5 Genetics of chromatin accessibility in macrophage immune response

Collaboration note

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

5.1 Introduction

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

Since the original caQTL experiment (Degner et al., 2012), other studies have followed looking at the genetics of histone modifications and transcription factor binding (Ding et al., 2014; Grubert et al., 2015; Waszak et al., 2015). However, due to the large cell numbers required by chromatin assays, all of these studies have been conducted in LCLs. Therefore, although the cell type and condition specificity of eQTLs is well established (Fairfax et al., 2012, 2014), how these effects manifest on the chromatin level and how they propagate down to gene expression is mostly unknown. The development of ATAC-seq (assay for transpose accessible chromatin) has made it possible to measure chromatin accessibility in much smaller number of cells, thus greatly increasing the number of cell types and conditions that can be profiled

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

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

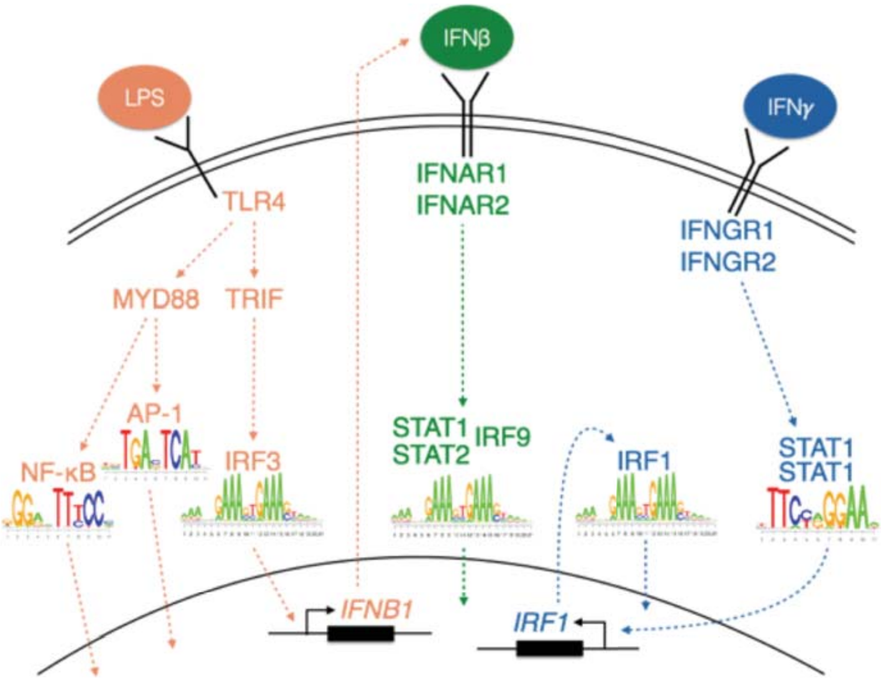

Figure 5.1: Main signalling pathways activated in macrophages after *Salmonella* infection and IFNɣ stimulation. Macrophages recognise LPS on the *Salmonella* cell wall via the TLR4 receptor that leads to the downstream activation of the nuclear factor kappa B (NF-κB) and activator protein 1 (AP-1) (Takeuchi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

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

5.2 Methods

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

5.2.1 ATAC-seq

Experimental procedures

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

Read alignment

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

Peak calling

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

Sample quality control

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

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

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

Differentially accessible regions

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

Motif enrichment

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

5.2.2 ChIP-seq data analysis

The public ChIP-seq datasets used in this study are summarised in section 'Summary of public ChIP-seq datasets used in the analyses'. Single-end datasets (Pham *et al* and Qiao *et al*) were aligned to the GRCh38 human reference genome using bwa aln v0.7.12 with default parameters. Paired-end datasets (Reschen *et al*, Schmidt *et al* and Wong *et al*) were aligned to the GRCh38 reference genome using bwa mem v0.7.12 with the -M flag set. Only properly paired reads were used for downstream analysis. Duplicate reads were removed with Picard v1.134 MarkDuplicates with the ‘REMOVE_DUPLICATES=true’ parameter set. I used bedtools v2.17.0 (Quinlan and Hall, 2010) to construct genome wide read (single-end) or fragment (paired-end) coverage tracks in BigWig format. I called peaks using MACS2 v2.1.0 with ‘-q 0.01’ option.

Summary of public ChIP-seq datasets used in the analyses

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

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

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

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

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

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

Detecting regions with differential H3K27Ac signal

I performed differential histone acetylation analysis on the Qiao et al (Qiao et al., 2013) dataset to compare it to our ATAC-seq data. As H3K27Ac peaks are generally broader than ATAC-seq peaks, I used MACS2 to call both broad and narrow peaks. Within each condition I only kept broad and narrow peaks that were detected at the 1% FDR threshold in both biological replicates. By visualising the data in a genome browser, I observed that at the 1% FDR threshold MACS2 called an excess of broad peaks compared to the narrow peaks so I further removed broad peaks that did not overlap any narrow peaks in the same condition. I then defined the union of broad peaks identified in each condition as the consensus set of peaks. I used featureCounts (Liao et al., 2014) to count the number of reads overlapping the consensus peaks in each sample. Finally, I used limma voom (Law et al., 2014) to identify peaks that showed at least 2-fold differential histone acetylation between naive and one of the stimulated states at 10% FDR. I used less stringent fold change and FDR thresholds for the histone acetylation data compared to the ATAC-seq data, because the broad histone peaks were less dynamic than the narrow ATAC peaks and because the histone dataset had only two biological replicates.

Peak overlap analysis

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

5.2.3 Chromatin accessibility QTL mapping

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

Filtering condition-specific caQTLs by effect size

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

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

QTL replicability between conditions

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

Motif disruption analysis

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

Identifying condition-specific dependent peaks

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

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

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

Similarly to condition-specific caQTL analysis, I used the first three principal components calculated separately for each condition as covariates in the model. I used the $\log_2$FC from RASQUAL as the measure of caQTL effect size. To identify true condition-specific dependent peaks, I further filtered the results by requiring the absolute $\log_2$FC of the master peak to be > 0.59 (1.5-fold) in the naive condition and the change in the $\log_2$FC for the dependent peak between the naive and stimulated condition to be > 0.59.

