Chapter 4

T-cell expression quantitative trait loci maps in Primary sclerosing cholangitis

4.1 Introduction

Colocalisation of GWAS risk loci with eQTLs provides a powerful way to indentify the functional role of the numerous non-coding risk loci by assigning molecular function to them. As shown in the previous chapter, colocalisation using published eQTL datasets for a variety of immune cell types and tissues has enabled the identification the genes perturbed by six of the studied PSC risk loci. The failure to identify the genes underlying the remaining risk loci may result, in part, from the failure to identify genetic variants that regulate gene-expression in cell-types and states relevant to PSC.

Whilst many eQTLs are shared across multiple tissues, some remain highly specific to a particular cell type, tissue, environment or activation state [130]. One of the ongoing challenges is to identify the correct cell-type or tissue in which to map eQTLs for colocalisation with GWAS risk loci. Indeed, it has been shown that when trying to unravel the molecular basis of disease-specific risk loci, the choice of disease-relevant tissues supports the finding of eQTLs enriched for disease-associated variants [190, 191]. However, colocalisation analysis remains limited by the availability of published eQTL summary statistics. Furthermore, since PSC is a rare disease, there are currently no published eQTL studies of the cell types perhaps most relevant to PSC, in the environments most relevant to PSC. Therefore eQTL mapping in PSC-specific cell types, in PSC-specific environments, is of great scientific interest.

Identification of the cell types of most potential relevance to the causal pathogenesis of a disease relies upon existing knowledge of disease pathogenesis, which unfortunately, in PSC remains limited. As with many immune-mediated diseases (IMDs), T-regulatory cells

have been implicated in the pathogenesis of PSC, not least supported by the finding of two PSC risk loci near the *IL2RA* and *IL2/IL21* genes, the protein products of which are expressed or involved in pathways of T-regulatory cells. Histological observations provide further evidence of potentially relevant cell types. PSC is histologically characterised by a T-cell rich portal infiltration with peri-ductal inflammation, portal fibrosis and progressive loss of the bile ducts, known as ductopenia [226]. Moreover, evidence for potentially relevant cell types comes from the strong link with IBD, which is present in 50-70% of patients with PSC [23]. The liver and colon are anatomically linked with 75% of the blood supply to the liver originating from the intestine via the portal vein. In PSC, it has been shown that 20% of liver-infiltrating lymphocytes express gut-specific ligands CCR9 and α4β7. The 'gut-homing T-cell hypothesis' suggests that these CCR9+ memory T-cells are originally activated by inflammation within the gut and are recruited to the liver due to the observed aberrant inflammation-induced expression of their receptors MAdCAM-1 and CCL25 [53, 79]. In support of this, the vast majority of these CCR9+ liver-infiltrating T-lymphocytes in PSC are CD45RA+ CCR7+CD11a(high) and secrete IFN-γ , in keeping with an effector-memory phenotype. After recruitment to the liver, Grant *et al* proposed that CCR9+ and α4β7+ gut-derived lymphocytes are likely to use other chemokines such as CXCL12 and CXCR6 to localise to the biliary epithelium where they mediate targeted inflammation of the bile ducts. To date, no existing studies have mapped eQTLs in any of the aforementioned cell types. Therefore, some of the most potentially relevant cell types for the focus of future eQTL mapping efforts in PSC include the CD4+ and CD8+ effector-memory T-cells with the CCR9+ phenotype. Furthermore, one of the means of evaluating cells in the PSC-specific activated state, most representative of the active disease transcriptional phenotype, is to derive those cells from individuals with the active inflammatory condition.

4.2 Chapter Overview

Many studies have sought to map genetic variants associated with quantitative changes in gene expression in order to assign molecular function to non-coding disease risk loci via colocalisation. However eQTLs are known to be specific to both tissue type and activation state. Thus, one means of better understanding the genetic risk loci associated with susceptibility to PSC is to explore eQTL maps specific to the tissues and activation states of the disease. In this chapter, I describe the generation of eQTL maps in six peripheral blood T-cells subtypes, currently hypothesised to be important in the causal pathogenesis of PSC. These cells are derived from patients with active PSC and the highly co-morbid condition, UC. I describe the entire study process from patient recruitment to sample preparation and RNA sequencing analysis. I perform differential gene expression

analysis based on disease status. I map eQTLs for each cell type and identify those shared across several T-cell subtypes and those specific to an individual T-cell subtype. Finally, I perform colocalisation with genetic risk loci for PSC, IBD and other immune-mediated diseases (IMDs) in order to identify the genes perturbed by disease-associated risk loci.

