

Lara-Astiaso, D., Weiner, A., Lorenzo-Vivas, E., Zaretsky, I., Jaitin, D.A., David, E., Keren-Shaul, H., Mildner, A., Winter, D., Jung, S., et al. (2014). Immunogenetics. Chromatin state dynamics during blood formation. Science *345*, 943–949.

Lavin, Y., Winter, D., Blecher-Gonen, R., David, E., Keren-Shaul, H., Merad, M., Jung, S., and Amit, I. (2014). Tissue-resident macrophage enhancer landscapes are shaped by the local microenvironment. Cell *159*, 1312–1326.

Law, C.W., Chen, Y., Shi, W., and Smyth, G.K. (2014). voom: Precision weights unlock linear model analysis tools for RNA-seq read counts. Genome Biol. *15*, R29.

Lawrence, M., Gentleman, R., and Carey, V. (2009). rtracklayer: an R package for interfacing with genome browsers. Bioinformatics *25*, 1841–1842.

Lee, M.N., Ye, C., Villani, A.-C., Raj, T., Li, W., Eisenhaure, T.M., Imboywa, S.H., Chipendo, P.I., Ran, F.A., Slowikowski, K., et al. (2014). Common genetic variants modulate pathogen-sensing responses in human dendritic cells. Science *343*, 1246980.

Leek, J.T., and Storey, J.D. (2007). Capturing heterogeneity in gene expression studies by surrogate variable analysis. PLoS Genet. *3*, 1724–1735.

Li, H. (2013). Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. arXiv [q-bio.GN].

Li, Y., Oosting, M., Deelen, P., Ricaño-Ponce, I., Smeekens, S., Jaeger, M., Matzaraki, V., Swertz, M.A., Xavier, R.J., Franke, L., et al. (2016a). Inter-individual variability and genetic influences on cytokine responses to bacteria and fungi. Nat. Med. 22, 952–960.

Li, Y.I., Knowles, D.A., and Pritchard, J.K. (2016b). LeafCutter: Annotation-free quantification of RNA splicing. bioRxiv 044107.

Li, Y.I., van de Geijn, B., Raj, A., Knowles, D.A., Petti, A.A., Golan, D., Gilad, Y., and Pritchard, J.K. (2016c). RNA splicing is a primary link between genetic variation and disease. Science *352*, 600–604.

Liao, Y., Smyth, G.K., and Shi, W. (2014). featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. Bioinformatics *30*, 923–930.

Love, M.I., Huber, W., and Anders, S. (2014). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol. *15*, 550.

MacArthur, S., Li, X.-Y., Li, J., Brown, J.B., Chu, H.C., Zeng, L., Grondona, B.P., Hechmer, A., Simirenko, L., Keränen, S.V.E., et al. (2009). Developmental roles of 21 Drosophila transcription factors are determined by quantitative differences in binding to an overlapping set of thousands of genomic regions. Genome Biol. *10*, R80.

Mantovani, A., Biswas, S.K., Galdiero, M.R., Sica, A., and Locati, M. (2013). Macrophage plasticity and polarization in tissue repair and remodelling. J. Pathol. *229*, 176–185.

Maranville, J.C., Luca, F., Richards, A.L., Wen, X., Witonsky, D.B., Baxter, S., Stephens, M., and Di Rienzo, A. (2011). Interactions between glucocorticoid treatment and cis-regulatory polymorphisms contribute to cellular response phenotypes. PLoS Genet. *7*, e1002162.

Marbach, D., Lamparter, D., Quon, G., Kellis, M., Kutalik, Z., and Bergmann, S. (2016). Tissue-specific regulatory circuits reveal variable modular perturbations across complex diseases. Nat. Methods *13*, 366–370.

Marioni, J.C., Mason, C.E., Mane, S.M., Stephens, M., and Gilad, Y. (2008). RNA-seq: an assessment of technical reproducibility and comparison with gene expression arrays. Genome Res. *18*, 1509–1517.

Martinez, F.O. (2012). Analysis of gene expression and gene silencing in human macrophages. Curr. Protoc. Immunol. *Chapter 14*, Unit 14.28.1–23.

