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).

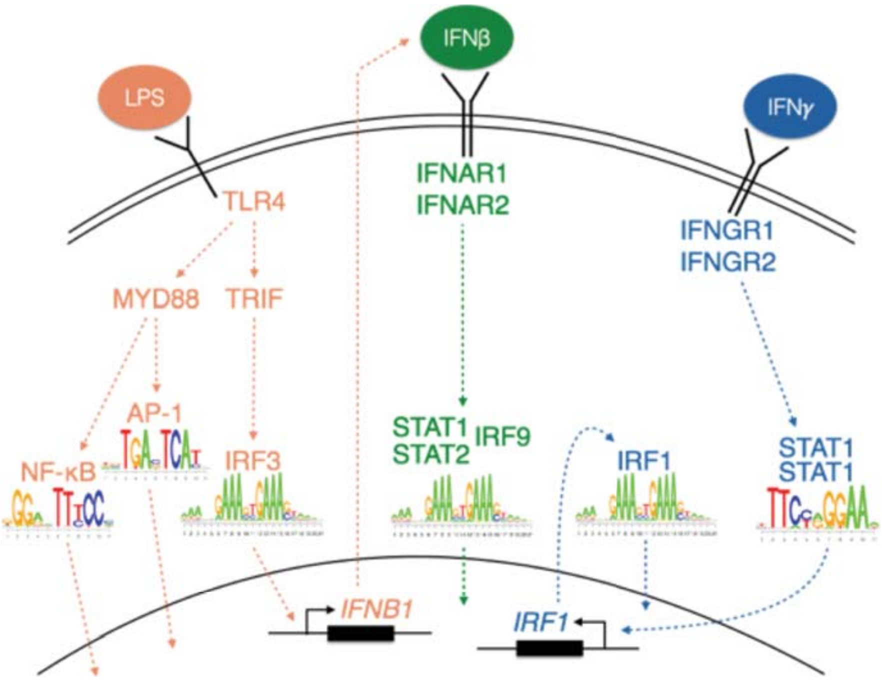


Figure 1.1: Main signalling pathways activated in macrophages after *Salmonella* infection and IFNɣ stimulation. Macrophages recognise LPS on the *Salmonella* cell wall via the TLR4 receptor (Medzhitov and Horng, 2009). Ligand binding to multiple TLRs such as TLR2, TLR3, TLR4 and TLR9 leads to downstream 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.

