4 Genetics of gene expression in macrophage immune response

Collaboration note

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

4.1 Introduction

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

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

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

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

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

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

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

4.2 Methods

4.2.1 Gene expression analysis

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

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

Differential expression analysis

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

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

Detecting hidden confounders with PEER

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

4.2.2 Gene expression QTL mapping

Preparing genotype data

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

Detecting eQTLs using linear model

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

Quantifying allele-specific expression

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

Detecting QTLs using RASQUAL

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

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

Comparing RASQUAL and FastQTL results

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

Detecting condition-specific QTLs with a linear model

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

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

Filtering and clustering QTLs based on effect size

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

4.2.3 Alternative transcription analysis

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

Quantifying the expression of annotated alternative transcripts

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

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

Constructing alternative transcription events from known annotations

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

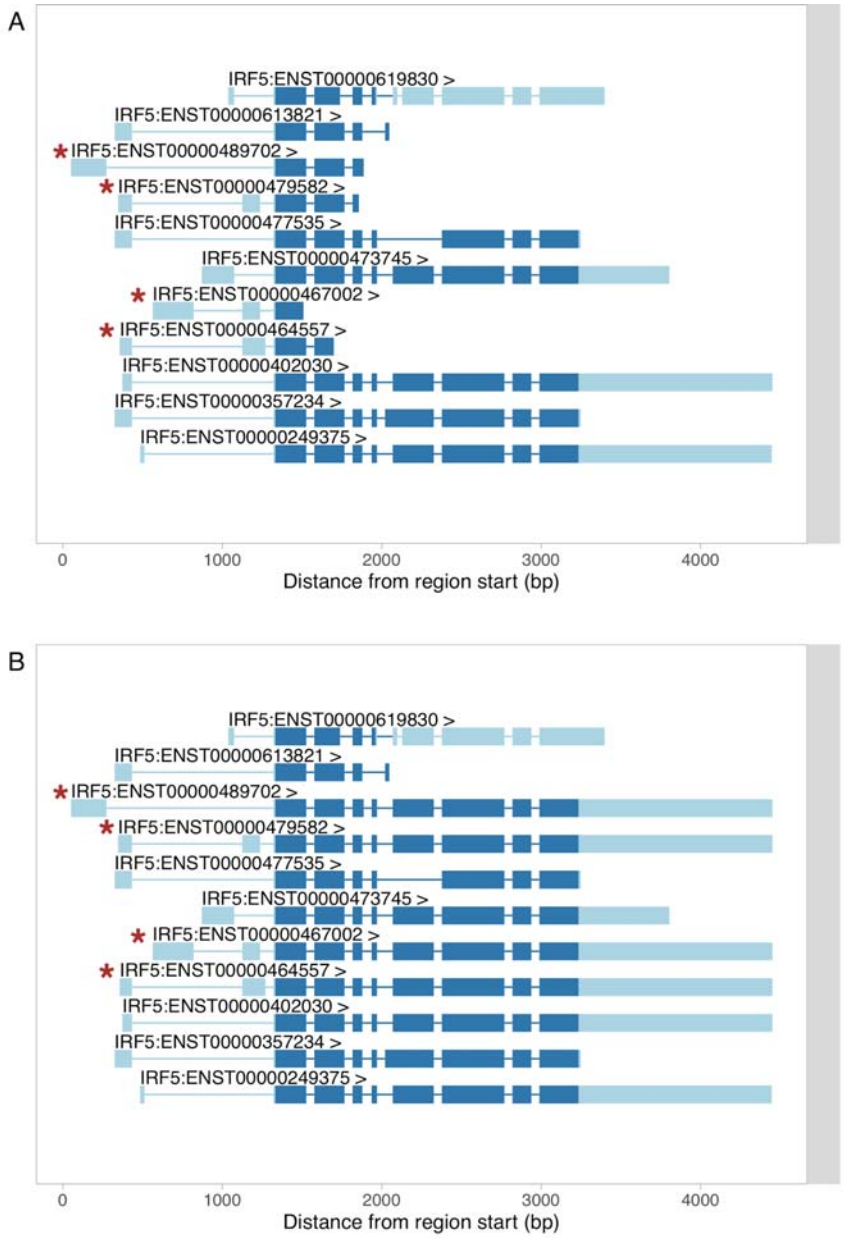

Figure 4.1: Extending truncated transcripts of the IRF5 gene. (A) Protein coding transcripts of the IRF5 gene from the Ensembl 85 gene set. The transcripts with annotated incomplete 3′ ends are marked with red asterisks. **(B)** Truncated transcripts have been extended using the exons from the transcript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