Martinez, F.O., Gordon, S., Locati, M., and Mantovani, A. (2006). Transcriptional profiling of the human monocyte-to-macrophage differentiation and polarization: new molecules and patterns of gene expression. The Journal of Immunology *177*, 7303–7311.

Maurano, M.T., Humbert, R., Rynes, E., Thurman, R.E., Haugen, E., Wang, H., Reynolds, A.P.,

Sandstrom, R., Qu, H., Brody, J., et al. (2012). Systematic localization of common disease-associated variation in regulatory DNA. Science *337*, 1190–1195.

Mayr, C., and Bartel, D.P. (2009). Widespread shortening of 3'UTRs by alternative cleavage and polyadenylation activates oncogenes in cancer cells. Cell *138*, 673–684.

McVicker, G., van de Geijn, B., Degner, J.F., Cain, C.E., Banovich, N.E., Raj, A., Lewellen, N., Myrthil, M., Gilad, Y., and Pritchard, J.K. (2013). Identification of Genetic Variants That Affect Histone Modifications in Human Cells. Science *342*, 747–749.

Medina, K.L., and Singh, H. (2005). Gene Regulatory Networks Orchestrating B Cell Fate Specification, Commitment, and Differentiation. In Molecular Analysis of B Lymphocyte Development and Activation, P.D.H. Singh, and P.D.R. Grosschedl, eds. (Springer Berlin Heidelberg), pp. 1–14.

Medzhitov, R., and Horng, T. (2009). Transcriptional control of the inflammatory response. Nat. Rev. Immunol. *9*, 692–703.

Miura, P., Shenker, S., Andreu-Agullo, C., Westholm, J.O., and Lai, E.C. (2013). Widespread and extensive lengthening of 3' UTRs in the mammalian brain. Genome Res. *23*, 812–825.

Mogensen, T.H. (2009). Pathogen recognition and inflammatory signaling in innate immune defenses. Clin. Microbiol. Rev. *22*, 240–273

Montgomery, S.B., Sammeth, M., Gutierrez-Arcelus, M., Lach, R.P., Ingle, C., Nisbett, J., Guigo, R., and Dermitzakis, E.T. (2010). Transcriptome genetics using second generation sequencing in a Caucasian population. Nature *464*, 773–777.

Mullen, A.C., Orlando, D.A., Newman, J.J., Lovén, J., Kumar, R.M., Bilodeau, S., Reddy, J., Guenther, M.G., DeKoter, R.P., and Young, R.A. (2011). Master transcription factors determine cell-type-specific responses to TGF-β signaling. Cell *147*, 565–576.

Munoz, D.M., Cassiani, P.J., Li, L., Billy, E., Korn, J.M., Jones, M.D., Golji, J., Ruddy, D.A., Yu, K., McAllister, G., et al. (2016). CRISPR Screens Provide a Comprehensive Assessment of Cancer Vulnerabilities but Generate False-Positive Hits for Highly Amplified Genomic Regions. Cancer Discov. 6, 900–913.

Murray, P.J., and Wynn, T.A. (2011). Protective and pathogenic functions of macrophage

subsets. Nat. Rev. Immunol. *11*, 723–737.

Muruve, D.A., Pétrilli, V., Zaiss, A.K., White, L.R., Clark, S.A., Ross, P.J., Parks, R.J., and Tschopp, J. (2008). The inflammasome recognizes cytosolic microbial and host DNA and triggers an innate immune response. Nature *452*, 103–107.

Naranbhai, V., Fairfax, B.P., Makino, S., Humburg, P., Wong, D., Ng, E., Hill, A.V.S., and Knight, J.C. (2015). Genomic modulators of gene expression in human neutrophils. Nat. Commun. *6*, 7545.

Nau, G.J., Richmond, J.F.L., Schlesinger, A., Jennings, E.G., Lander, E.S., and Young, R.A. (2002). Human macrophage activation programs induced by bacterial pathogens. Proc. Natl. Acad. Sci. U. S. A. *99*, 1503–1508.

Negishi, H., Fujita, Y., Yanai, H., Sakaguchi, S., Ouyang, X., Shinohara, M., Takayanagi, H., Ohba, Y., Taniguchi, T., and Honda, K. (2006). Evidence for licensing of IFN-gamma-induced IFN regulatory factor 1 transcription factor by MyD88 in Toll-like receptor-dependent gene induction program. Proc. Natl. Acad. Sci. U. S. A. *103*, 15136–15141.