5.3 Quantifying chromatin accessibility

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

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

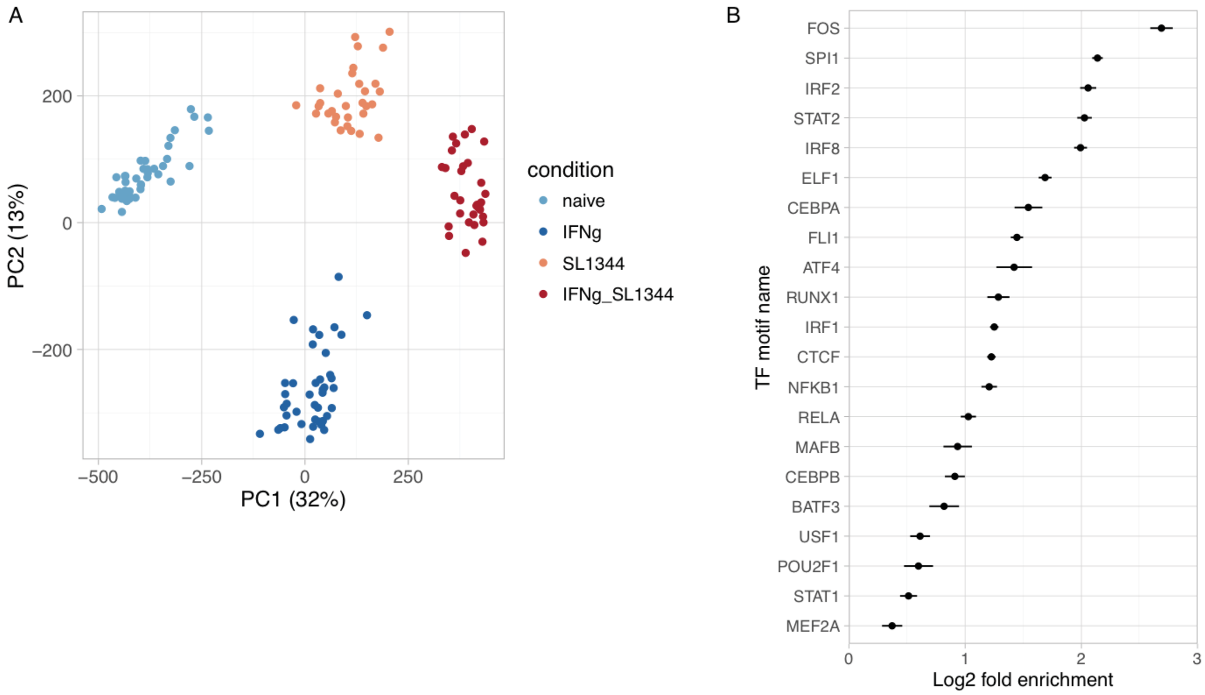

Figure 5.2: Summary of chromatin accessibility data. (A) PCA of macrophage chromatin accessibility data in four conditions. Axis labels indicate the percentage of variance explained by the first two principal components. **(B)** A selection of 21 representative TF motifs that are enriched in macrophage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

5.3.1 Differential chromatin accessibility between conditions

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

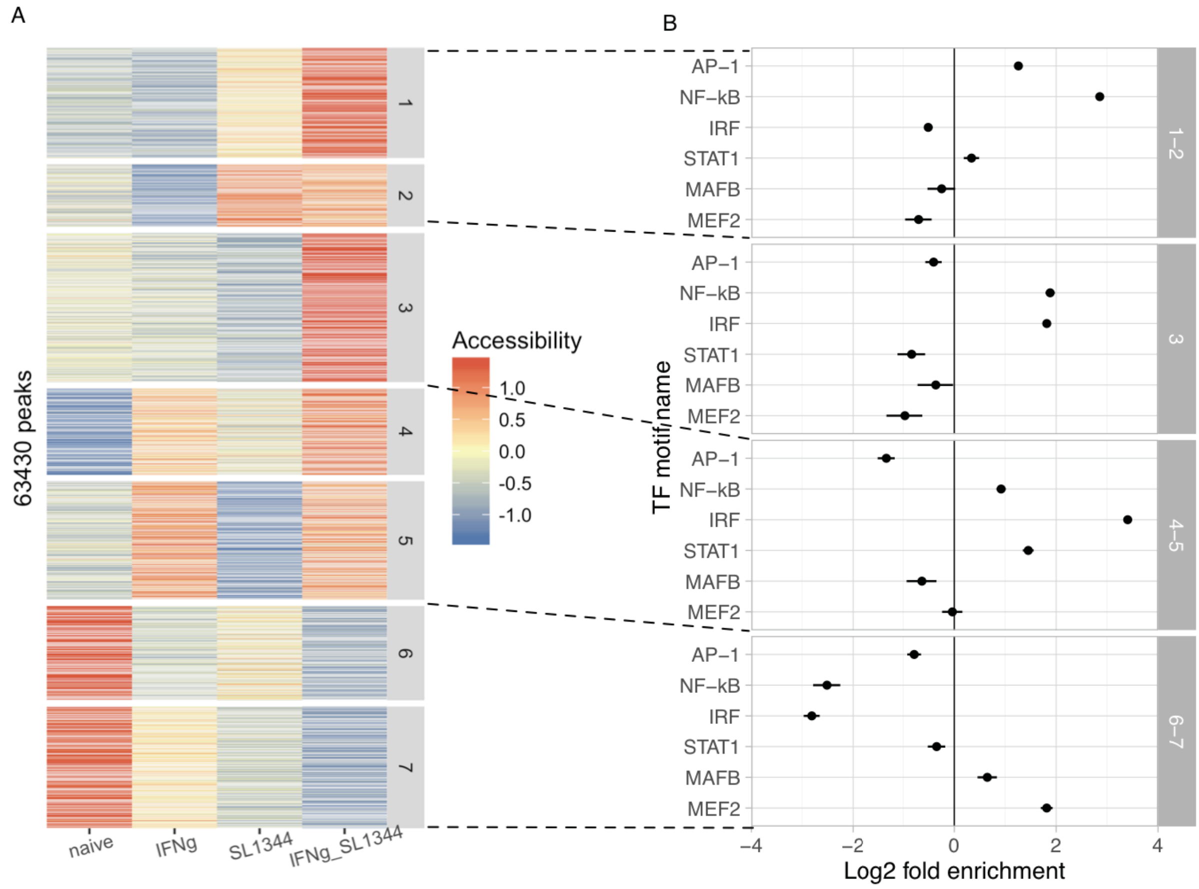

Figure 5.3: Dynamics of chromatin accessibility between conditions. (A) The 63,350 differentially accessible open chromatin regions were clustered into seven distinct patterns using c-means clustering implemented in the MFuzz packages. The clusters have been grouped into four groups according to whether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

Clusters 1 and 2, both of which became more accessible after *Salmonella* infection, were specifically enriched for NF-κB and AP-1 motifs, the two main TFs activated downstream of TLR4 signalling (Takeuchi and Akira, 2010). Cluster 3, which became accessible only after both of the stimuli were present, was enriched for the IRF (ISRE) and NF-κB motifs, suggesting possible collaborative interactions between IFNɣ-induced IRF1 and TLR4-activated NF-κB TFs that have been previously reported (Negishi et al., 2006). However, the motif analysis that I have performed does not distinguish between IRF1 and other IRF factors, because all IRF factors have similar sequence preferences. In contrast, clusters 4 and 5 were activated by IFNɣ and were enriched for IRF and STAT1 motifs, consistent with the activation of STAT1 and IRF1 downstream of IFNɣ signalling (Schroder et al., 2004).