4.3 Methods

4.3.1 Sample type and Patient recruitment

The PSC-specific cell-types chosen for analysis in this study were; T-regulatory cells (T-regs), non-activated memory T-cells (T-mems) and activated CD4+ and CD8+ effector-memory T-cells that are positive and negative for the gut-homing ligand, CCR9 (CD4+CCR9-, CD4+CCR9+, CD8+CCR9-, CD8+CCR9+) [53]. The surface marker phenotype of each cell subtype is shown in Table 4.1. I aimed to recruit a total of 80 patients for this study, based upon evidence that previous studies with similar numbers of individuals have identified eQTLs. For example, the GTEx Consortium pilot study of post mortem tissues was able to detect tissue-specific quantitative genetic traits for a median sample size of 105 for the 9 high-priority tissues [176]. Furthermore, the HapMap study of genetic variants underlying variation in gene expression detected an abundance of *cis*-regulatory variants in the human genome with a median sample size of just 40 individuals in each population group [120]. However, PSC is a rare disease with a prevalence of 1 in 10,000 and there are predicted to be just 7,000 patients living with PSC in the UK. Due to the rarity of PSC it is therefore difficult to recruit large numbers of PSC patients with a homogenous, active, disease phenotype. To address this difficulty, I aimed to recruit a total of 80 patients, 40 with PSC and concomitant UC and a further 40 with UC alone. Both PSC-UC and UC patients harbour increased numbers of CCR9+ effector-memory T-cells that have been activated in the inflamed colon [78, 80], and thus this combined cohort would facilitate a sample size large enough to detect eQTLs.

I recruited patients for this study from the Autoimmune liver disease clinic in the Department of Gastroenterology at the Norfolk and Norwich University Hospital. I was granted prior ethical approval for the study by the Norfolk and Norwich University Hospital Human Tissue Bank (reference number: 20122013-57 HT). Given that the ultimate aim of this study was to perform colocalisation with loci associated with risk of PSC in European populations, all patients were of white European ancestry. In order to minimise immune influences on the transcriptome, patients on steroids or biologic therapy, as well as those with previous cancer diagnoses, were excluded. In addition, given that one of the important cell types under investigation was the CCR9+ effector-memory T-cell activated within the inflamed colon, patients with previous colectomy were also excluded. Finally, all recruited

Table 4.1: Fluorochrome-labelled antibody panel defining six subtypes of T-cell by FACS

Cell type	Abbreviation	Antibody panel
T-regulatory cells	T-reg	CD3+CD4+CD25+CD45RO+CD127low
Memory T-cells (non-activated)	T-mem	CD3+CD4+CD45RO+CD25-
CD4+ CCR9- effector memory T-cells	CD4+CCR9-	CD3+CD4+CD62L-CD45RO+CD199-
CD8+ CCR9- effector memory T-cells	CD8+CCR9-	CD3+CD8+CD62L-CD45RO+CD199-
CD4+ CCR9+ effector memory T-cells	CD4+CCR9+	CD3+CD4+CD62L-CD45RO+CD199+
CD8+ CCR9+ effector memory T-cells	CD8+CCR9+	CD3+CD8+CD62L-CD45RO+CD199+

patients had a serum alkaline phosphatase raised above the reference range for the upper limit of normal, but no histological or radiological evidence of cirrhosis to ensure an active PSC transcriptome. A total of seventy-nine donors were recruited; forty-four with PSC and UC and thirty with lone UC. Five healthy controls (HC) for the pilot study set-up which were also included for analysis.

4.3.2 Sample preparation

I drew 50mls of peripheral blood from each donor, and processed this immediately at 4°C to prevent activation or degradation of cells. From whole blood, I separated peripheral blood mononuclear cells (PBMCs) over Ficoll and stained them with a fluorochrome labelled antibody panel designed to isolate the six T-cell subtypes, using three rounds of two-way sorting, as shown in Table 4.1. I sorted cells directly into chilled cell lysis buffer (Buffer RLT Plus, *Qiagen*) using a Sony SH800 fluorescent activated cell sorter (FACS). Samples were then immediately stored at -80°C. An example of the standard FACS gating strategy used is shown in Figure 4.2.