Neph, S., Vierstra, J., Stergachis, A.B., Reynolds, A.P., Haugen, E., Vernot, B., Thurman, R.E., John, S., Sandstrom, R., Johnson, A.K., et al. (2012). An expansive human regulatory lexicon encoded in transcription factor footprints. Nature *489*, 83–90.

Ngo, M., and Ridgway, N.D. (2009). Oxysterol binding protein–related protein 9 (ORP9) is a cholesterol transfer protein that regulates Golgi structure and function. Mol. Biol. Cell *20*, 1388–1399.

Nica, A.C., Parts, L., Glass, D., Nisbet, J., Barrett, A., Sekowska, M., Travers, M., Potter, S., Grundberg, E., Small, K., et al. (2011). The architecture of gene regulatory variation across multiple human tissues: the MuTHER study. PLoS Genet. *7*, e1002003.

Odom, D.T., Zizlsperger, N., Benjamin Gordon, D., Bell, 1. George W., Rinaldi, N.J., Murray, H.L., Volkert, 1. Tom L., Schreiber, J., Alexander Rolfe, P., Gifford, D.K., et al. (2004). Control of Pancreas and Liver Gene Expression by HNF Transcription Factors. Science *303*.

O'Neill, L.A., Sheedy, F.J., and McCoy, C.E. (2011). MicroRNAs: the fine-tuners of Toll-like receptor signalling. Nat. Rev. Immunol. *11*, 163–175.

Ongen, H., Buil, A., Brown, A.A., Dermitzakis, E.T., and Delaneau, O. (2016). Fast and efficient QTL mapper for thousands of molecular phenotypes. Bioinformatics 32, 1479–1485.

Orozco, L.D., Bennett, B.J., Farber, C.R., Ghazalpour, A., Pan, C., Che, N., Wen, P., Qi, H.X., Mutukulu, A., Siemers, N., et al. (2012). Unraveling inflammatory responses using systems genetics and gene-environment interactions in macrophages. Cell *151*, 658–670.

Ostuni, R., Piccolo, V., Barozzi, I., Polletti, S., Termanini, A., Bonifacio, S., Curina, A., Prosperini, E., Ghisletti, S., and Natoli, G. (2013). Latent enhancers activated by stimulation in differentiated cells. Cell *152*, 157–171.

Parnas, O., Jovanovic, M., Eisenhaure, T.M., Herbst, R.H., Dixit, A., Ye, C.J., Przybylski, D., Platt, R.J., Tirosh, I., Sanjana, N.E., et al. (2015). A Genome-wide CRISPR Screen in Primary Immune Cells to Dissect Regulatory Networks. Cell 162, 675–686.

Parts, L., Stegle, O., Winn, J., and Durbin, R. (2011). Joint genetic analysis of gene expression data with inferred cellular phenotypes. PLoS Genet. *7*, e1001276.

Patro, R., Mount, S.M., and Kingsford, C. (2014). Sailfish enables alignment-free isoform quantification from RNA-seq reads using lightweight algorithms. Nat. Biotechnol. *32*, 462–464.

Patro, R., Duggal, G., Love, M.I., Irizarry, R.A., and Kingsford, C. (2016). Salmon provides accurate, fast, and bias-aware transcript expression estimates using dual-phase inference. bioRxiv 021592.

Paull, D., Sevilla, A., Zhou, H., Hahn, A.K., Kim, H., Napolitano, C., Tsankov, A., Shang, L., Krumholz, K., Jagadeesan, P., et al. (2015). Automated, high-throughput derivation, characterization and differentiation of induced pluripotent stem cells. Nat. Methods *12*, 885–892.

Pham, T.-H., Benner, C., Lichtinger, M., Schwarzfischer, L., Hu, Y., Andreesen, R., Chen, W., and Rehli, M. (2012). Dynamic epigenetic enhancer signatures reveal key transcription factors associated with monocytic differentiation states. Blood *119*, e161–e171.