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

5.3.2 Overlap with ChIP-seq signals

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

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

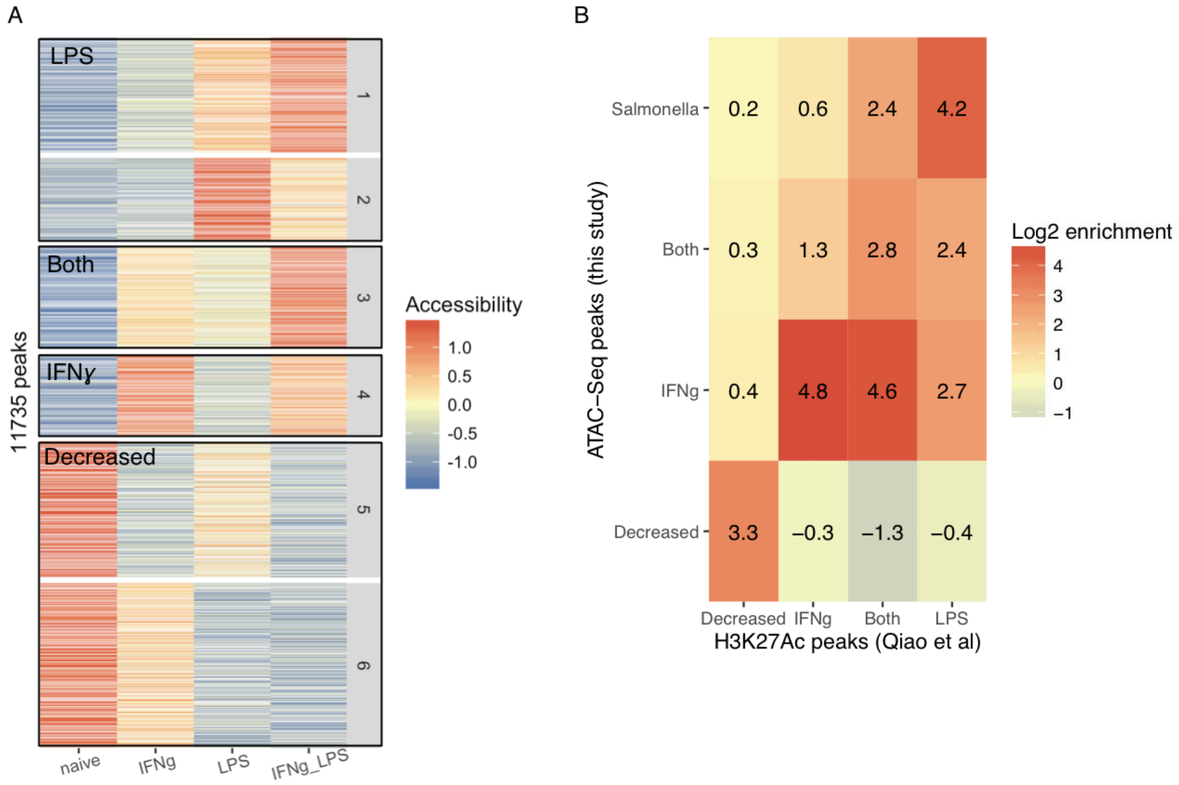

Figure 5.4: Concordance of chromatin changes between IPSDMs and MDMs. **(A)** Clustering of differential H3K27Ac peaks from (Qiao et al., 2013) study. The six clusters identified by MFuzz have been grouped into four groups based on whether the H3k27ac signal increases after LPS stimulation, IFNɣ stimulation,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

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

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

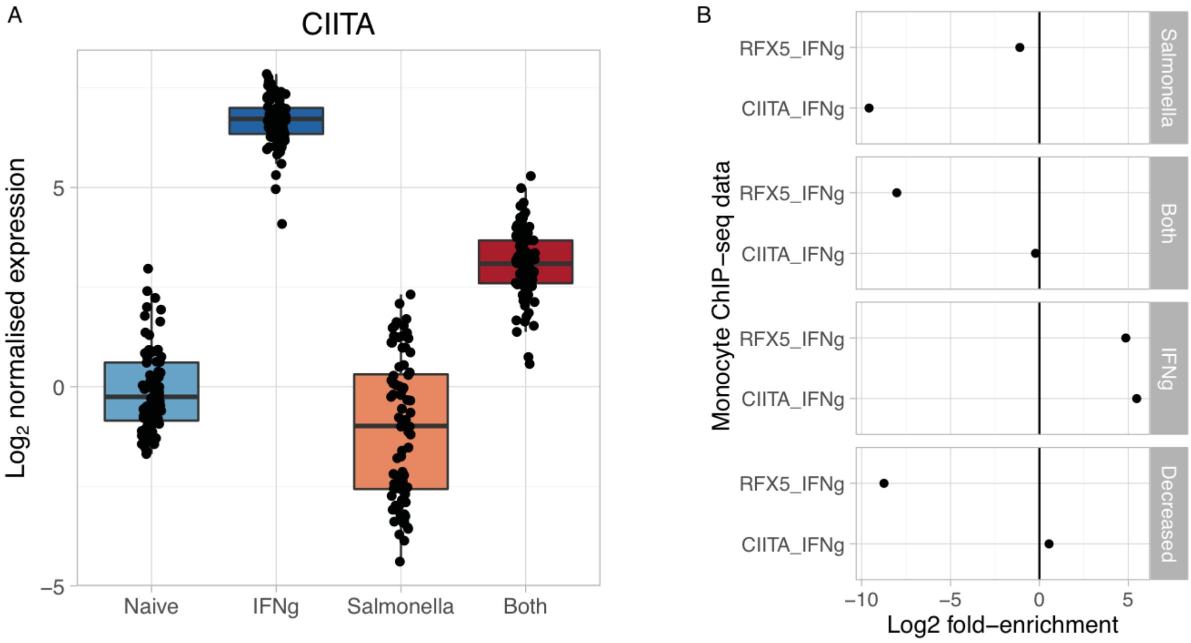

Figure 5.5: Regulation of MHC class II expression in IPSDMs. (A) Expression level of CIITA gene in IPSDMs in the four conditions. **(B)** Enrichment of monocyte RFX5 and CIITA ChIP-seq peaks (Wong et al., 2014) in IPSDM ATAC-seq peak clusters from Figure 5.3A.