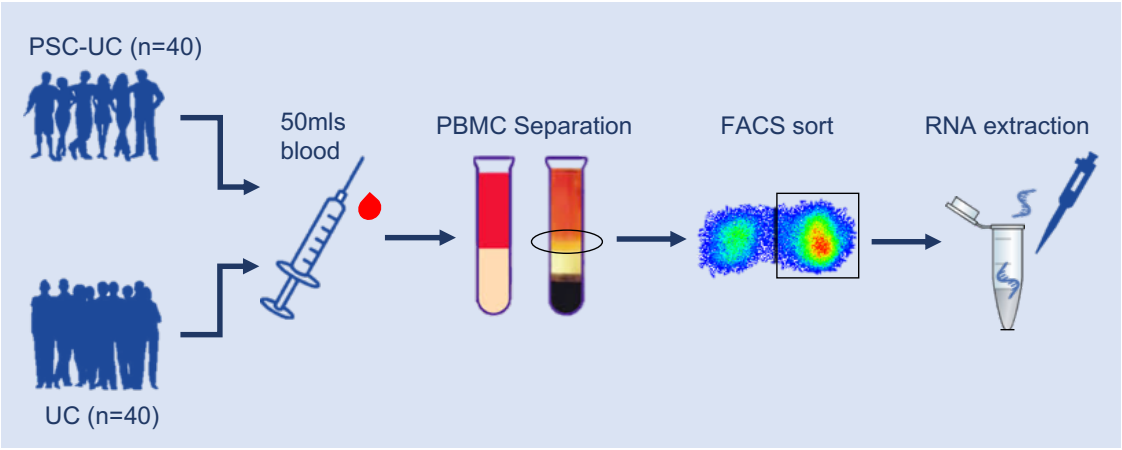

Figure 4.1: Sample preparation pipeline.

During the set-up phase, I verified a small subset of twelve samples (two of each cell type) to >95% purity by performing repeated FACS on already-sorted samples, using the

same gating strategy. To minimise cellular perturbation, I performed all cell sorting using a 100µm nozzle at low sorting pressures using chilled, preservative-free Hank's Balanced Salt Solution (HBSS). Maximum time from acquisition of the whole blood sample to freezing of lysed, FACS sorted, T-cell samples, was six hours. Technical failure of the cell-sorter calibration on two occasions resulted in the loss of all T-cell samples from three individuals (two with PSC-UC and one with lone UC). Therefore, in total 456 T-cell samples were isolated from 76 individuals; 42 with PSC-UC, 29 with UC and 5 healthy controls.