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

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

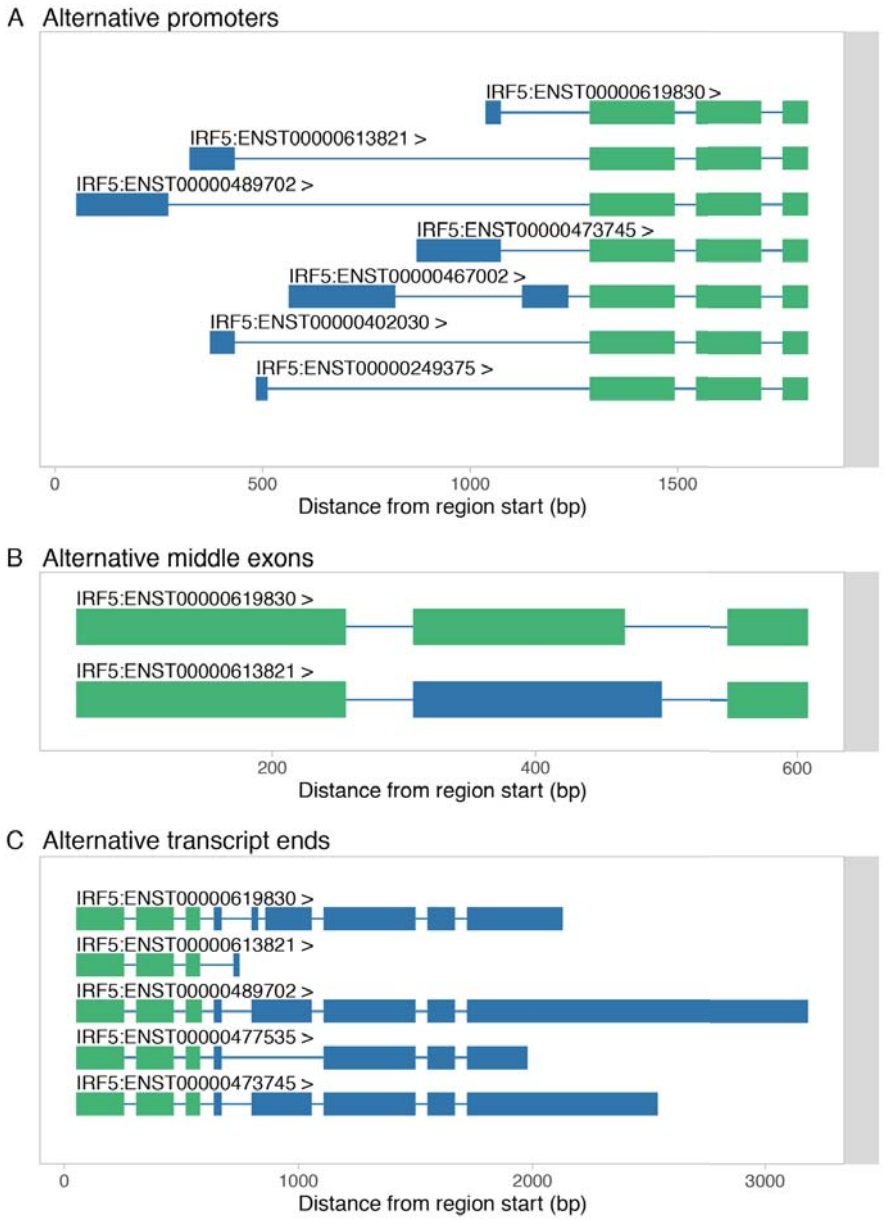

Figure 4.2: Alternative transcription events constructed from the 11 annotated transcripts (Figure 4.1B) of the IRF5 gene. Exons shared by all alternative events are highlighted in green and exons specific to some events are shown in blue.

Quantifying rates of intron retention

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

4.2.4 Transcript ratio QTL mapping

Data normalisation

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

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

Detecting transcript ratio QTLs

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

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

4.2.5 Overlap analysis with the NHGRI-EBI GWAS catalogue

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

4.2.6 QTL replicability between conditions

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

4.3 Quantifying gene expression and alternative transcription

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

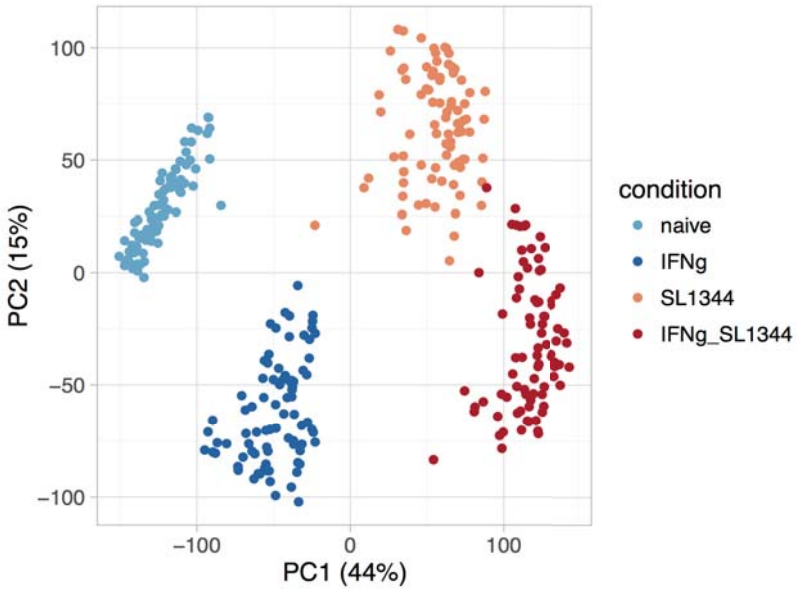

Figure 4.3. Principal component analysis of normalised and standardised gene expression data.

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

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

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

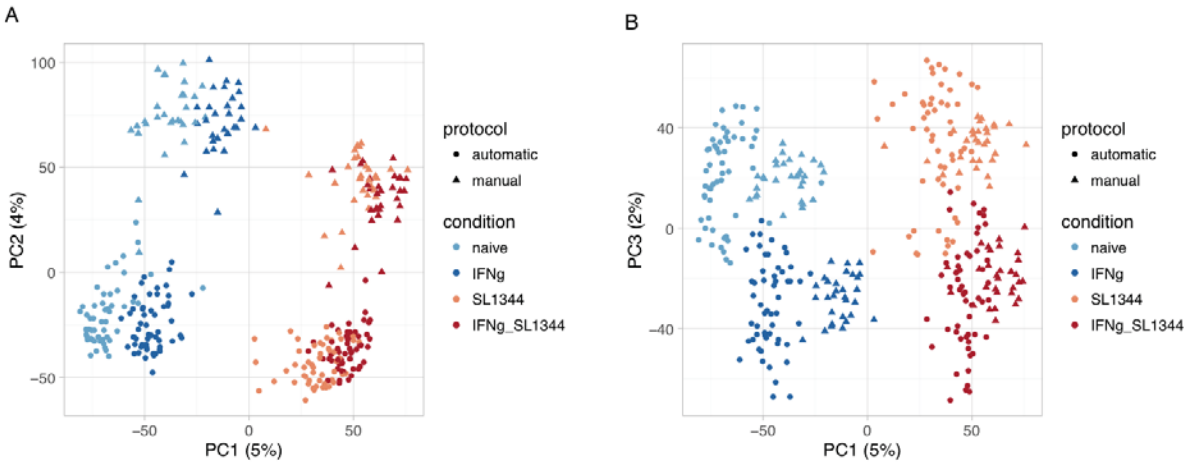

Figure 4.4: Principal component analysis of normalised intron excision ratios. (A) PC1 plotted against PC2. **(B)** PC1 plotted against PC3. **Protocol** - type of RNA-seq library construction protocol used, either manual or automatic.