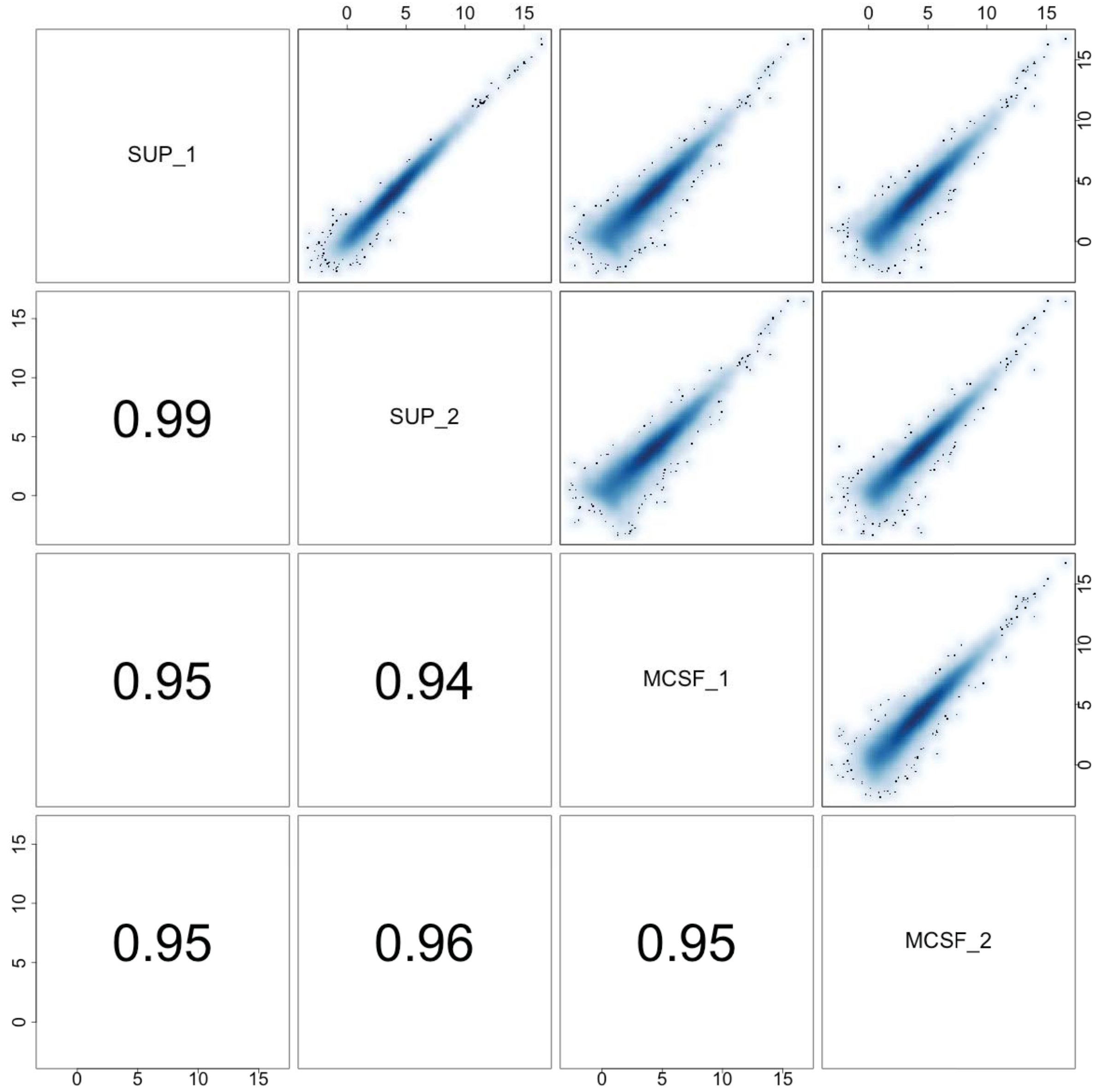


Figure 2.1. Biological reproducibility of IPSDM differentiation. Two biological replicates of FSPS10C-derived IPSDMs differentiated with either supernatant (SUP_1 and SUP_2) or recombinant M-CSF (MCSF_1 and MCSF_2). Above diagonal: pairwise scatterplots of expressed genes (transcripts per million (TPM) > 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.

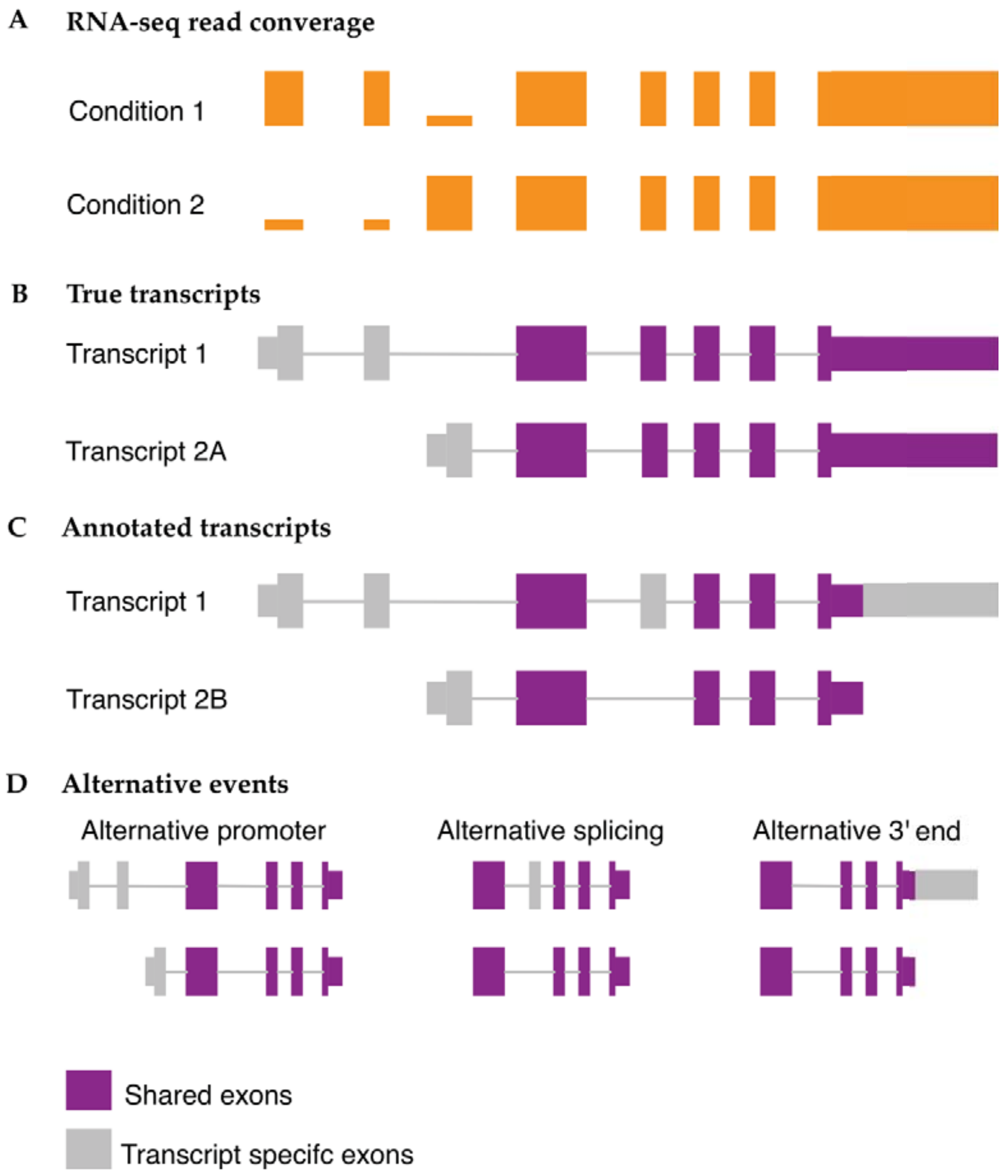


Figure 2.2. Constructing alternative transcription events from annotated transcripts. (A) Hypothetical RNA-seq read coverage over a gene indicating that there is switch from proximal to distal promoter between conditions 1 and 2. **(B)** True transcript annotations generating the read coverage observed 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.