Pham, T.-H., Minderjahn, J., Schmidl, C., Hoffmeister, H., Schmidhofer, S., Chen, W., Längst, G., Benner, C., and Rehli, M. (2013). Mechanisms of in vivo binding site selection of the hematopoietic master transcription factor PU.1. Nucleic Acids Res. *41*, 6391–6402.

Pickrell, J.K., Marioni, J.C., Pai, A.A., Degner, J.F., Engelhardt, B.E., Nkadori, E., Veyrieras, J.-

B., Stephens, M., Gilad, Y., and Pritchard, J.K. (2010). Understanding mechanisms underlying human gene expression variation with RNA sequencing. Nature *464*, 768–772.

Picelli, S., Faridani, O.R., Björklund, A.K., Winberg, G., Sagasser, S., and Sandberg, R. (2014). Full-length RNA-seq from single cells using Smart-seq2. Nat. Protoc. *9*, 171–181.

Platanias, L.C. (2005). Mechanisms of type-I- and type-II-interferon-mediated signalling. Nat. Rev. Immunol. *5*, 375–386.

Polach, K.J., and Widom, J. (1996). A model for the cooperative binding of eukaryotic regulatory proteins to nucleosomal target sites. J. Mol. Biol. *258*, 800–812.

Polo, J.M., Liu, S., Figueroa, M.E., Kulalert, W., Eminli, S., Tan, K.Y., Apostolou, E., Stadtfeld, M., Li, Y., Shioda, T., et al. (2010). Cell type of origin influences the molecular and functional properties of mouse induced pluripotent stem cells. Nat. Biotechnol. *28*, 848–855.

Qiao, Y., Giannopoulou, E.G., Chan, C.H., Park, S.-H., Gong, S., Chen, J., Hu, X., Elemento, O., and Ivashkiv, L.B. (2013). Synergistic activation of inflammatory cytokine genes by interferon-γ-induced chromatin remodeling and toll-like receptor signaling. Immunity *39*, 454–469.

Quinlan, A.R., and Hall, I.M. (2010). BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics *26*, 841–842.

Ramsauer, K., Farlik, M., Zupkovitz, G., Seiser, C., Kröger, A., Hauser, H., and Decker, T. (2007). Distinct modes of action applied by transcription factors STAT1 and IRF1 to initiate transcription of the IFN-gamma-inducible gbp2 gene. Proc. Natl. Acad. Sci. U. S. A. *104*, 2849–2854.

Reimand, J., Arak, T., and Vilo, J. (2011). g:Profiler--a web server for functional interpretation of gene lists (2011 update). Nucleic Acids Res. *39*, W307–W315.

Reimand, J., Arak, T., Adler, P., Kolberg, L., Reisberg, S., Peterson, H., and Vilo, J. (2016). g:Profiler—a web server for functional interpretation of gene lists (2016 update). Nucleic Acids Res. *44*, W83–W89.

Reith, W., LeibundGut-Landmann, S., and Waldburger, J.-M. (2005). Regulation of MHC class II gene expression by the class II transactivator. Nat. Rev. Immunol. *5*, 793–806.

Reschen, M.E., Gaulton, K.J., Lin, D., Soilleux, E.J., Morris, A.J., Smyth, S.S., and O'Callaghan, C.A. (2015). Lipid-induced epigenomic changes in human macrophages identify a coronary artery disease-associated variant that regulates PPAP2B Expression through Altered C/EBP-beta binding. PLoS Genet. *11*, e1005061.

Rigamonti, A., Repetti, G.G., Sun, C., Price, F.D., Reny, D.C., Rapino, F., Weisinger, K., Benkler, C., Peterson, Q.P., Davidow, L.S., et al. (2016). Large-Scale Production of Mature Neurons from Human Pluripotent Stem Cells in a Three-Dimensional Suspension Culture System. Stem Cell Reports *6*, 993–1008.

Roadmap Epigenomics Consortium, Kundaje, A., Meuleman, W., Ernst, J., Bilenky, M., Yen, A., Heravi-Moussavi, A., Kheradpour, P., Zhang, Z., Wang, J., et al. (2015). Integrative analysis of 111 reference human epigenomes. Nature *518*, 317–330.

Roberts, A., Trapnell, C., Donaghey, J., Rinn, J.L., and Pachter, L. (2011a). Improving RNA-Seq expression estimates by correcting for fragment bias. Genome Biol. *12*, R22.