5.4 Genetics of chromatin accessibility

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

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

Salmonella	31	5267	5493
Both	31	3782	4337

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

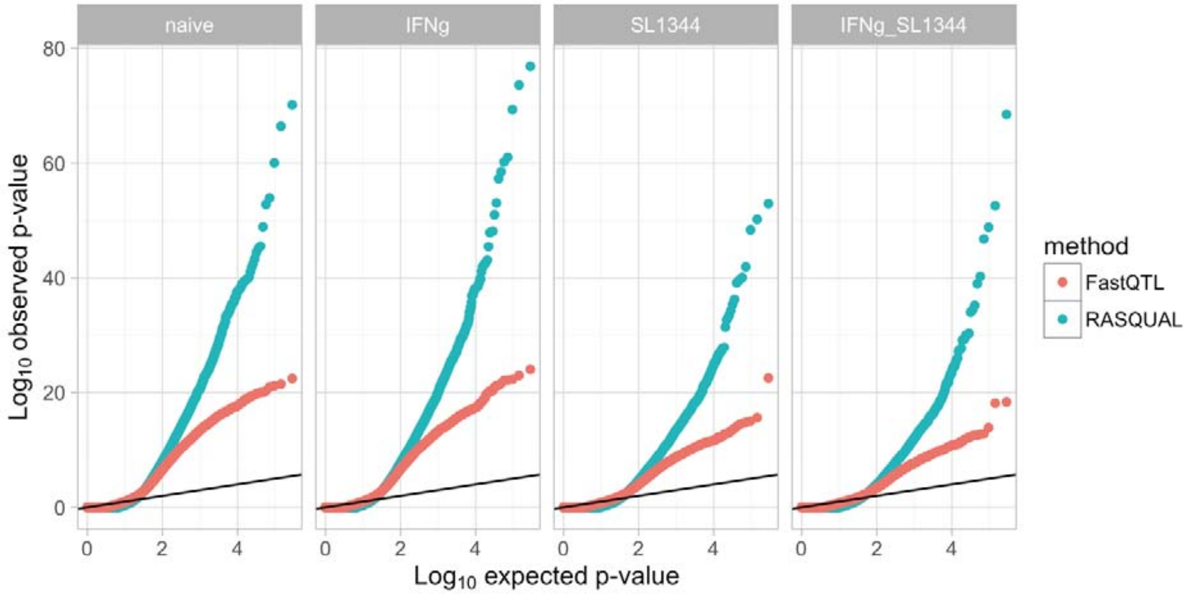

Figure 5.6: Q-Q plots of caQTLs identified by RASQUAL and FastQTL in each of the four conditions. On each plot, the solid line corresponds to the expected distribution of p-values under the null model of no association. The FastQTL and RASQUAL p-values have been corrected for the number of independent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

5.4.1 Fine mapping putative causal variants

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

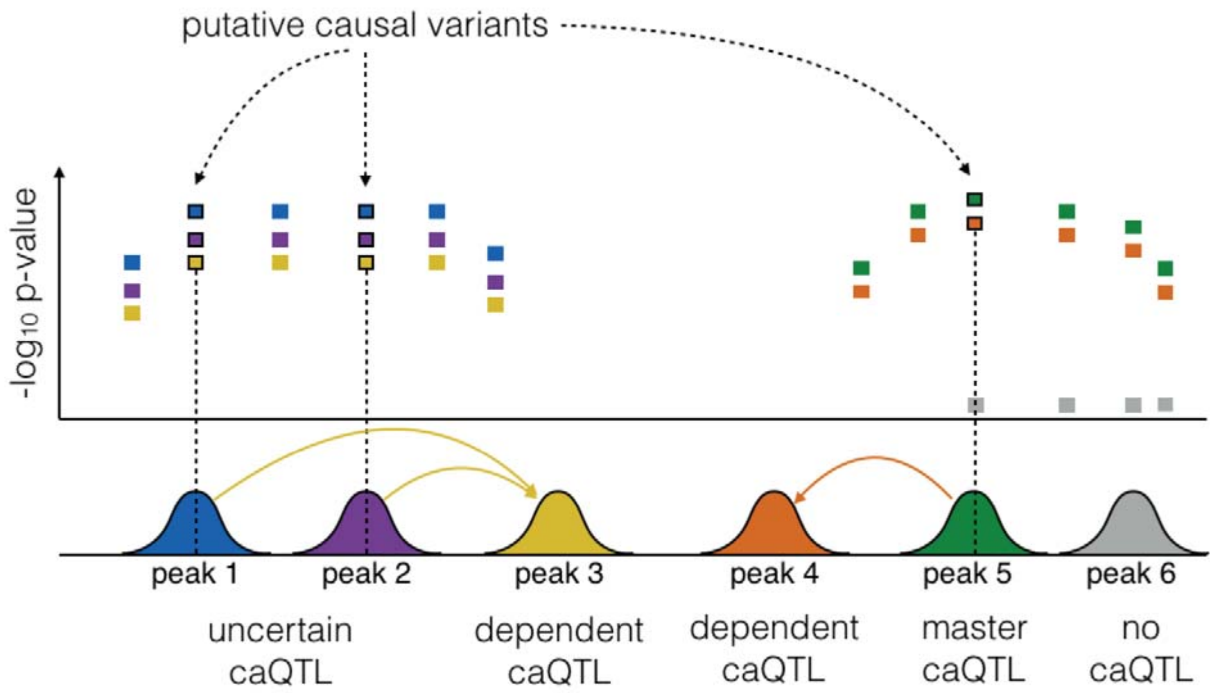

Figure 5.7: Heuristic approach to identify master and dependent caQTLs and their putative causal variants. Peak 5 is a *master caQTL peak*, because all of the variants in its credible set (green squares) overlap only peak 5 and no other caQTL region. Peak 1 and 2 are uncertain caQTLs, because the credible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

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

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

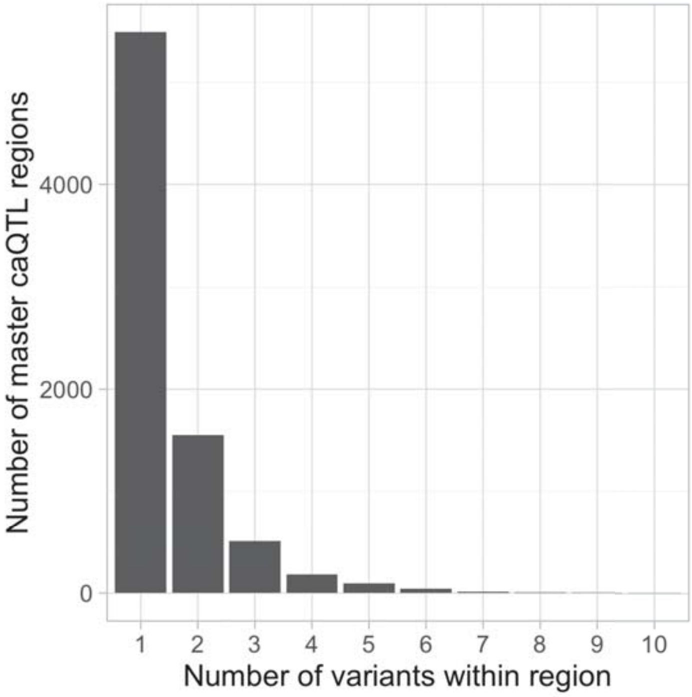

Figure 5.8. Histogram of the number of associated variants overlapping 7,903 putative master caQTL peaks.

Next, to identify dependent peaks, I focussed on the 1,840 caQTL peaks whose credible sets did not overlap the region itself. I found that for 753/1,840 peaks the credible set overlapped one of the putative master caQTL peaks identified above. This suggests that ~10% of the putative master caQTLs regions also have a dependent caQTL. However, this is likely to be an underestimate, because dependent caQTLs generally have smaller effects than master caQTLs and we are less powered to detect them with our small sample size.