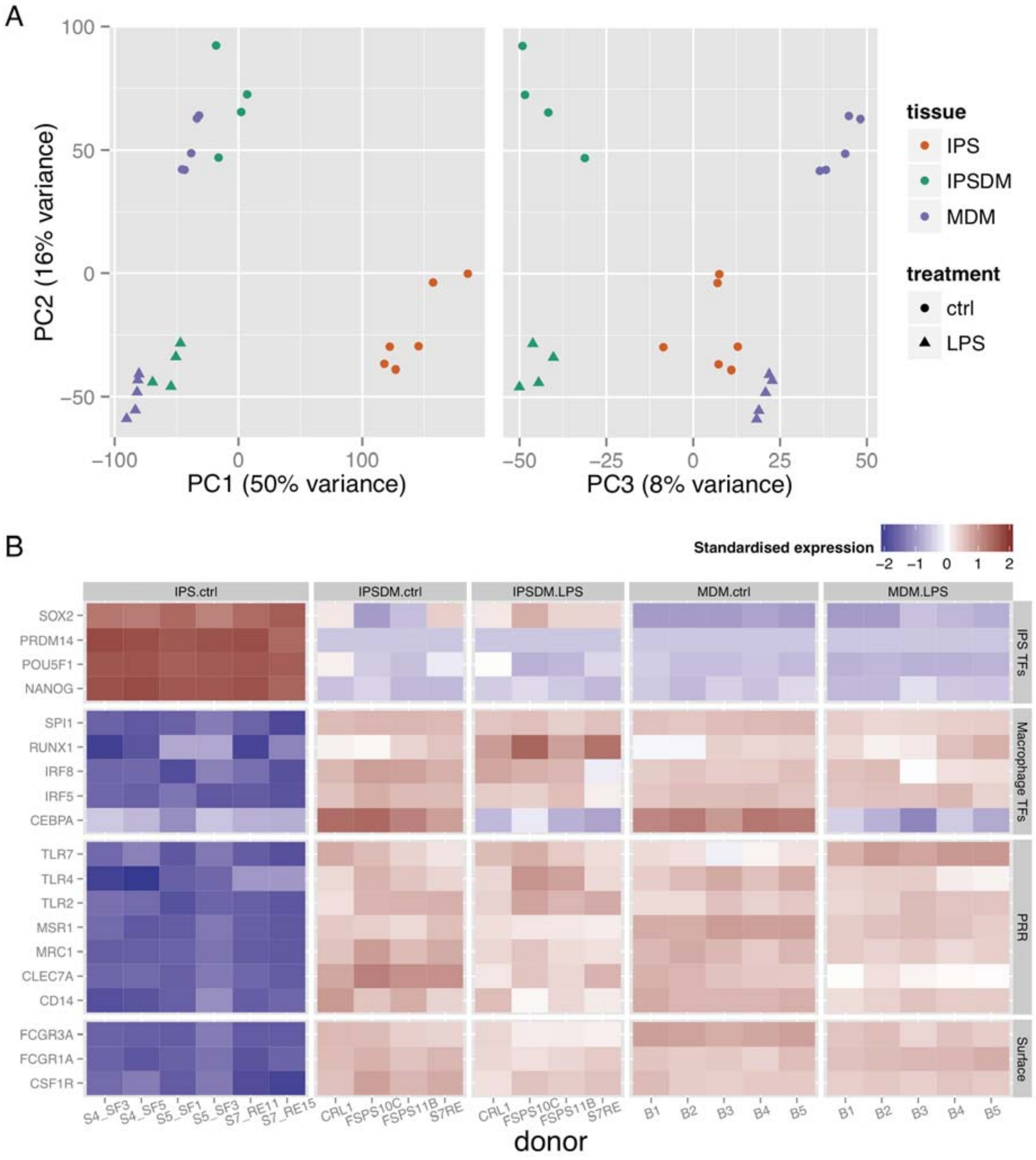
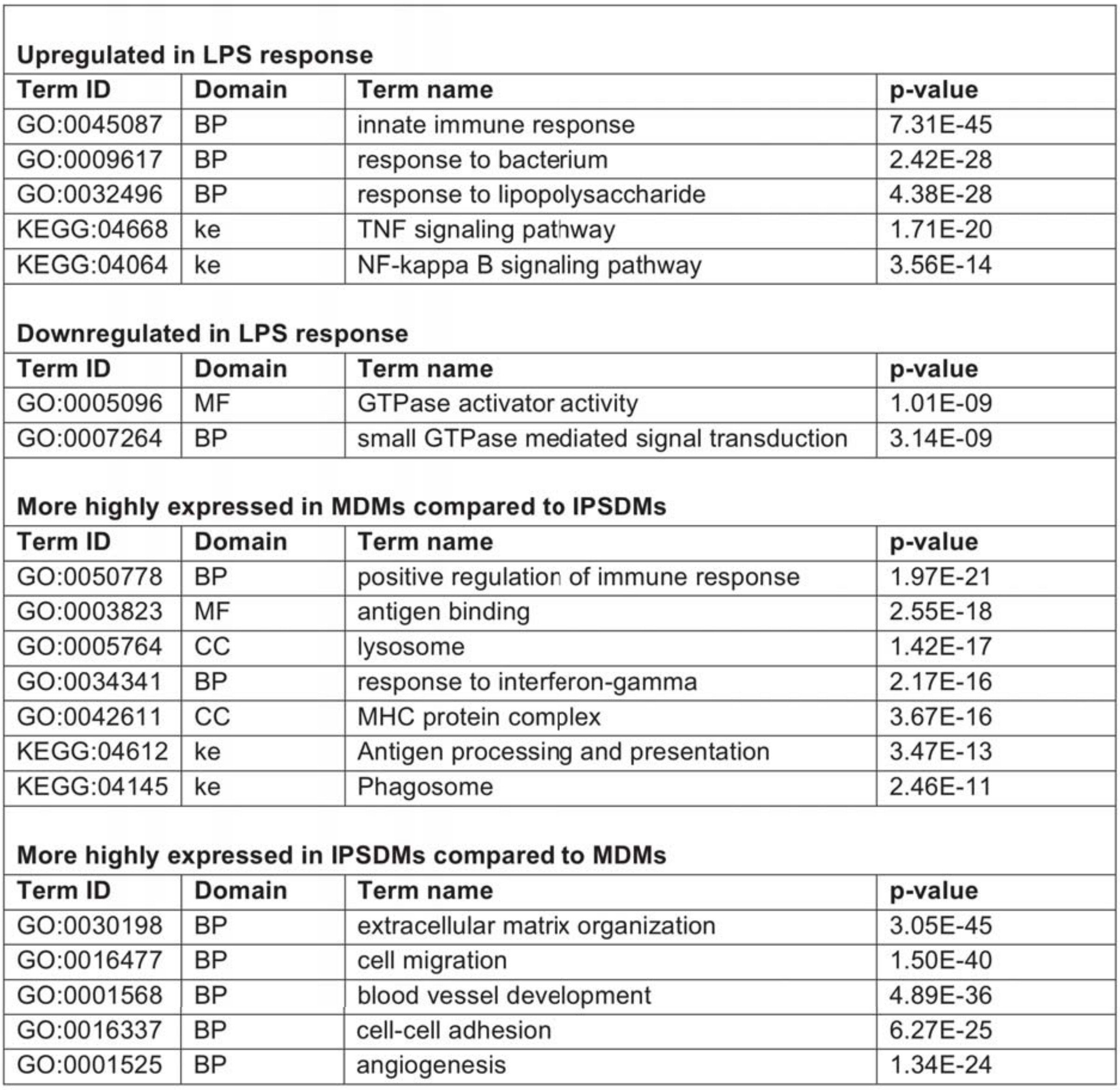