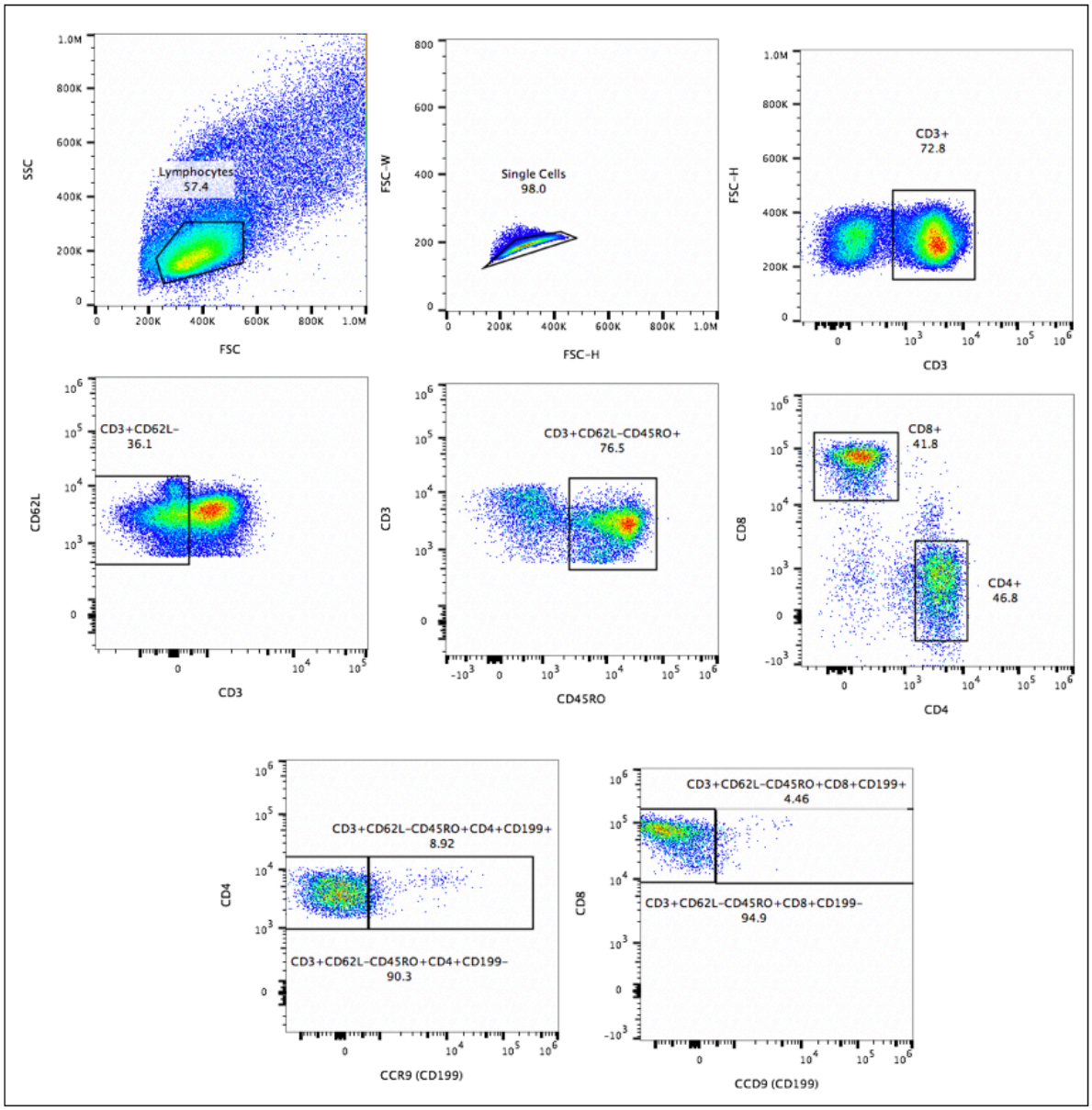

Figure 4.2: Gating strategy used for FACS separation of CD4+CCR9-, CD4+CCR9+, CD8+CCR9- and CD8+CCR9+ central effector T-cells from peripheral blood mononuclear cells.

4.3.3 RNA extraction, library preparation and sequencing

I sequenced six different cell-type samples from seventy-six donors giving a total of 456 libraries. I performed RNA extraction using the *Qiagen* RNAeasy Micro plus kit. I checked RNA concentration and quality on a 20% subset of samples (equally representative of all cell-types) using the Agilent 2100 Bioanalyser, confirming RNA integrity number (RIN)

of >8.0. All samples were then sent to the Wellcome Sanger Institute RNA Pipelines for library preparation and RNA sequencing. Library preparation was performed by Sanger Pipelines using NEBNext Ultra II Directional RNA kit, with a poly(A) pulldown using oligo d(T) beads. Samples were then sequenced using 75 base-pair, paired-end read sequencing, performed on the Illumina HiSeq 4000. Four plates, each containing 96 samples, were pooled at 96-plex and run over twelve lanes (eight samples sequenced per lane) and the fifth plate containing 76 samples was run at 76-plex across ten lanes (7.6 samples per lane). The expected number of reads per samples was ∼60 million reads.

4.3.4 Read alignment, counts and quality control

I aligned reads to the human genome and transcriptome, using *STAR* (Spliced Transcripts Alignment to a Reference) software [227] and the reference genome; Genome Reference Consortium Human Build 38 Release 29 (GRCh38.p12). This is a comprehensive reference transcriptome, which includes protein coding RNA, all known non-coding RNA, non-sense mediated decay transcripts, and both processed and unprocessed pseudogenes. The reference genome is however incomplete, particularly around centromeres, meaning that reads can be incorrectly mapped to other places within the genome (albeit with low quality) resulting in false positive calls. I therefore included decoy contigs, known true human genome sequence that is not included within the reference genome, to map reads that would otherwise map to other regions of the genome.

Read counts were assigned to genes using *FeatureCounts*, implemented in R [228]. For RNA samples, greater than 75% alignment of the total number of reads to the genome was considered successful [229]. Samples with less than 60% of reads aligned to the genome were immediately removed from the analysis, and those between 60-75% aligned initially retained, but ultimately excluded following further quality control (QC) steps described below. Across all samples, the mean proportion of the total reads mapping to exons was 0.79, with a median of 0.80. Samples with a proportion of exonic mapped reads less than 0.6 were also removed from the analysis. Following these preliminary QC steps, 6 T-cell samples were removed from the analysis (Figure 4.3).