4.3.1 Differential expression analysis reveals expected pathways

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

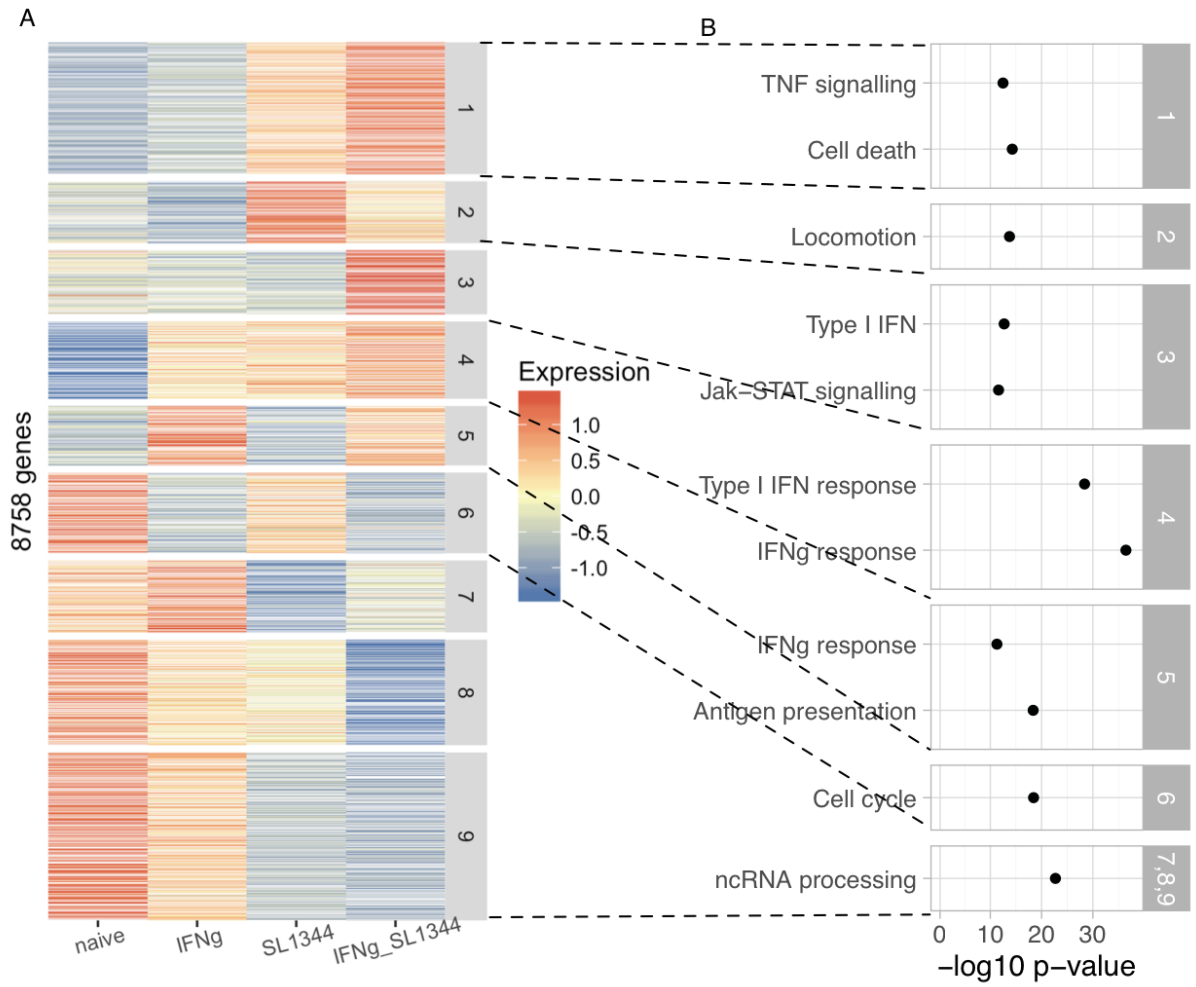

Figure 4.5: Differential gene expression between the four experimental conditions. (A) Heatmap of 8758 differentially expressed genes clustered into nine distinct patterns of expression. **(B)** A selection of Gene Ontology (GO) terms specifically enriched in each cluster. Only enrichments with $p <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

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

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

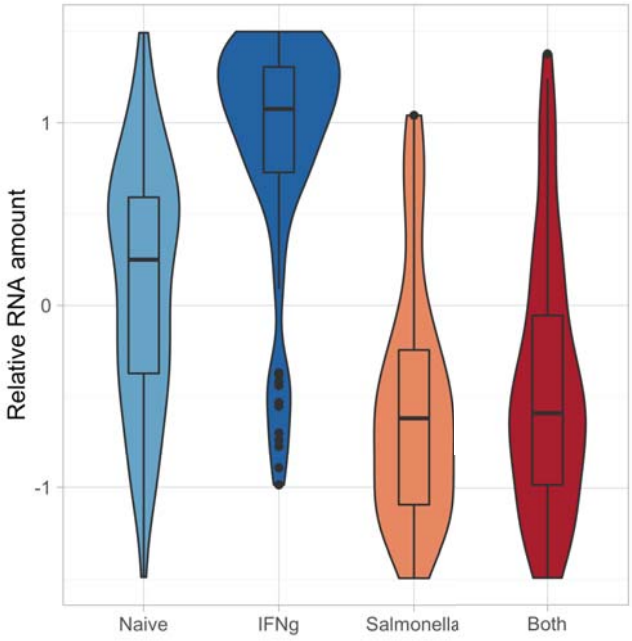

Figure 4.6: Relative amount of RNA obtained from each condition across 84 macrophage lines. I quantified the total amount of RNA obtained from each sample. For all four samples from a single line (corresponding to four conditions) I then subtracted the mean RNA amount across conditions and divided by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