Figure 2.3. Gene expression variation between iPSCs, IPSDMs and MDMs. (A) Principal Component Analysis of expressed genes (TPM > 2) in iPSCs, IPSDMs and MDMs. **(B)** Heatmap showing the gene expression of selected iPSC-specific transcription factors (TFs), macrophage specific TFs, pattern recognition 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.

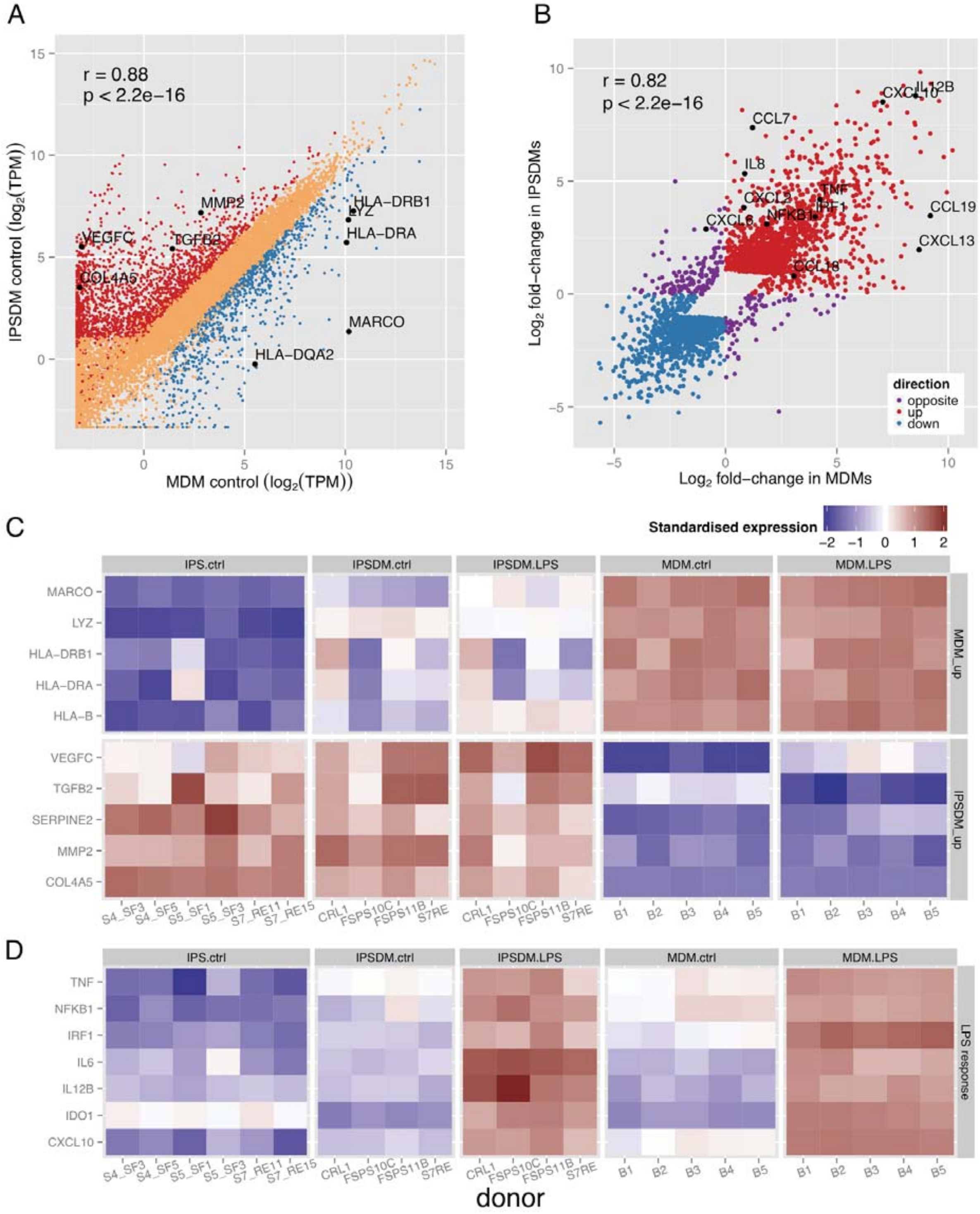


Figure 2.4. Differential expression analysis of IPSDMs vs MDMs. (A) Scatter plot of gene expression levels between MDMs and IPSDMs. Genes that are significantly more highly expressed in IPSDMs are shown in red and genes that are significantly more highly expressed in MDMs are shown in blue. **(B)** Scatter 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 = 4.5×10^{-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).