Roberts, A., Pimentel, H., Trapnell, C., and Pachter, L. (2011b). Identification of novel transcripts in annotated genomes using RNA-Seq. Bioinformatics *27*, 2325–2329.

Romanoski, C.E., Link, V.M., Heinz, S., and Glass, C.K. (2015). Exploiting genomics and natural genetic variation to decode macrophage enhancers. Trends Immunol. *36*, 507–518.

Rosenberger, C.M., Scott, M.G., Gold, M.R., Hancock, R.E., and Finlay, B.B. (2000). Salmonella typhimurium infection and lipopolysaccharide stimulation induce similar changes in macrophage gene expression. J. Immunol. *164*, 5894–5904.

Rouhani, F., Kumasaka, N., de Brito, M.C., Bradley, A., Vallier, L., and Gaffney, D. (2014). Genetic background drives transcriptional variation in human induced pluripotent stem cells. PLoS Genet. *10*, e1004432.

Rozowsky, J., Abyzov, A., Wang, J., Alves, P., Raha, D., Harmanci, A., Leng, J., Bjornson, R., Kong, Y., Kitabayashi, N., et al. (2011). AlleleSeq: analysis of allele-specific expression and binding in a network framework. Mol. Syst. Biol. *7*, 522.

Sandberg, R., Neilson, J.R., Sarma, A., Sharp, P.A., and Burge, C.B. (2008). Proliferating cells express mRNAs with shortened 3' untranslated regions and fewer microRNA target sites. Science *320*, 1643–1647.

Schildberger, A., Rossmanith, E., Eichhorn, T., Strassl, K., and Weber, V. (2013). Monocytes, peripheral blood mononuclear cells, and THP-1 cells exhibit different cytokine expression patterns following stimulation with lipopolysaccharide. Mediators Inflamm. *2013*, 697972.

Schmidt, S.V., Krebs, W., Ulas, T., Xue, J., Baßler, K., Günther, P., Hardt, A.-L., Schultze, H., Sander, J., Klee, K., et al. (2016). The transcriptional regulator network of human inflammatory macrophages is defined by open chromatin. Cell Res. *26*, 151–170.

Schmieder, A., Michel, J., Schönhaar, K., Goerdt, S., and Schledzewski, K. (2012). Differentiation and gene expression profile of tumor-associated macrophages. Semin. Cancer Biol. *22*, 289–297.

Schroder, K., Hertzog, P.J., Ravasi, T., and Hume, D.A. (2004). Interferon-gamma: an overview of signals, mechanisms and functions. J. Leukoc. Biol. *75*, 163–189.

Schroder, K., Irvine, K.M., Taylor, M.S., Bokil, N.J., Le Cao, K.-A., Masterman, K.-A., Labzin, L.I., Semple, C.A., Kapetanovic, R., Fairbairn, L., et al. (2012). Conservation and divergence in Toll-like receptor 4-regulated gene expression in primary human versus mouse macrophages. Proc. Natl. Acad. Sci. U. S. A. *109*, E944–E953.

Shao, W., Halachmi, S., and Brown, M. (2002). ERAP140, a conserved tissue-specific nuclear receptor coactivator. Mol. Cell. Biol. *22*, 3358–3372.

Shin, H.Y., Willi, M., Yoo, K.H., Zeng, X., Wang, C., Metser, G., and Hennighausen, L. (2016). Hierarchy within the mammary STAT5-driven Wap super-enhancer. Nat. Genet. 48, 904–911.

Shukla, S., Kavak, E., Gregory, M., Imashimizu, M., Shutinoski, B., Kashlev, M., Oberdoerffer, P., Sandberg, R., and Oberdoerffer, S. (2011). CTCF-promoted RNA polymerase II pausing links DNA methylation to splicing. Nature *479*, 74–79.

Smith, E.N., and Kruglyak, L. (2008). Gene-environment interaction in yeast gene expression. PLoS Biol. *6*, e83.

Soehnlein, O., and Lindbom, L. (2010). Phagocyte partnership during the onset and resolution of inflammation. Nat. Rev. Immunol. *10*, 427–439.