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

5.4.2 Assessing condition-specificity of caQTLs

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

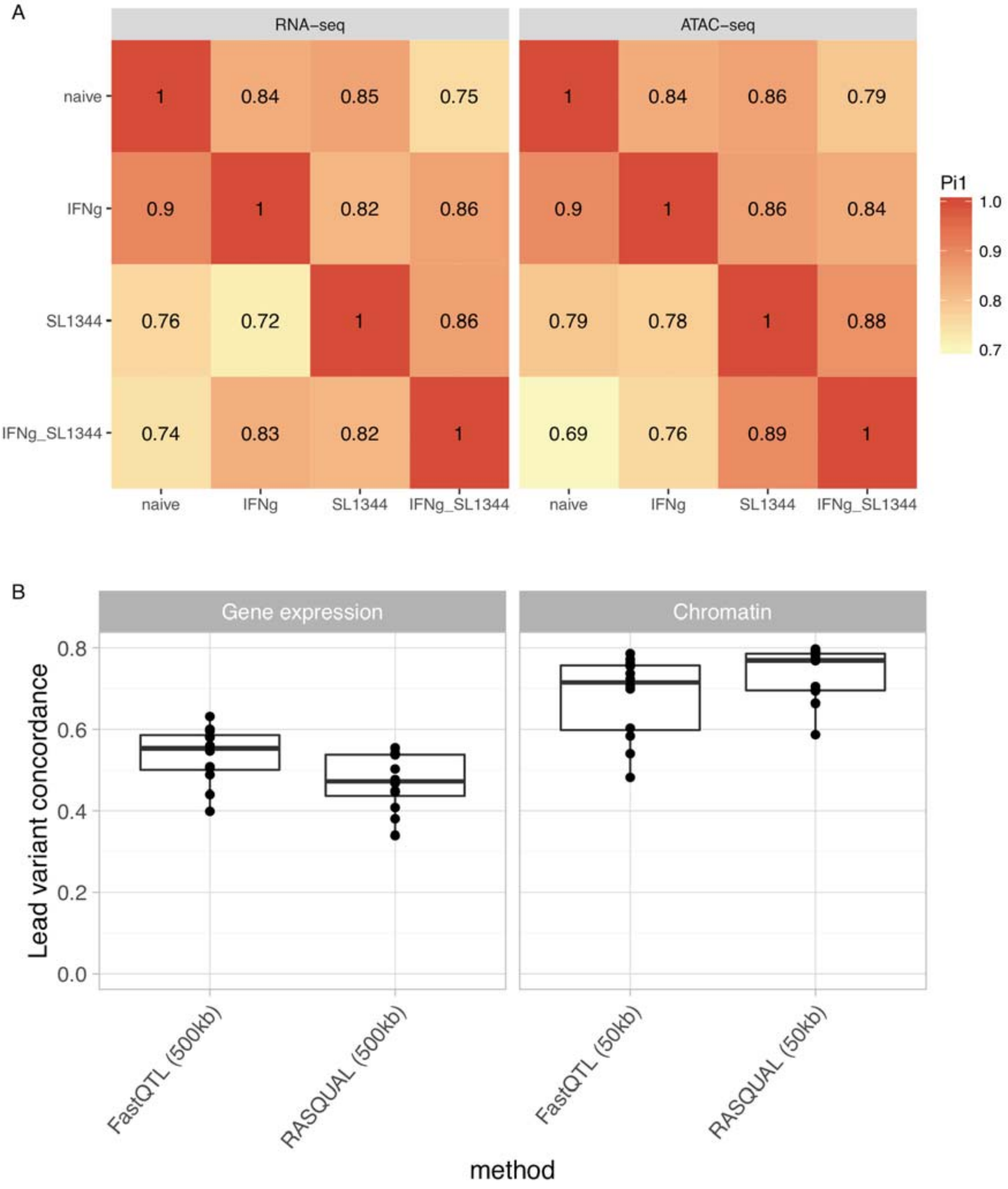

Figure 5.9: Replicability of eQTLs and caQTLs between conditions. (A) Feature-level replicability between conditions using the Storey's π_1 statistic. The π_1 statistic was calculated based on the FastQTL permutation p-values. **(B)** Pairwise concordance of the lead eQTL and caQTL variants for each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

To identify individual peaks that have condition-specific caQTLs, I compiled all independent (R^2 < 0.8) variant-peak pairs across conditions and used two-way ANOVA to test for interactions between genotype and condition. Using sex and first three principal components of the dataset as covariates, I found that 4,947/16,924 (28%) caQTLs had significant interactions. After filtering out interactions with small effects, I identified 1,990 highly condition-specific caQTLs of which 1,113 appeared after stimulation ($\log_2 FC_{naive} < 1$) and 887 disappeared after stimulation ($\log_2 FC_{naive} > 1$).

I then clustered the condition-specific caQTLs based on their relative $\log_2$FC across conditions. For the caQTLs that appeared after stimulation, I identified six distinct clusters of peaks (Figure 5.10A). I then tested if the likely causal variants for the condition-specific caQTLs were enriched for disrupting specific TF binding motifs compared to all caQTLs (Figure 5.10B). For this analysis I focussed only on the unique master peaks identified in Section 5.1 that had 1 to 3 likely causal variants overlapping the peak. I found that *Salmonella*-specific clusters 2 and 3 were enriched for disrupting NF-κB and AP-1 motifs whereas IFNɣ-specific clusters 5 and 6 were enriched for disrupting the ISRE motif. Furthermore, all condition-specific caQTLs were depleted for disrupting PU.1 binding motif (Figure 5.10B). This analysis suggests that condition-specific caQTLs are at least partly driven by variants that disrupt the binding sites of condition-specific TFs that are not active in the naive state. However, despite observing these motif enrichments, only ~15% condition specific caQTL could be explained by a motif disruption event at the thresholds that I used. Interestingly, I observed that almost all condition-specific caQTL peaks on Figure 5.10A were completely inaccessible in the naive condition and became most accessible in the condition with the largest caQTL effect size (Figure 5.11B). On the other hand, we observed no such relationship in the gene expression data where the genes with condition-specific eQTLs were on average equally highly expressed in all four conditions (Figure 5.11A).