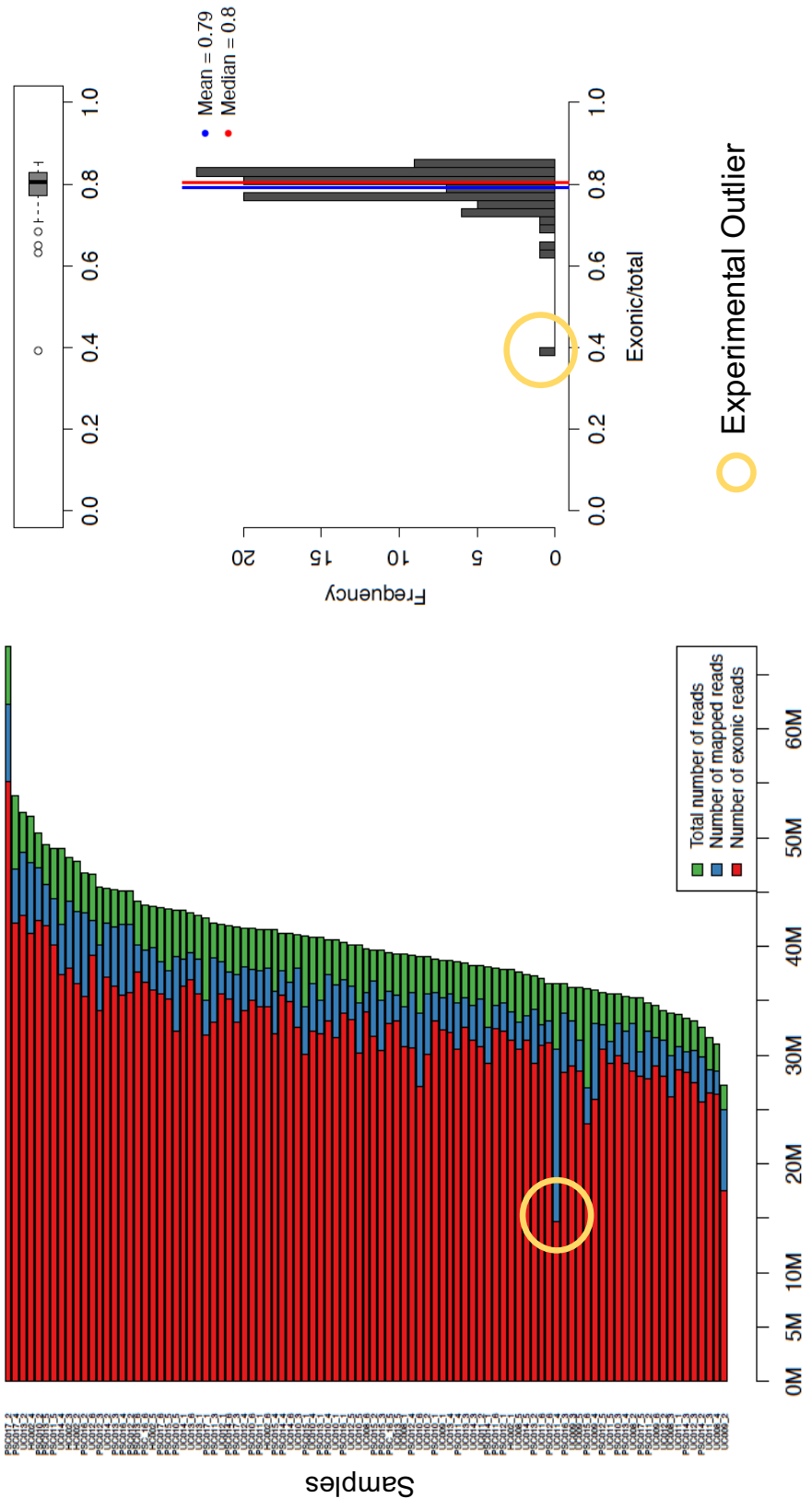

Figure 4.3: Proportion of reads mapped to exons for a subset of 96 of the total 456 samples, highlighting an experimental outlier which was subsequently excluded due to a low proportion of reads mapped to exons compared to the mean.