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

4.4 Genetics of gene expression

4.4.1 Gene expression QTL mapping

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

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

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

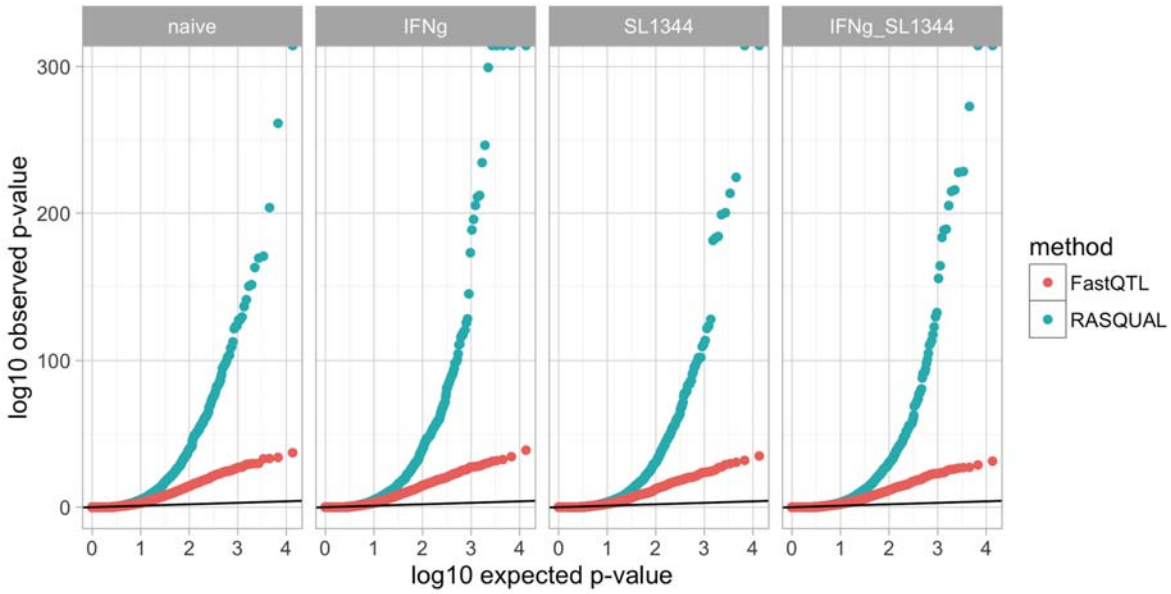

Figure 4.7: Quantile-quantile plots for the p-values of eQTLs detected either with RASQUAL or FastQTL. Solid lines represent the expected distribution of p-values under the null model.

4.4.2 Transcript ratio QTL mapping

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

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

Condition	LeafCutter	Salmon + Ensembl 85	Salmon + reviseAnnotations

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

4.4.3 Concordance of QTLs detected by different methods

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

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

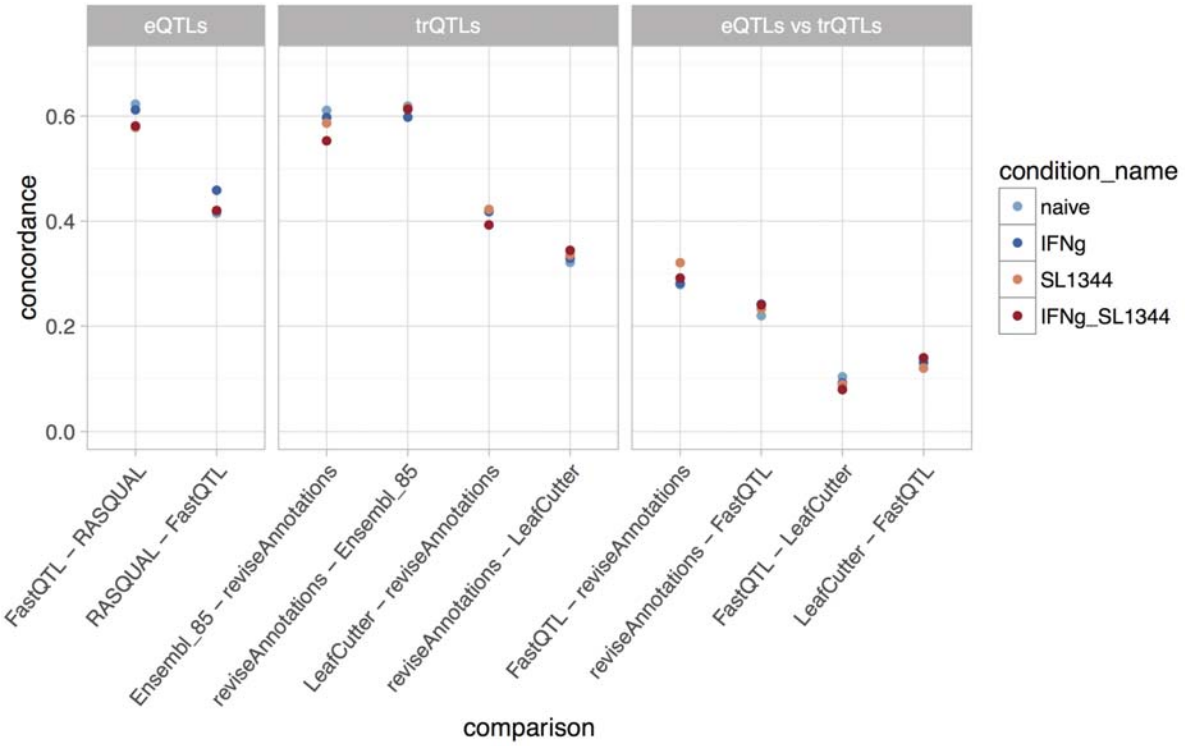

Figure 4.8: Concordance of lead QTL variants detected by different methods. In the gene expression (eQTL) comparison (left panel) I used FastQTL and RASQUAL lead variants from +/-500kb cis-window. For the eQTL and trQTL comparison (rightmost panel) I reran FastQTL eQTL mapping in a 100kb around the gene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