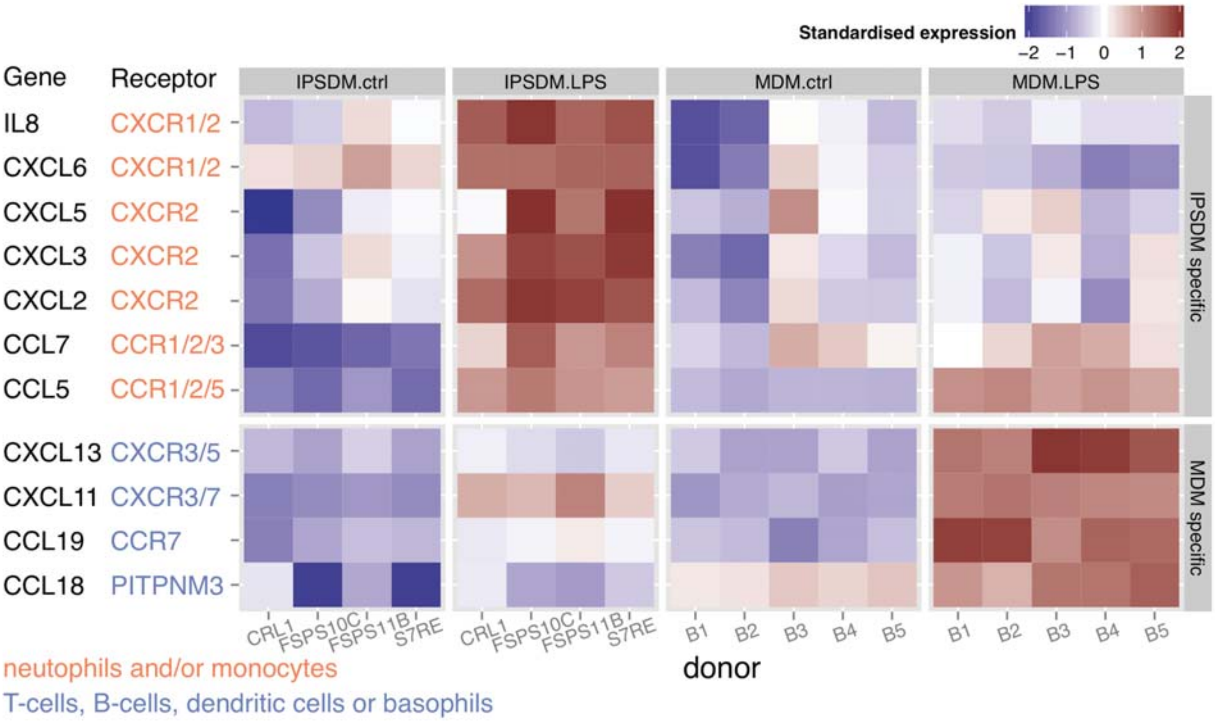


Figure 2.5. Chemokine genes that were particularly upregulated in either IPSDMs or MDMs in LPS response. Their annotated receptors and target cell types were taken from the literature (Soehnlein and Lindbom, 2010; Zlotnik and Yoshie, 2012).

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.