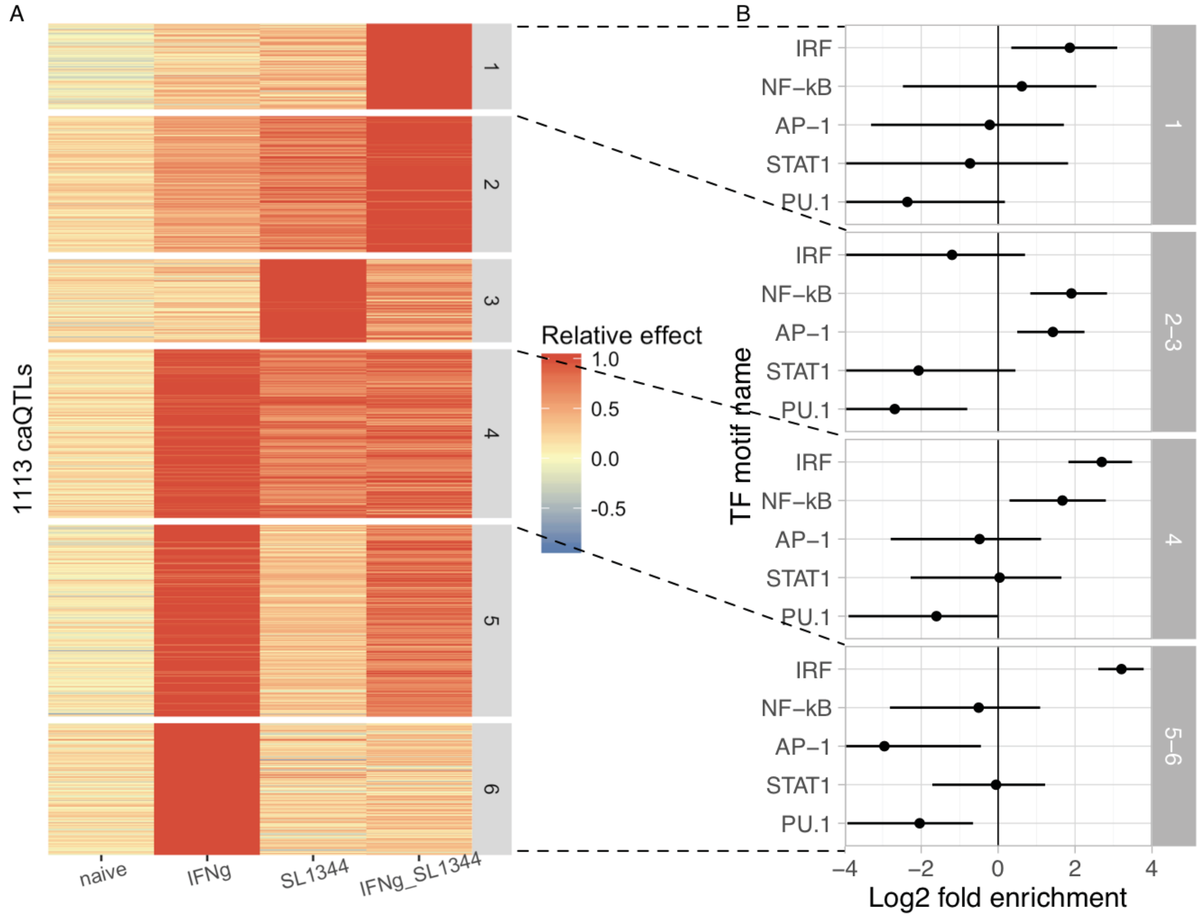

Figure 5.10: Identifying condition-specific caQTLs. (A) Condition-specific caQTLs clustered by their relative effect size. **(B)** Enrichment of TF motif disruptions in each cluster of caQTLs. The six cluster were grouped into four groups based on the caQTL activity pattern.

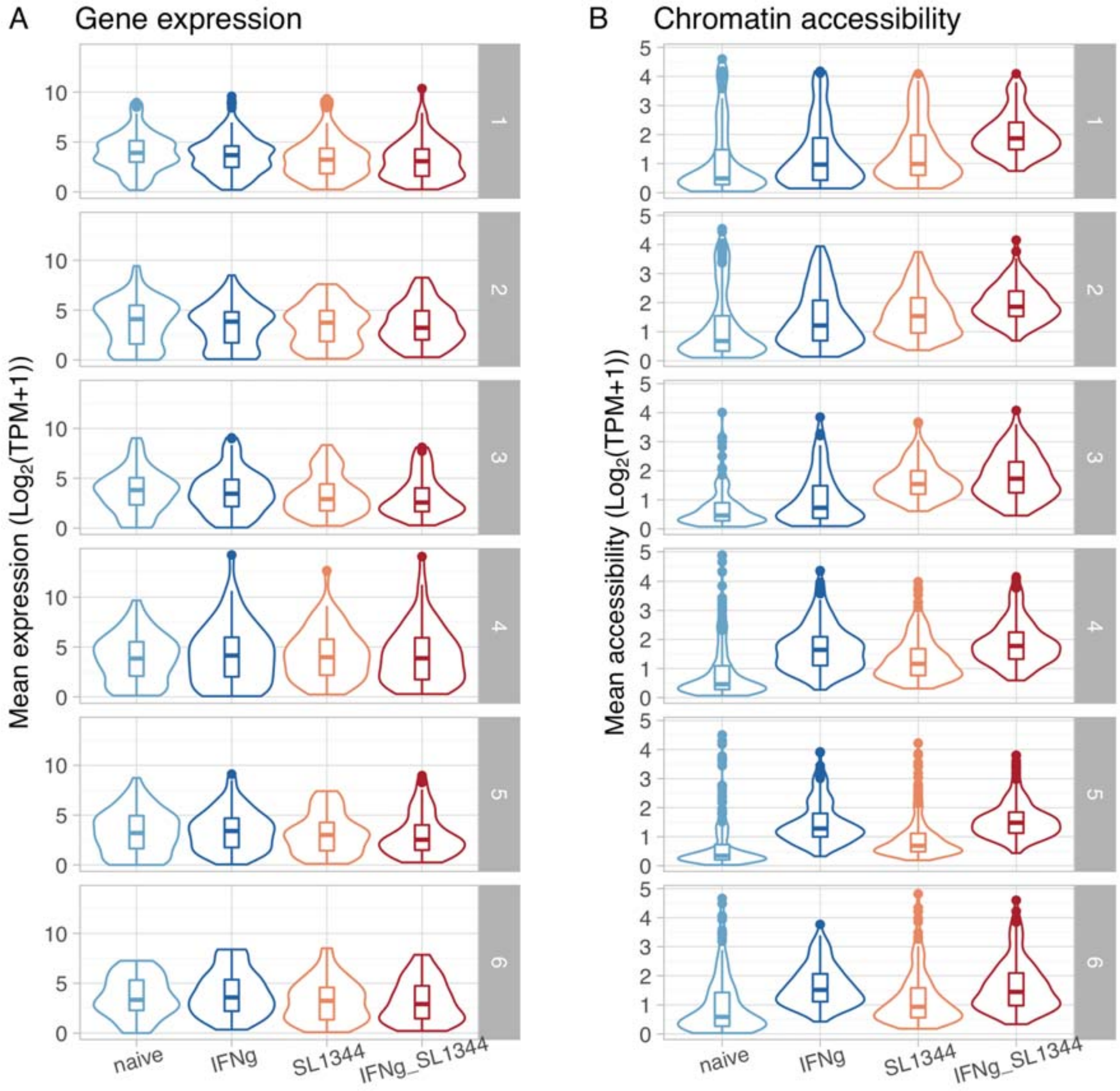

Figure 5.11: Relationship between QTL condition-specificity and mean gene expression or chromatin accessibility in each of the four conditions. (A) The distribution of mean gene expression values in each condition for the genes with conditions specific eQTLs from Figure 4.11C in Chapter 4. The numbered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