4.4.4 Condition specificity of eQTLs and trQTLs

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

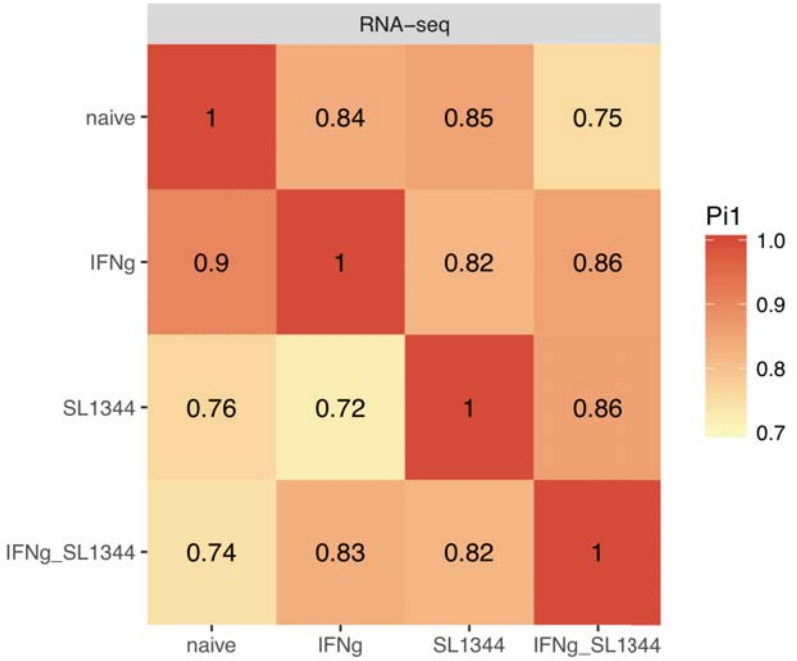

Figure 4.9: Replicability of eGenes between conditions. The heatmap shows the pairwise Storey's π_1 statistic for eQTLs detected between conditions.

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

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

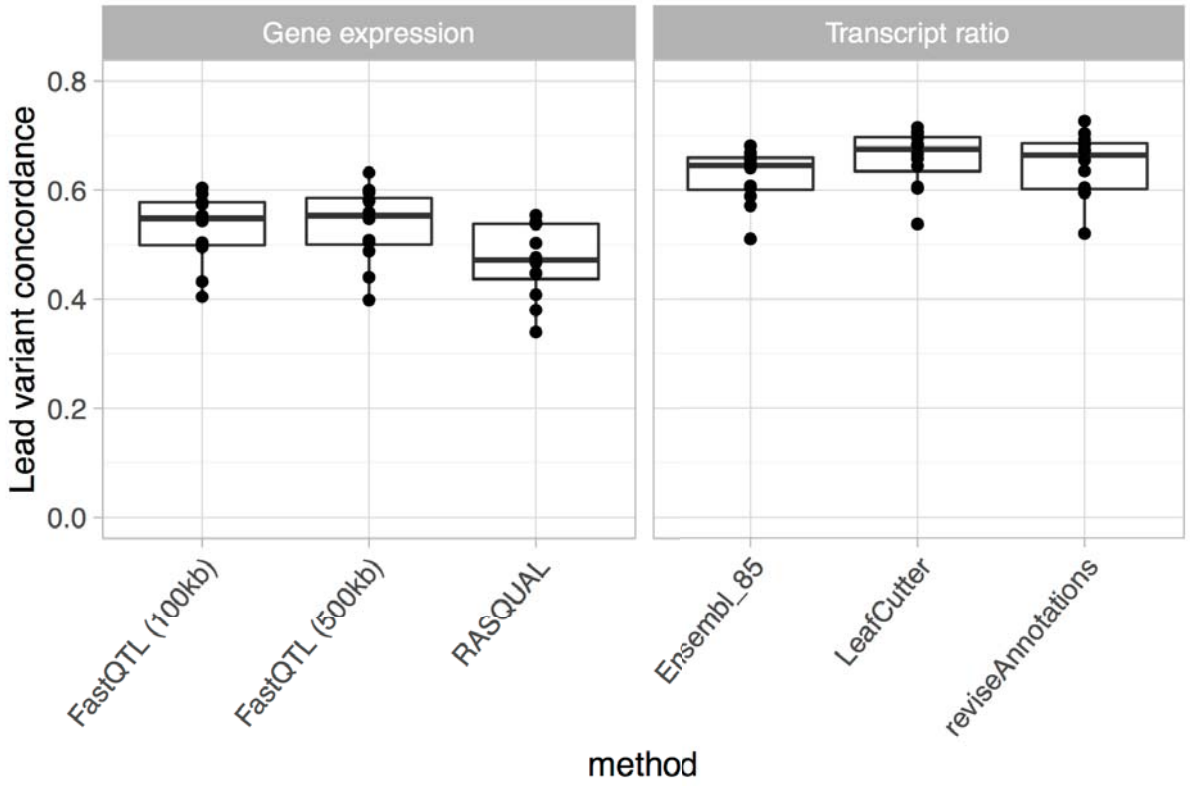

Figure 4.10: Concordance of QTL lead variants between pairs of conditions detected by different QTL mapping methods. Each dot represents one pairwise comparison between conditions (such as IFNɣ vs naive). I mapped eQTLs with FastQTL in both +/- 500kb and +/- 100kb cis-windows to match the 100 kb window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