Soucie, E.L., Weng, Z., Geirsdóttir, L., Molawi, K., Maurizio, J., Fenouil, R., Mossadegh-Keller, N., Gimenez, G., VanHille, L., Beniazza, M., et al. (2016). Lineage-specific enhancers activate

self-renewal genes in macrophages and embryonic stem cells. Science 351, aad5510.

Stegle, O., Parts, L., Durbin, R., and Winn, J. (2010). A Bayesian framework to account for complex non-genetic factors in gene expression levels greatly increases power in eQTL studies. PLoS Comput. Biol. *6*, e1000770.

Stegle, O., Parts, L., Piipari, M., Winn, J., and Durbin, R. (2012). Using probabilistic estimation of expression residuals (PEER) to obtain increased power and interpretability of gene expression analyses. Nat. Protoc. *7*, 500–507.

Stojnic, R., and Diez, D. (2015). PWMEnrich: PWM enrichment analysis. R package version 4.8.2.

Su, X., Yu, Y., Zhong, Y., Giannopoulou, E.G., Hu, X., Liu, H., Cross, J.R., Rätsch, G., Rice, C.M., and Ivashkiv, L.B. (2015). Interferon-γ regulates cellular metabolism and mRNA translation to potentiate macrophage activation. Nat. Immunol. 16, 838–849.

Takeuchi, O., and Akira, S. (2010). Pattern recognition receptors and inflammation. Cell *140*, 805–820.

Tan, G., and Lenhard, B. (2016). TFBSTools: an R/bioconductor package for transcription factor binding site analysis. Bioinformatics 32, 1555–1556.

The ENCODE Project Consortium (2012). An integrated encyclopedia of DNA elements in the human genome. Nature *489*, 57–74.

The GTEx Consortium (2015). The Genotype-Tissue Expression (GTEx) pilot analysis: Multitissue gene regulation in humans. Science *348*, 648–660.

Trapnell, C., Pachter, L., and Salzberg, S.L. (2009). TopHat: discovering splice junctions with RNA-Seq. Bioinformatics *25*, 1105–1111.

Trapnell, C., Williams, B.A., Pertea, G., Mortazavi, A., Kwan, G., van Baren, M.J., Salzberg, S.L., Wold, B.J., and Pachter, L. (2010). Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. Nat. Biotechnol. *28*, 511–515.

Trapnell, C., Hendrickson, D.G., Sauvageau, M., Goff, L., Rinn, J.L., and Pachter, L. (2013). Differential analysis of gene regulation at transcript resolution with RNA-seq. Nat. Biotechnol.

31, 46–53.

Trompouki, E., Bowman, T.V., Lawton, L.N., Fan, Z.P., Wu, D.-C., DiBiase, A., Martin, C.S., Cech, J.N., Sessa, A.K., Leblanc, J.L., et al. (2011). Lineage regulators direct BMP and Wnt pathways to cell-specific programs during differentiation and regeneration. Cell *147*, 577–589.

Trynka, G., Sandor, C., Han, B., Xu, H., Stranger, B.E., Liu, X.S., and Raychaudhuri, S. (2013). Chromatin marks identify critical cell types for fine mapping complex trait variants. Nat. Genet. *45*, 124–130.

Tsuchiya, S., Kobayashi, Y., Goto, Y., Okumura, H., Nakae, S., Konno, T., and Tada, K. (1982). Induction of maturation in cultured human monocytic leukemia cells by a phorbol diester. Cancer Res. *42*, 1530–1536.

Turro, E., Su, S.-Y., Gonçalves, Â., Coin, L.J.M., Richardson, S., and Lewin, A. (2011). Haplotype and isoform specific expression estimation using multi-mapping RNA-seq reads. Genome Biol. *12*, R13.

Turro, E., Astle, W.J., and Tavaré, S. (2014). Flexible analysis of RNA-seq data using mixed effects models. Bioinformatics *30*, 180–188.

Vallier, L., Touboul, T., Brown, S., Cho, C., Bilican, B., Alexander, M., Cedervall, J., Chandran, S., Ahrlund-Richter, L., Weber, A., et al. (2009). Signaling pathways controlling pluripotency and early cell fate decisions of human induced pluripotent stem cells. Stem Cells *27*, 2655–2666.