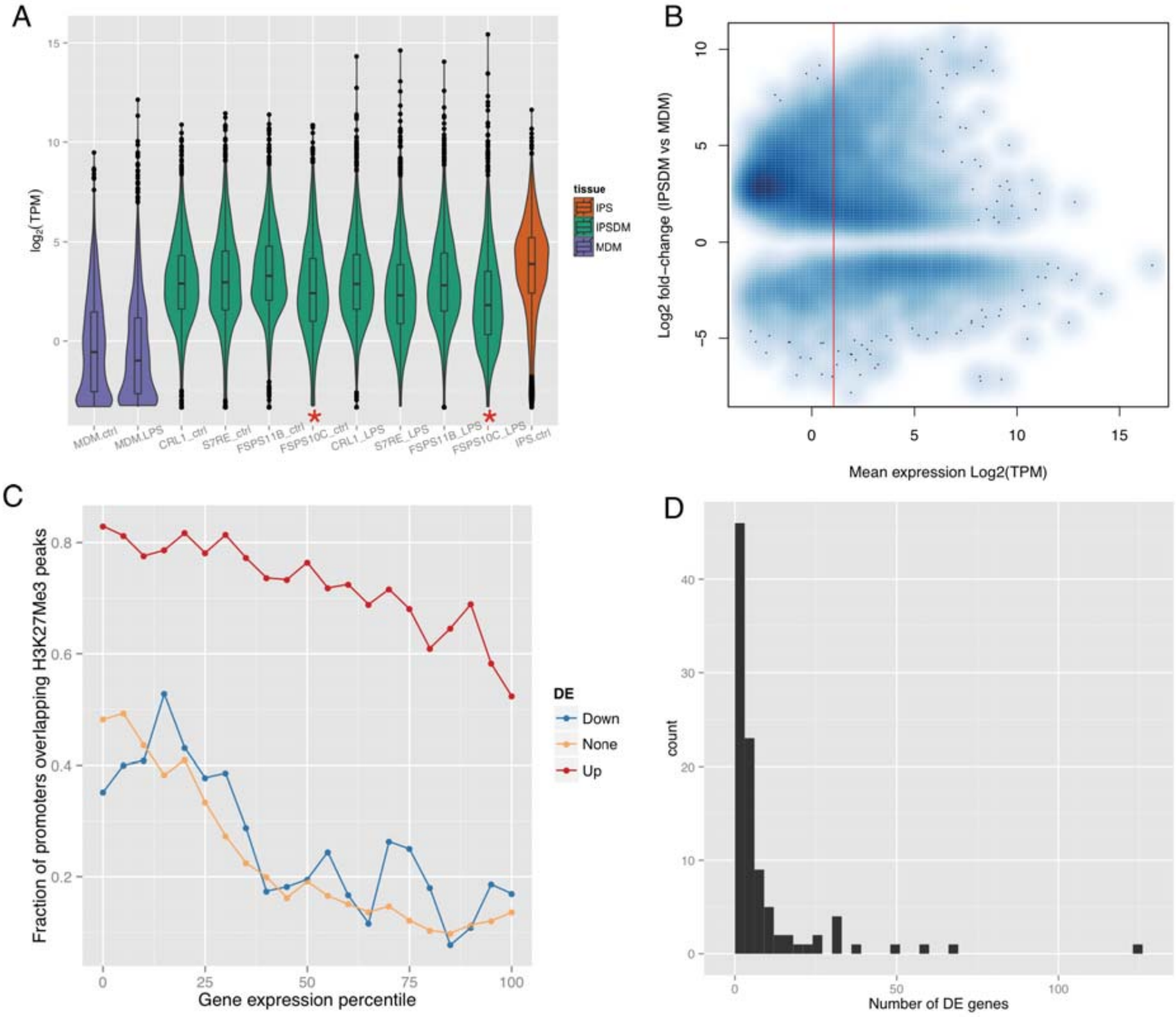


Figure 2.6: Mechanisms underlying differential expression between MDMs and IPSDMs. (A) Expression levels of genes that were more highly expressed in IPSDMs compared to MDMs (TPM > 2). Purple violin plots show the mean expression of these genes in MDMs and orange in IPS cells. Red asterisks mark IPSDM 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.

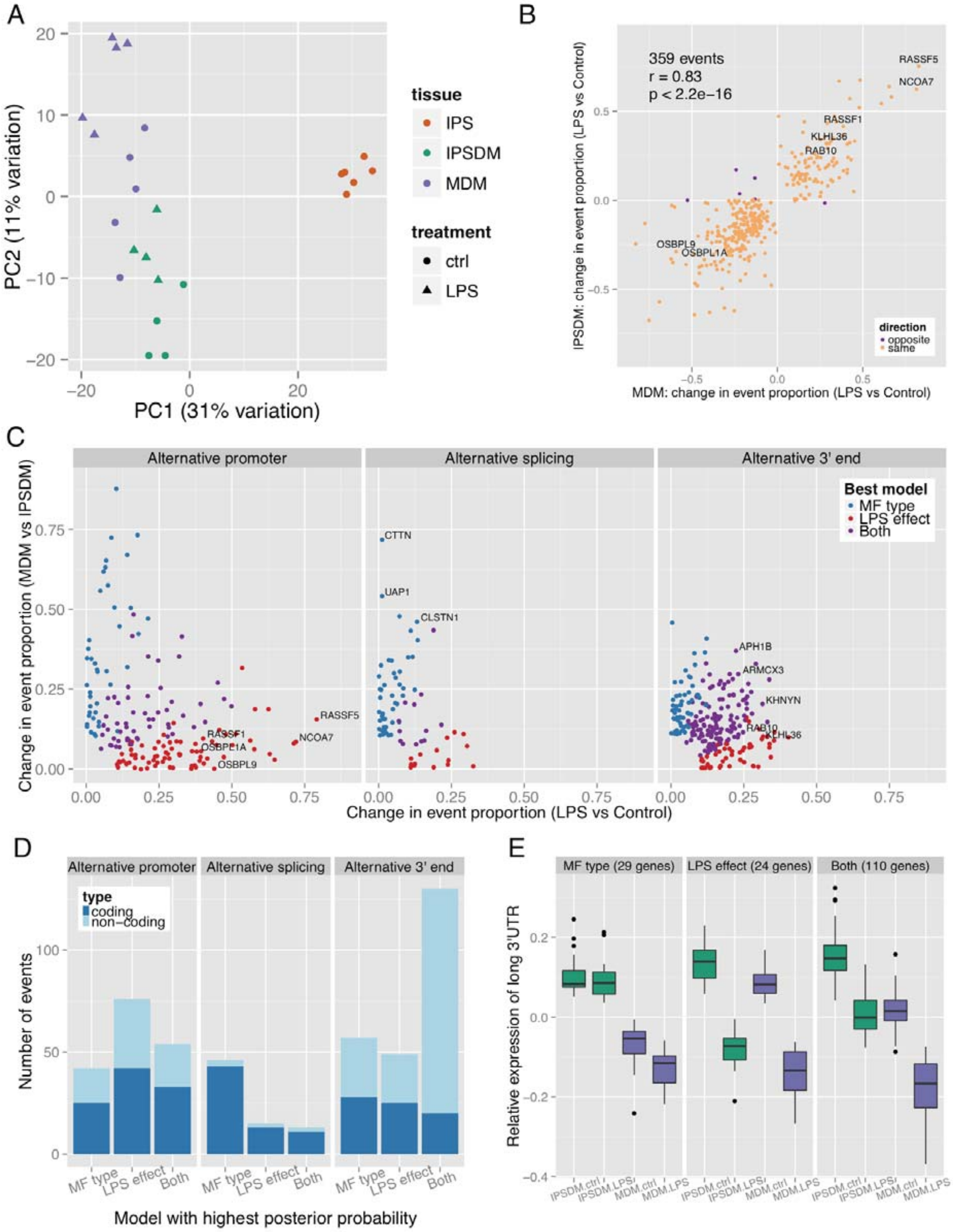


Figure 2.7. Alternative transcription in IPSDMs and MDMs. (A) PCA of relative transcript proportions in iPSCs, IPSDMs and MDMs. Only genes with mean TPM > 2 in all conditions were

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.