Identifying condition-specific eQTLs

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

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

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

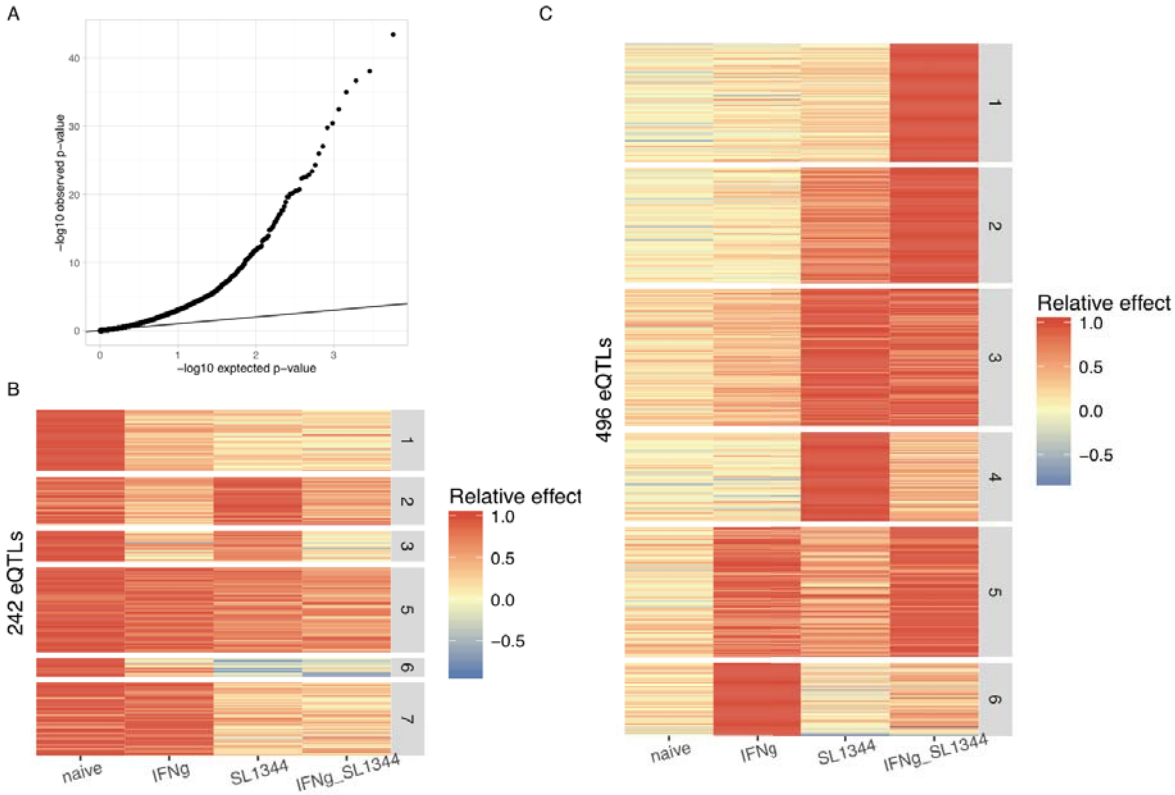

Figure 4.11: Condition-specific eQTLs clustered by their effect size. **(A)** Quantile-quantile plot of the expected and observed p-values for the interaction test **(B)** Effect size heatmap of the seven clusters of eQTLs that disappeared after stimulation. **(C)** Effect size heatmap of the six clusters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