5.4.3 Condition-specific dependent peaks

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

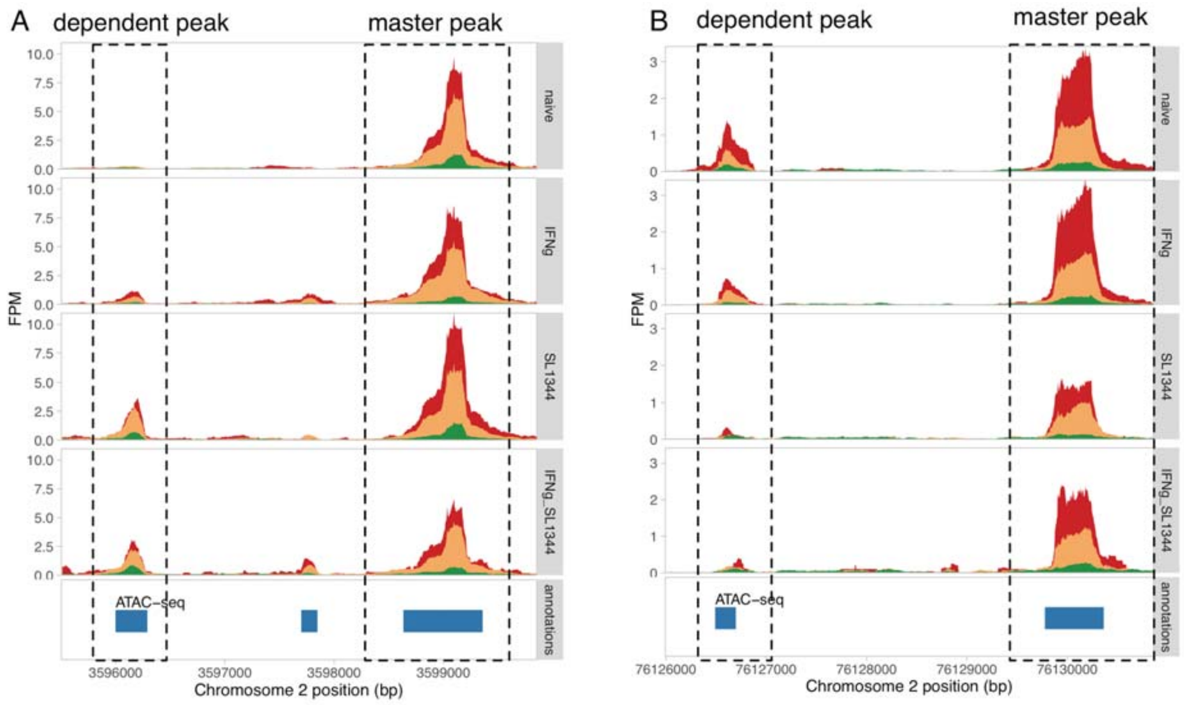

Figure 5.12: Two examples of condition-specific dependent peaks. (A) Dependent peak appears after *Salmonella* infection. **(B)** Dependent peak disappears after *Salmonella* infection.

5.5 Linking chromatin accessibility to the transcriptome

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

5.5.1 Linking caQTLs to eQTLs

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

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

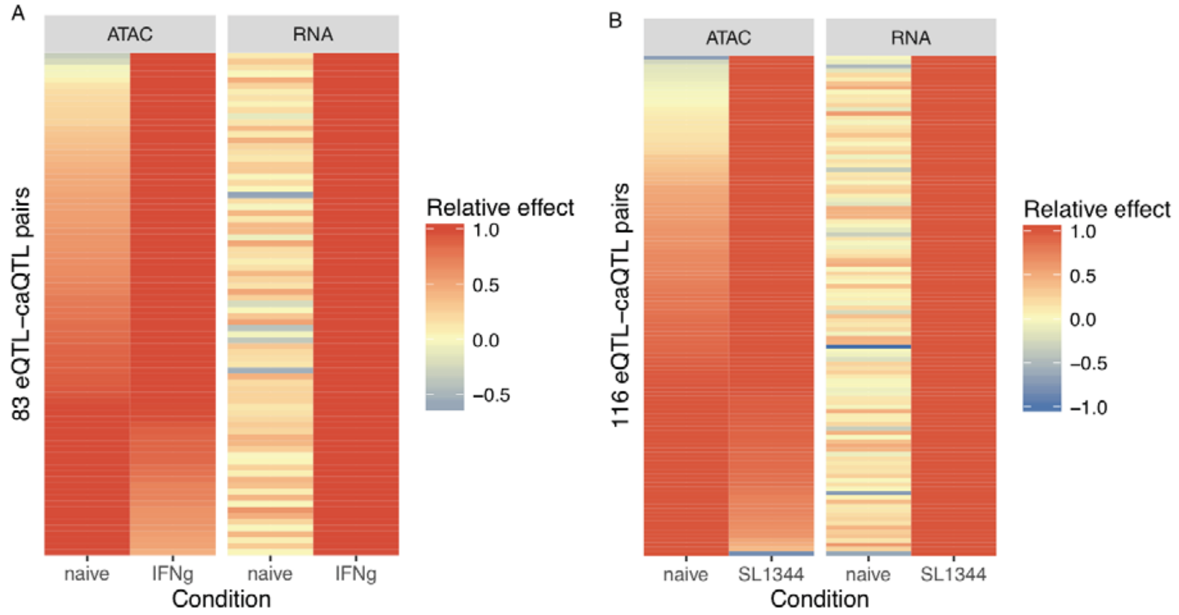

Figure 5.13: Comparison of effect sizes between condition-specific eQTLs and their corresponding caQTLs. (A) IFNɣ-specific eQTLs and their corresponding caQTLs. **(B)** *Salmonella*-specific eQTLs and their corresponding caQTLs.