Venables, J.P., Lapasset, L., Gadea, G., Fort, P., Klinck, R., Irimia, M., Vignal, E., Thibault, P., Prinos, P., Chabot, B., et al. (2013). MBNL1 and RBFOX2 cooperate to establish a splicing programme involved in pluripotent stem cell differentiation. Nat. Commun. *4*, 2480.

Verma, I.M., Stevenson, J.K., Schwarz, E.M., Van Antwerp, D., and Miyamoto, S. (1995). Rel/NF-kappa B/I kappa B family: intimate tales of association and dissociation. Genes Dev. *9*, 2723–2735.

Visscher, P.M., Hill, W.G., and Wray, N.R. (2008). Heritability in the genomics era--concepts and misconceptions. Nat. Rev. Genet. *9*, 255–266.

Wagner, G.P., Kin, K., and Lynch, V.J. (2012). Measurement of mRNA abundance using RNA-seq data: RPKM measure is inconsistent among samples. Theory Biosci. *131*, 281–285.

Wallace, C., Rotival, M., Cooper, J.D., Rice, C.M., Yang, J.H.M., McNeill, M., Smyth, D.J., Niblett, D., Cambien, F., Cardiogenics Consortium, et al. (2012). Statistical colocalization of monocyte gene expression and genetic risk variants for type 1 diabetes. Hum. Mol. Genet. *21*, 2815–2824.

Wang, E.T., Sandberg, R., Luo, S., Khrebtukova, I., Zhang, L., Mayr, C., Kingsmore, S.F., Schroth, G.P., and Burge, C.B. (2008). Alternative isoform regulation in human tissue transcriptomes. Nature *456*, 470–476.

Waszak, S.M., Delaneau, O., Gschwind, A.R., Kilpinen, H., Raghav, S.K., Witwicki, R.M., Orioli, A., Wiederkehr, M., Panousis, N.I., Yurovsky, A., et al. (2015). Population Variation and Genetic Control of Modular Chromatin Architecture in Humans. Cell 162, 1039–1050.

Wehrspaun, C.C., Ponting, C.P., and Marques, A.C. (2014). Brain-expressed 3′UTR extensions strengthen miRNA cross-talk between ion channel/transporter encoding mRNAs. Front. Genet. *5*.

Weirauch, M.T., Yang, A., Albu, M., Cote, A.G., Montenegro-Montero, A., Drewe, P., Najafabadi, H.S., Lambert, S.A., Mann, I., Cook, K., et al. (2014). Determination and Inference of Eukaryotic Transcription Factor Sequence Specificity. Cell *158*, 1431–1443.

Welter, D., MacArthur, J., Morales, J., Burdett, T., Hall, P., Junkins, H., Klemm, A., Flicek, P., Manolio, T., Hindorff, L., et al. (2014). The NHGRI GWAS Catalog, a curated resource of SNP-trait associations. Nucleic Acids Res. *42*, D1001–D1006.

Whalen, S., Truty, R.M., and Pollard, K.S. (2016). Enhancer-promoter interactions are encoded by complex genomic signatures on looping chromatin. Nat. Genet. *48*, 488–496.

van Wilgenburg, B., Browne, C., Vowles, J., and Cowley, S.A. (2013). Efficient, long term production of monocyte-derived macrophages from human pluripotent stem cells under partly-defined and fully-defined conditions. PLoS One *8*, e71098.

Wilkie, G.S., Dickson, K.S., and Gray, N.K. (2003). Regulation of mRNA translation by 5'- and 3'-UTR-binding factors. Trends Biochem. Sci. *28*, 182–188.

Wong, D., Lee, W., Humburg, P., Makino, S., Lau, E., Naranbhai, V., Fairfax, B.P., Chan, K., Plant, K., and Knight, J.C. (2014). Genomic mapping of the MHC transactivator CIITA using an integrated ChIP-seq and genetical genomics approach. Genome Biol. *15*, 494.

Wynn, T.A., Chawla, A., and Pollard, J.W. (2013). Macrophage biology in development, homeostasis and disease. Nature *496*, 445–455.