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

4.5 Case study: genetics of IRF5 transcription

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

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

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

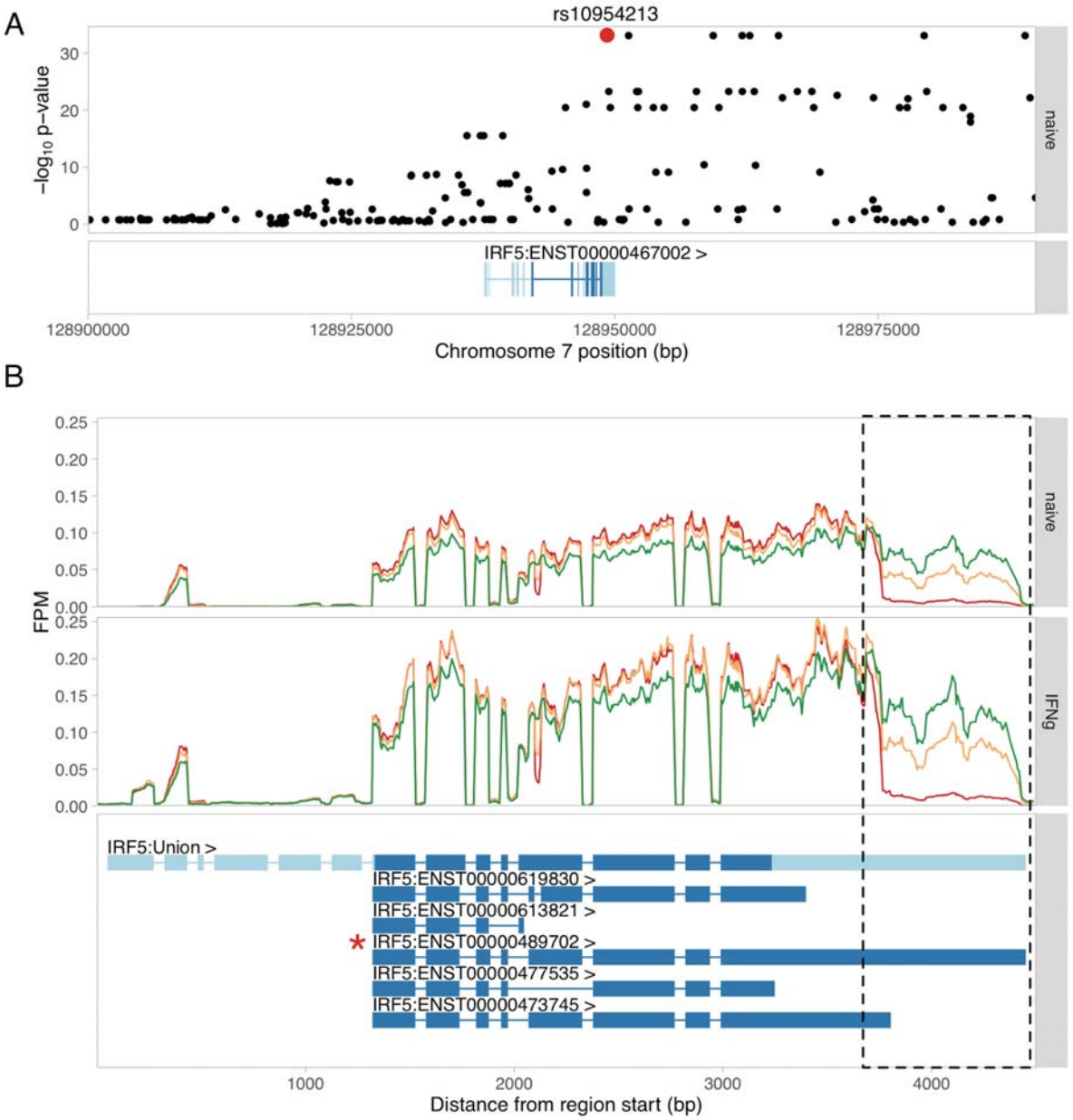

Figure 4.12: Example of a trQTL for the IRF5 gene that influences the proximal polyadenylation site usage. (A) Manhattan plot of the associated variants around the IRF5 gene in the naive condition. The lead variant rs10954213 disrupts the proximal polyadenylation site motif. **(B)** RNA-seq read coverage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

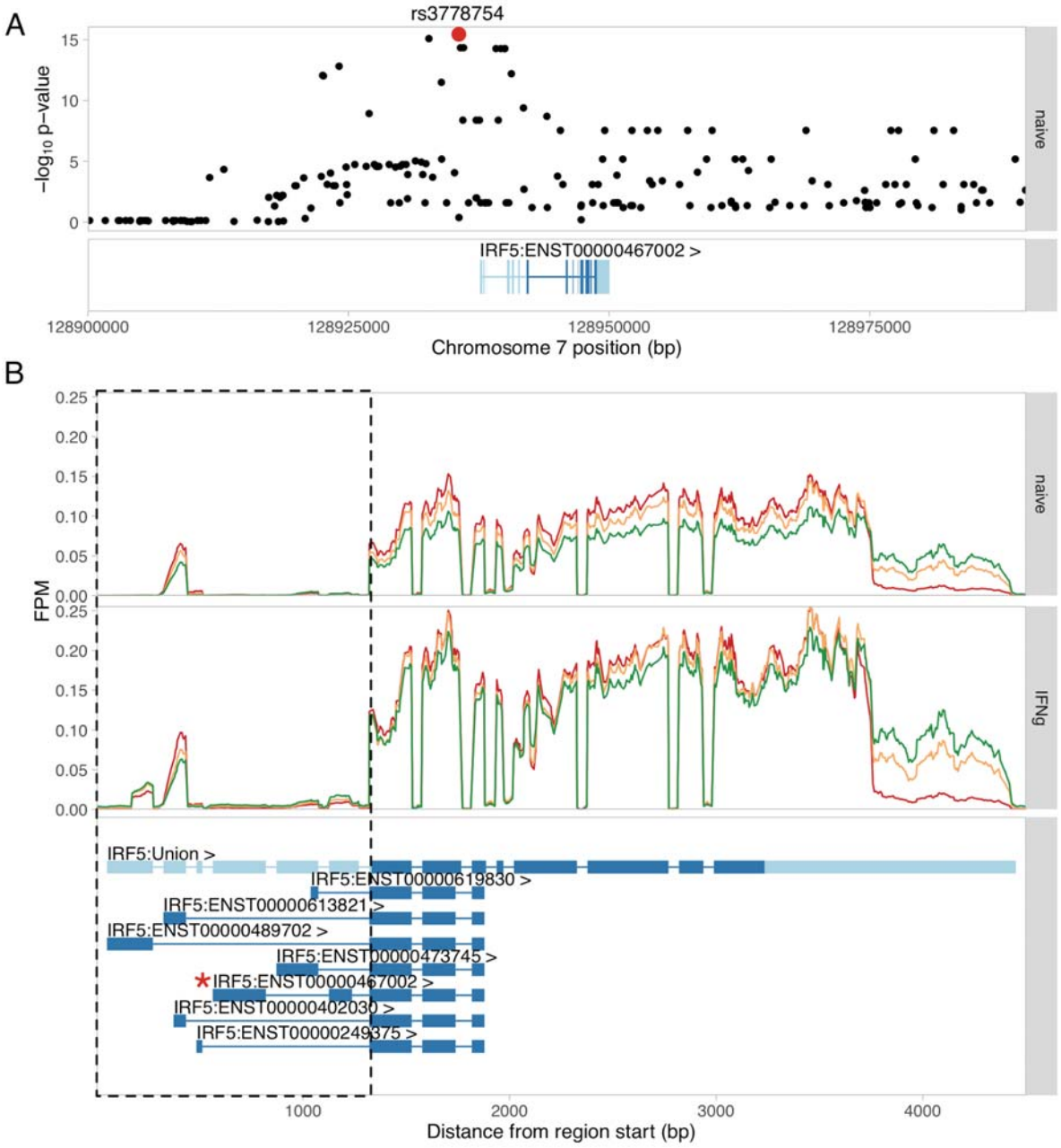

Figure 4.13: Alternative promoter QTL for the IRF5 gene. (A) Manhattan plot of the associated variants upstream of the IRF5 promoter in the naive condition. **(B)** RNA-seq read coverage across the IRF5 gene stratified by the genotype of the lead promoter QTL variant (rs3778754). The panel below the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