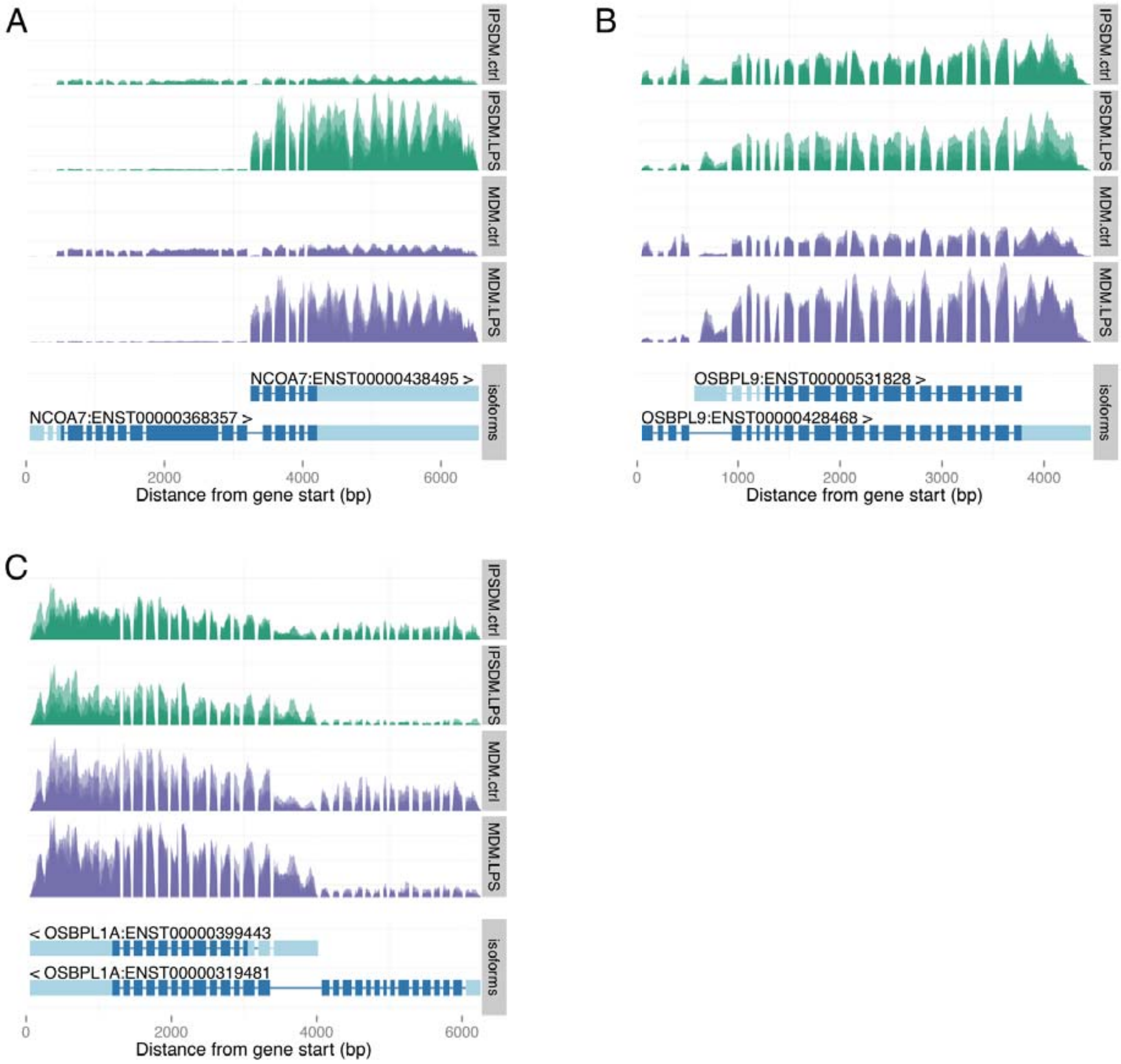


Figure 2.8. Examples of alternative promoter usage in LPS response. Each plot shows normalised read depth across the gene body in IPSDMs (green) and MDMs (purple) with gene structure in the panel beneath each plot. Introns have been compressed relative to exons to facilitate visualisation. (A-C) Alternative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