Duplicated genes within the pseudo-autosomal regions (PAR) were removed and normalisation performed by calculating transcripts per million (TPM). The number of reads mapping to a genes is affected by both sequencing depth (as each library has different sequencing depth) and gene length. TPM is a normalisation method that allows comparisons of genes across samples by normalising for both length of each gene and sequencing depth. Genes not expressed, or expressed at extremely low levels, defined as a sum of TPMs across all samples of <0.5, were removed. Because the presence of lowly expressed genes can decrease the sensitivity to detect differentially expressed genes, I performed a further filtering step, retaining only genes with a mean TPM of ≥ 1 in at least one disease condition.

In order to visualise samples that were experimental outliers, I performed principal component analysis (PCA) of the top 500 most variably expressed genes across all samples of all cell-types, implemented in *DESeq2* [230]. PCA uses linear combinations of gene expression values to define a new set of unrelated variables called principal components. Principal components (PCs) are orthogonal variables, where the PCs are ordered by the proportion of variation they explain in the data. This allows the description of a dataset and its variance by using a reduced number of variables, with the first two components describing the largest variability. The distances in the projection of the space defined by the principal components correlates with the similarities between the samples and thus the transcriptomes of different cell types. PC1 enabled CD4+ T-cells to be distinguished from CD8+ T-cells, explaining 52% of the variance (Figure 4.4). PC2 enabled samples from males and females to be distinguished (8% variance) and PC3 enabled CCR9+ and CCR9- cells to be distinguished (7% variance) (Figures 4.4 and 4.5). PCA was also used to identify experimental outliers, by performing PCA of the top 500 most variably expressed genes for all samples labelled according to sex, disease type and cell type. This process identified two outlying samples which did not cluster with the other samples of the same cell type (samples A and B shown in Figure 4.4), and therefore they were removed from the analysis. PCA also identified a further four outlying samples derived from two patients, which did not cluster with other samples of the expected sex (Figure 4.5). These four outlying samples were collected on the same day, and PCA confirmed that each sample clustered with the expected cell type and were therefore likely to be a direct swap or mislabelling of four samples between two patients. These samples were retained within the experiment for subsequent analysis using the *MBV* module of *QTLtools* [231] which matches genotype with transcriptome data (discussed later in Sample mismatch and amplification bias section).