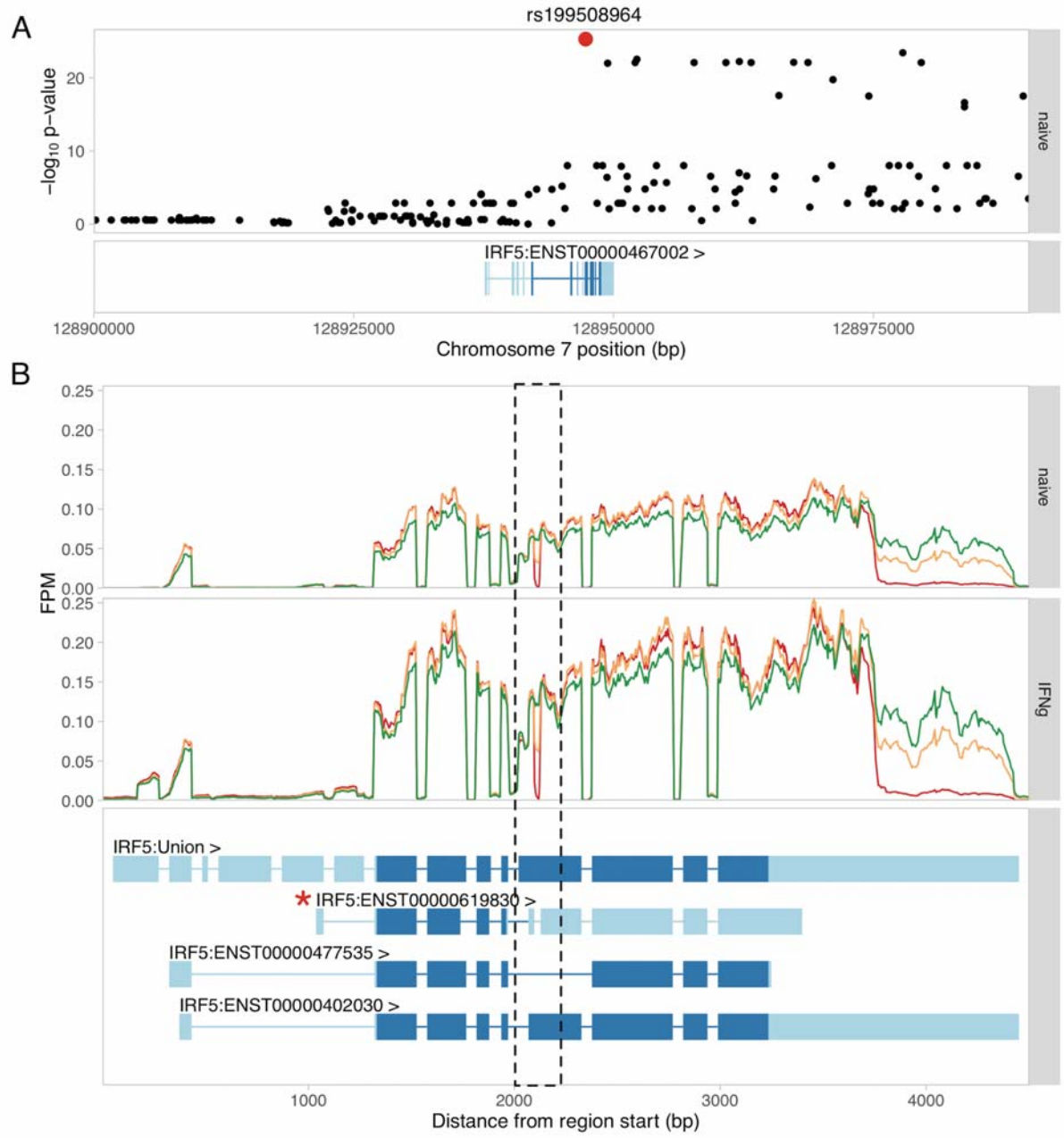

Figure 4.14: Intron excision QTL in the IRF5 gene. (A) Manhattan plot of eQTL p-values from RASQUAL in the naive condition. **(B)** Read coverage across the IRF5 gene stratified by the genotype of the lead QTL variant (rs199508964). The alternatively excised intron is highlighted by the dashed box.

4.6 Overlap with GWAS hits

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

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

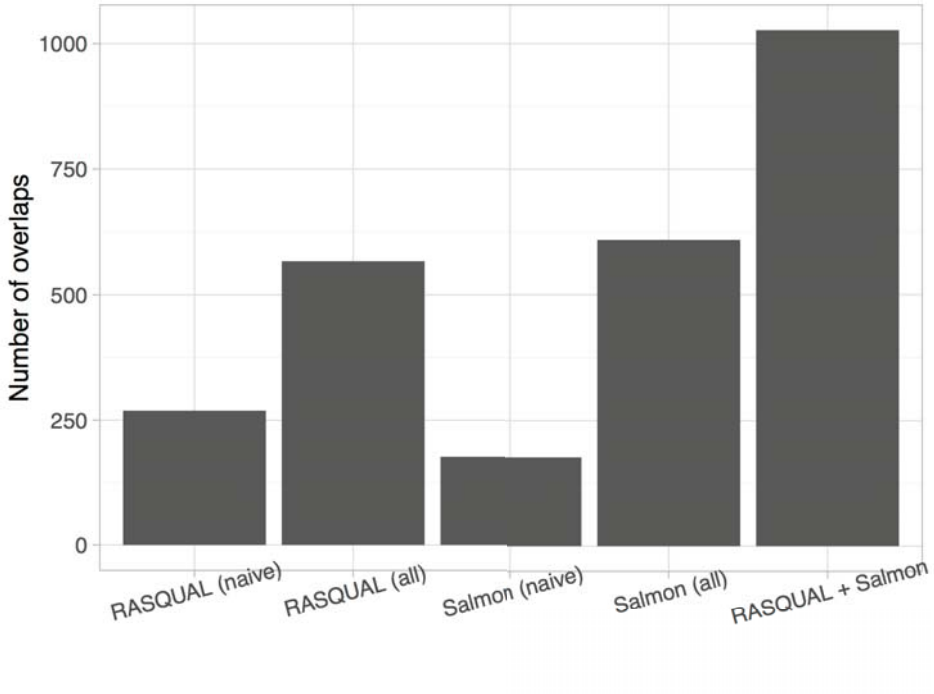

Figure 4.15: Number of RASQUAL eQTLs and Salmon trQTLs overlapping GWAS hits. 'Naive' represents QTLs from the unstimulated condition only while 'all' stands for all independent (R^2 < 0.8) QTLs across conditions. Lead QTL and GWAS variants were considered to be overlapping if the distance between the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

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

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

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

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

4.7 Discussion

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

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

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

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