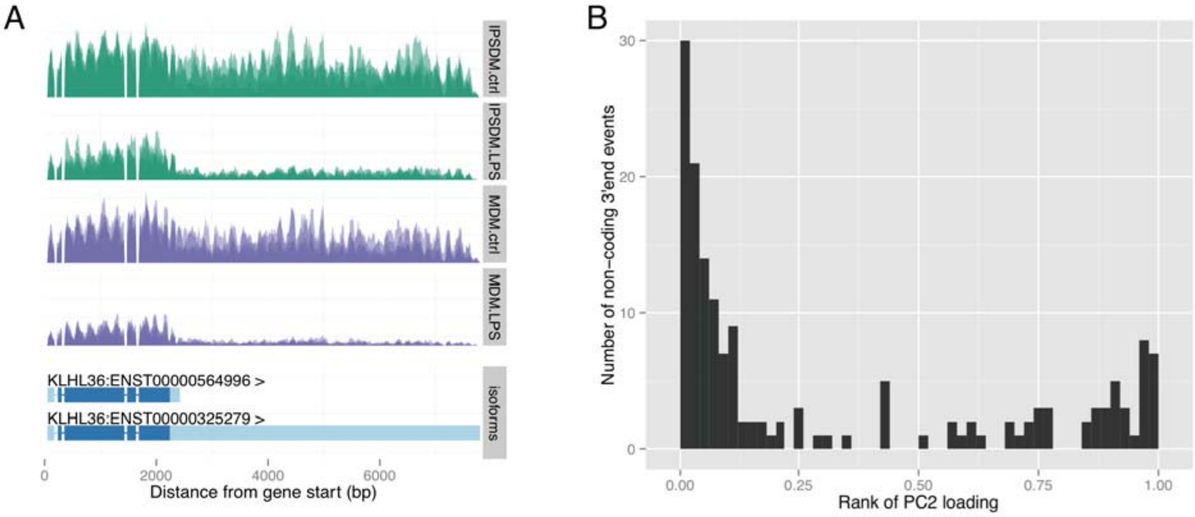


Figure 2.9. 3′ UTR shortening in LPS response. (A) Examples of 3′ UTR shortening in LPS response. The plot shows normalised read depth across the gene body in IPSDMs (green) and MDMs (purple) with gene structure in the panel beneath the plot. Introns have been compressed relative to exons to facilitate 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).

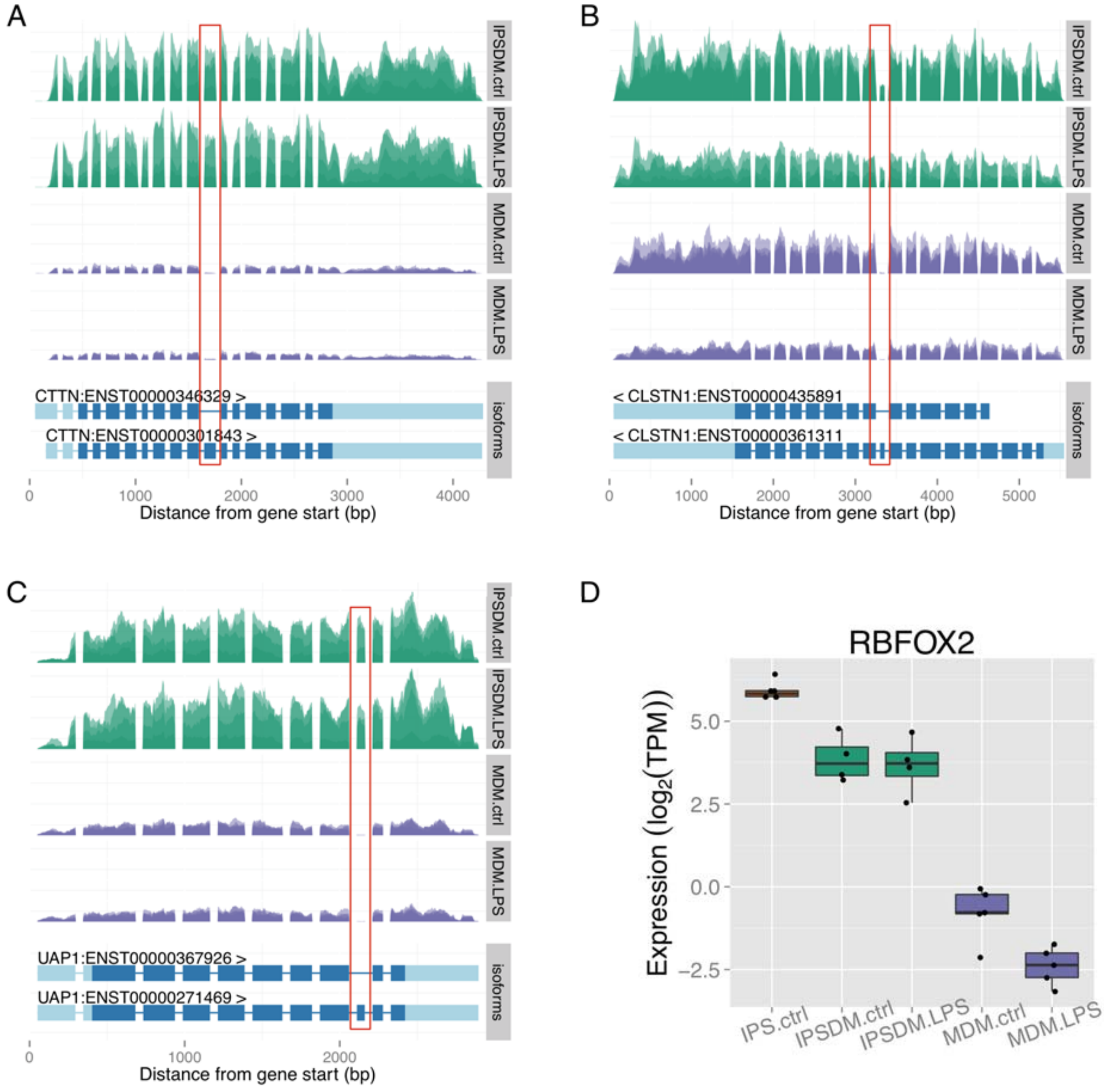


Figure 2.10. Alternative splicing between IPSDMs and MDMs. (A-C) Examples of alternative splicing between MDMs and IPSDMs. The alternatively spliced exon is marked with the red rectangle. **(D)** Expression of RBFOX2 gene in iPSCs, IPSDMs and MDMs.

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.