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

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

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

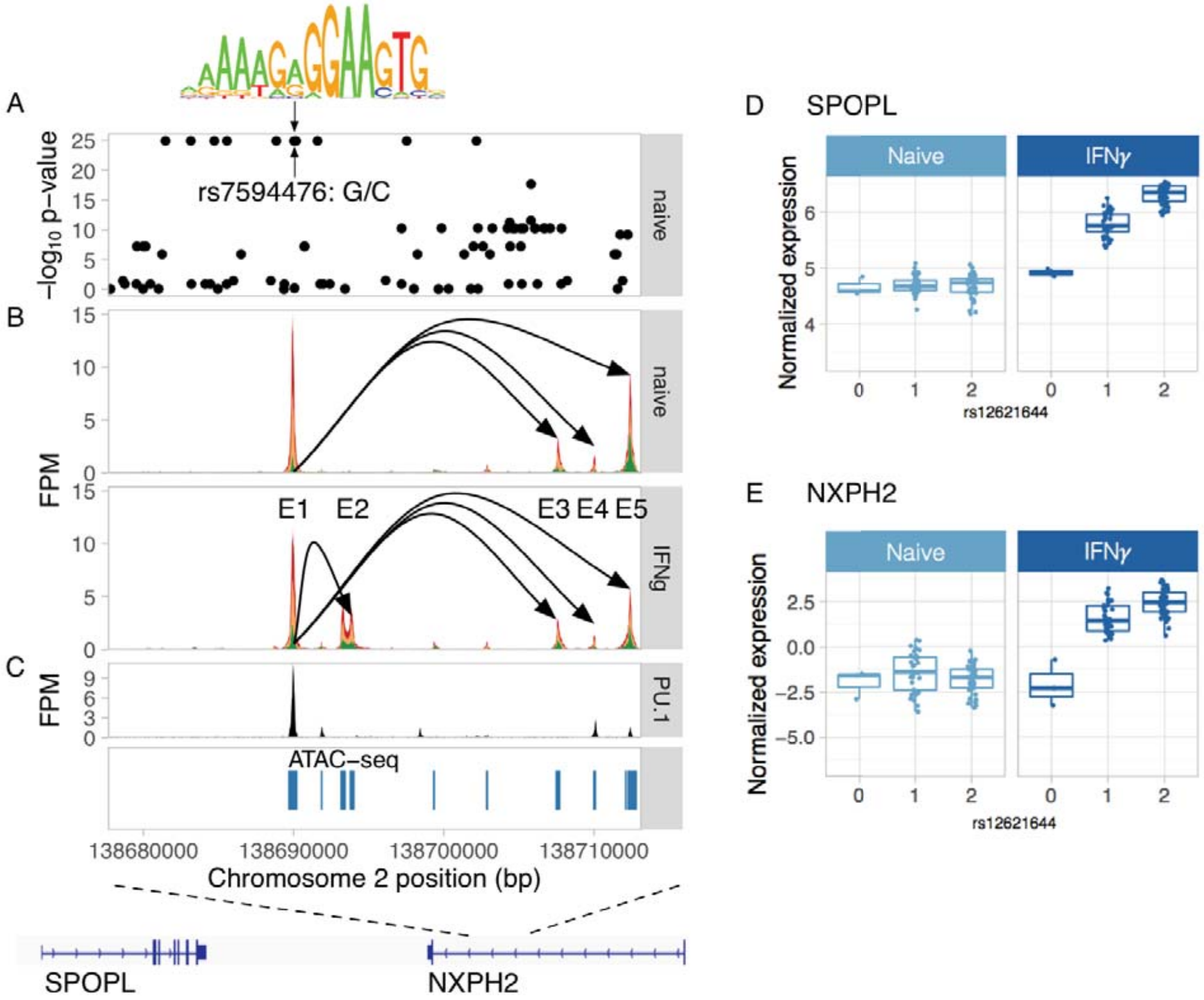

Figure 5.14: Example of a single QTL that influences chromatin accessibility at multiple peaks and the expression of two genes. **(A)** Manhattan plot of variants associated with the accessibility of the master caQTL peak E1. Only two of the associated variants overlap the E1 peak, and only rs7594476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

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

5.5.2 Using caQTLs to fine map causal variants for GWAS hits

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

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

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

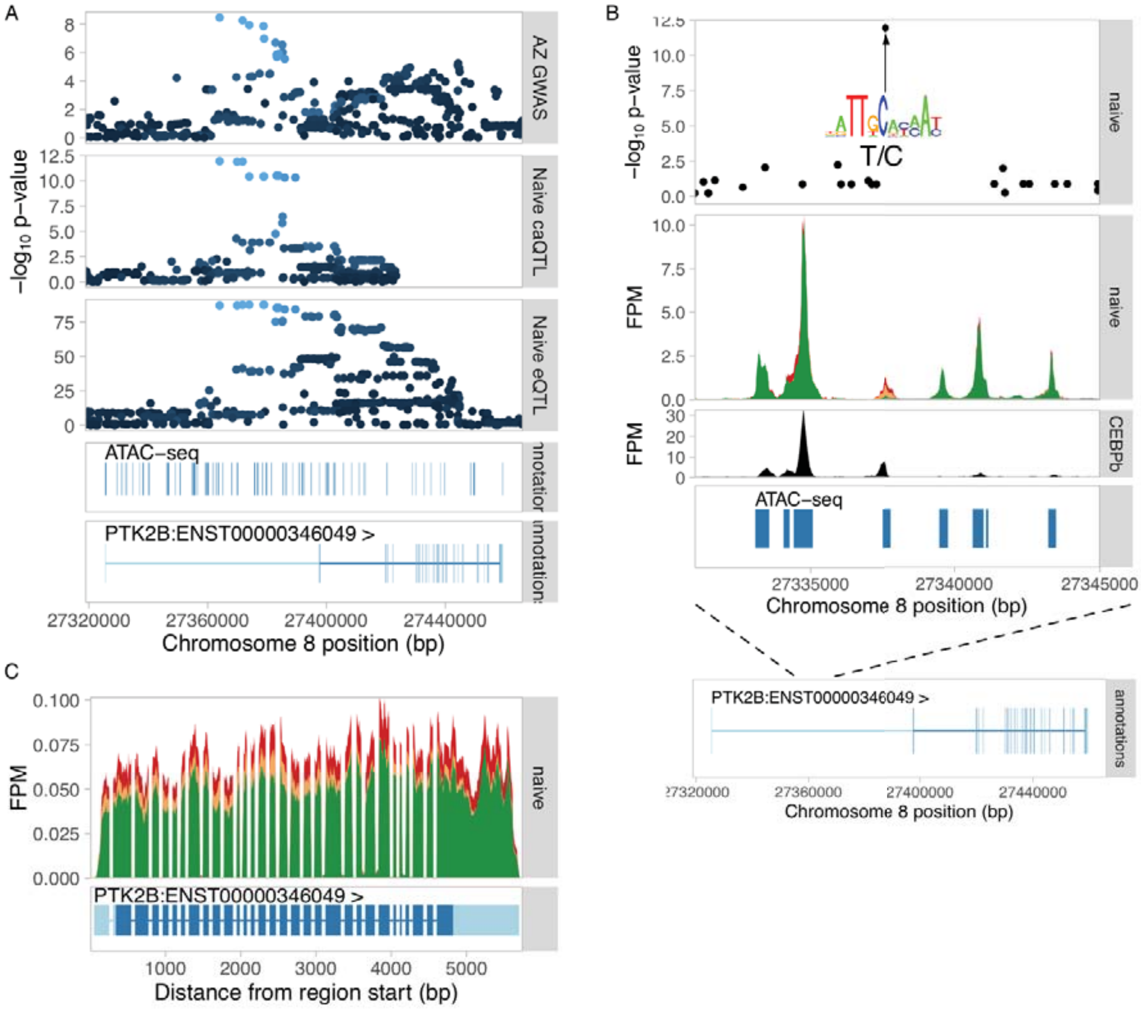

Figure 5.15: Dissecting the Alzheimer's disease causal variant at the PTK2B locus. (A) Manhattan plots for the Alzheimer's GWAS hit (top panel), colocalised caQTL (second panel) and colocalised eQTL for PTK2B gene (third panel). The bottom two tracks show all ATAC-seq peaks in the region as well was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

5.6 Discussion

We have shown that, similarly to gene expression, (Chapters 2 and 4), the chromatin accessibility dynamics of IPSDMs also closely resemble that of primary macrophages. Evidence for this comes from the motif enrichment analysis where constitutive and condition-specific macrophage ATAC peaks were enriched for expected macrophage-specific TF motifs such as PU.1, AP-1, NF-κB, STAT1 and ISRE motif representing multiple IRF factors. Secondly, overlap analysis with multiple public ChIP-seq datasets confirmed that overlapping regions changed their activity in IFNɣ and LPS response. Future studies where IPSDMs and MDMs are measured in the same experiment are needed to reliable detect any differential chromatin accessibility between the two cell types and identify TFs responsible for those differences.

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

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

Similarly to previous studies (Grubert et al., 2015; Kumasaka et al., 2016; Waszak et al., 2015), we also found widespread evidence of single caQTL variants regulating the accessibility of multiple dependent caQTL peaks, often multiple kb away from the master peak. In total, we were able to detect at least one dependent caQTL peak for ~10% of the master caQTL peaks, although this number is likely to increase with larger sample sizes. Importantly, measuring chromatin accessibility in multiple conditions allowed us to also identified a small number of dependent peaks that appeared or disappeared with stimulation. A number of those occurred in the SPOPL-NXPH2 locus (Figure 5.14), where the appearance of dependent caQTL peaks correlated with lead variant also becoming an eQTL for the two genes. This is consistent with a recently established model of hierarchical enhancer activation where signal-dependent transcription factors bind at or near primed enhancers to activate gene expression (Heinz et al., 2013; Romanoski et al., 2015).

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

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