Xaus, J., Cardó, M., Valledor, A.F., Soler, C., Lloberas, J., and Celada, A. (1999). Interferon gamma induces the expression of p21waf-1 and arrests macrophage cell cycle, preventing induction of apoptosis. Immunity *11*, 103–113.

Xia, Z., Donehower, L.A., Cooper, T.A., Neilson, J.R., Wheeler, D.A., Wagner, E.J., and Li, W. (2014). Dynamic analyses of alternative polyadenylation from RNA-seq reveal a 3'-UTR landscape across seven tumour types. Nat. Commun. *5*, 5274.

Xue, J., Schmidt, S.V., Sander, J., Draffehn, A., Krebs, W., Quester, I., De Nardo, D., Gohel, T.D., Emde, M., Schmidleithner, L., et al. (2014). Transcriptome-based network analysis reveals a spectrum model of human macrophage activation. Immunity *40*, 274–288.

Yanagimachi, M.D., Niwa, A., Tanaka, T., Honda-Ozaki, F., Nishimoto, S., Murata, Y., Yasumi, T., Ito, J., Tomida, S., Oshima, K., et al. (2013). Robust and Highly-Efficient Differentiation of Functional Monocytic Cells from Human Pluripotent Stem Cells under Serum- and Feeder Cell-Free Conditions. PLoS One *8*, e59243.

Yates, A., Akanni, W., Amode, M.R., Barrell, D., Billis, K., Carvalho-Silva, D., Cummins, C., Clapham, P., Fitzgerald, S., Gil, L., et al. (2016). Ensembl 2016. Nucleic Acids Res. 44, D710–D716.

Yoon, O.K., Hsu, T.Y., Im, J.H., and Brem, R.B. (2012). Genetics and regulatory impact of alternative polyadenylation in human B-lymphoblastoid cells. PLoS Genet. *8*, e1002882.

Yu, L., Croze, E., Yamaguchi, K.D., Tran, T., Reder, A.T., Litvak, V., and Volkert, M.R. (2014). Induction of a unique isoform of the NCOA7 oxidation resistance gene by interferon β-1b. J. Interferon Cytokine Res. 35, 186–199.

Zhang, H., Xue, C., Shah, R., Bermingham, K., Hinkle, C.C., Li, W., Rodrigues, A., Tabita-Martinez, J., Millar, J.S., Cuchel, M., et al. (2015). Functional analysis and transcriptomic profiling of iPSC-derived macrophages and their application in modeling Mendelian disease. Circ. Res. 117, 17–28.

Zhang, Y., Liu, T., Meyer, C.A., Eeckhoute, J., Johnson, D.S., Bernstein, B.E., Nusbaum, C., Myers, R.M., Brown, M., Li, W., et al. (2008b). Model-based analysis of ChIP-Seq (MACS).

Genome Biol. *9*, R137.

Zhao, H., Sun, Z., Wang, J., Huang, H., Kocher, J.-P., and Wang, L. (2014). CrossMap: a versatile tool for coordinate conversion between genome assemblies. Bioinformatics *30*, 1006–1007.

Zheng, X., Levine, D., Shen, J., Gogarten, S.M., Laurie, C., and Weir, B.S. (2012). A high-performance computing toolset for relatedness and principal component analysis of SNP data. Bioinformatics *28*, 3326–3328.

Zhernakova, D.V., de Klerk, E., Westra, H.-J., Mastrokolias, A., Amini, S., Ariyurek, Y., Jansen, R., Penninx, B.W., Hottenga, J.J., Willemsen, G., et al. (2013). DeepSAGE reveals genetic variants associated with alternative polyadenylation and expression of coding and non-coding transcripts. PLoS Genet. *9*, e1003594.

Zhou, J., and Troyanskaya, O.G. (2015). Predicting effects of noncoding variants with deep learning-based sequence model. Nat. Methods *12*, 931–934.

Zhu, Z., Zhang, F., Hu, H., Bakshi, A., Robinson, M.R., Powell, J.E., Montgomery, G.W., Goddard, M.E., Wray, N.R., Visscher, P.M., et al. (2016). Integration of summary data from GWAS and eQTL studies predicts complex trait gene targets. Nat. Genet. 48, 481–487.

Zlotnik, A., and Yoshie, O. (2012). The chemokine superfamily revisited. Immunity *36*, 705–716.