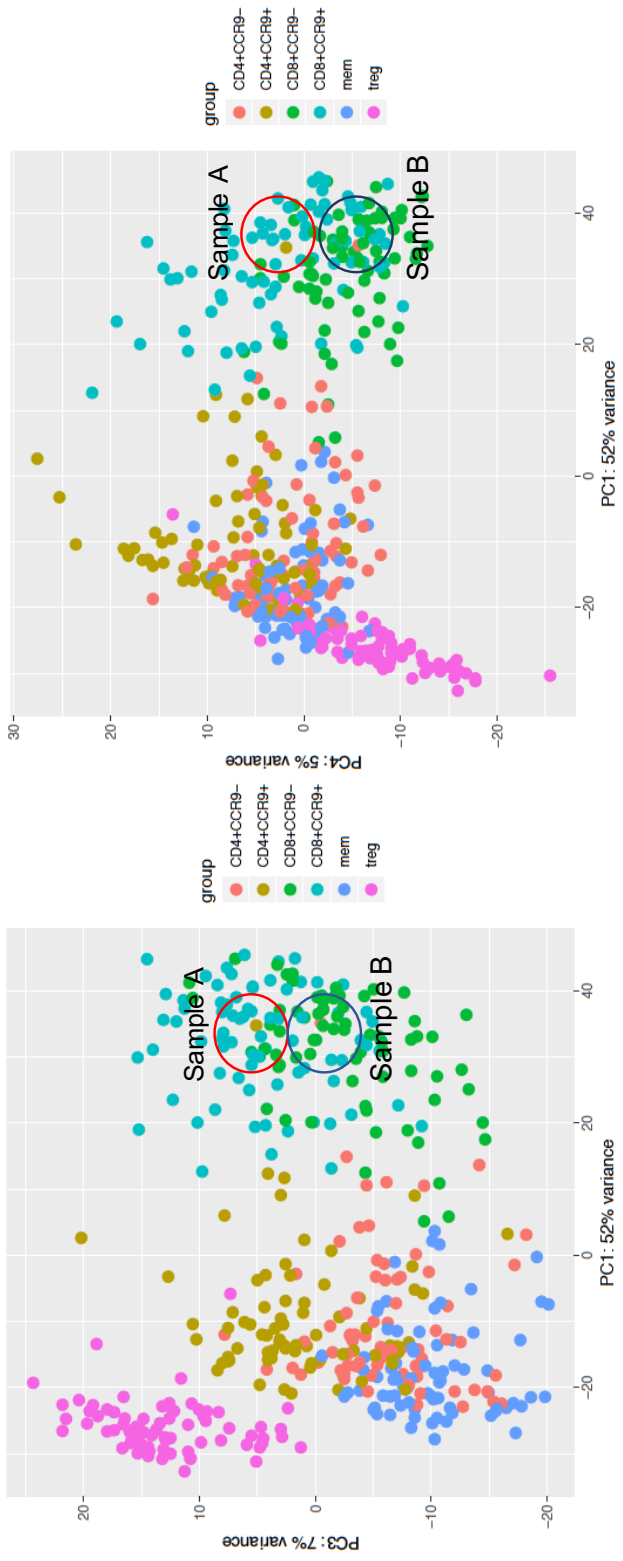

Figure 4.4: Principal component analysis of the top 500 most variably expressed genes, identifying two experimental outliers which did not cluster with their expected cell types.

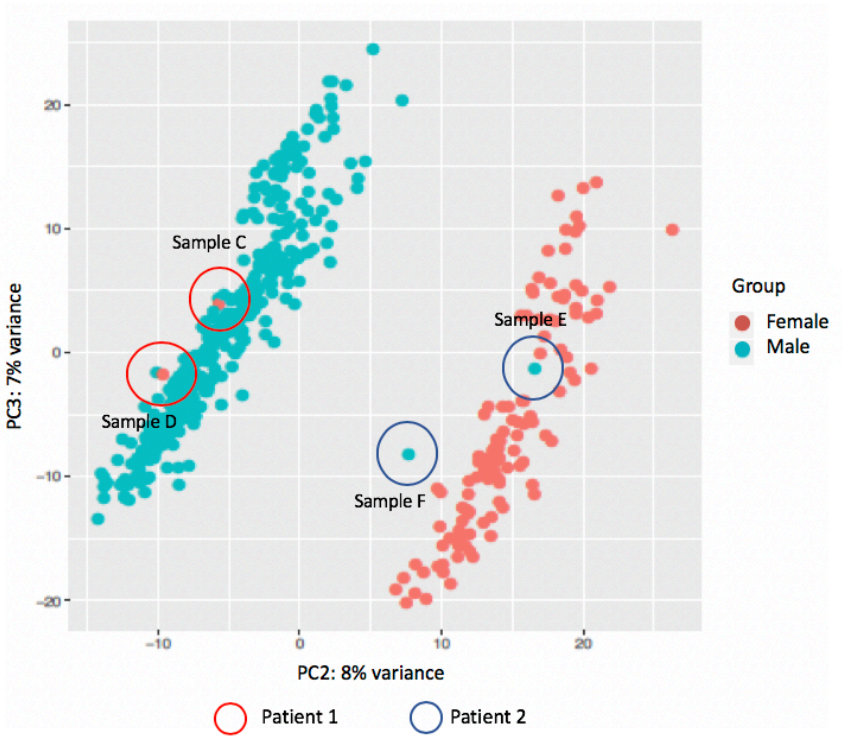

Figure 4.5: PCA analysis of the top 500 most variably expressed genes, identifying four experimental outliers from two patients.

To confirm that the gating strategy and FACS had successfully isolated the expected T-cell subtypes, I compared expression of known marker genes such as *CD4, CD8, CCR9* and *FOXP3* across all cell types. For this, I used the *PlotCounts* function implemented in *DESeq2* to visualise normalised counts of marker genes according to each cell type. This demonstrated good correlation between expected and observed marker gene expression for all cell types (Figure 4.6). The four CD4+ cell subtypes were shown to express high levels of *CD4* in comparison with the two CD8+ cell subtypes, which in turn expressed high levels of *CD8*. *FOXP3* is a transcription factor important in the development of T-regs. The T-regs in this study expressed high levels of *FOXP3*, compared to the other five cell types. *CCR9* expression was high in the two CCR9+ cell subtypes and the T-reg cell population and low in the T-mems, CD4+CCR9- and CD8+CCR9- cell types.

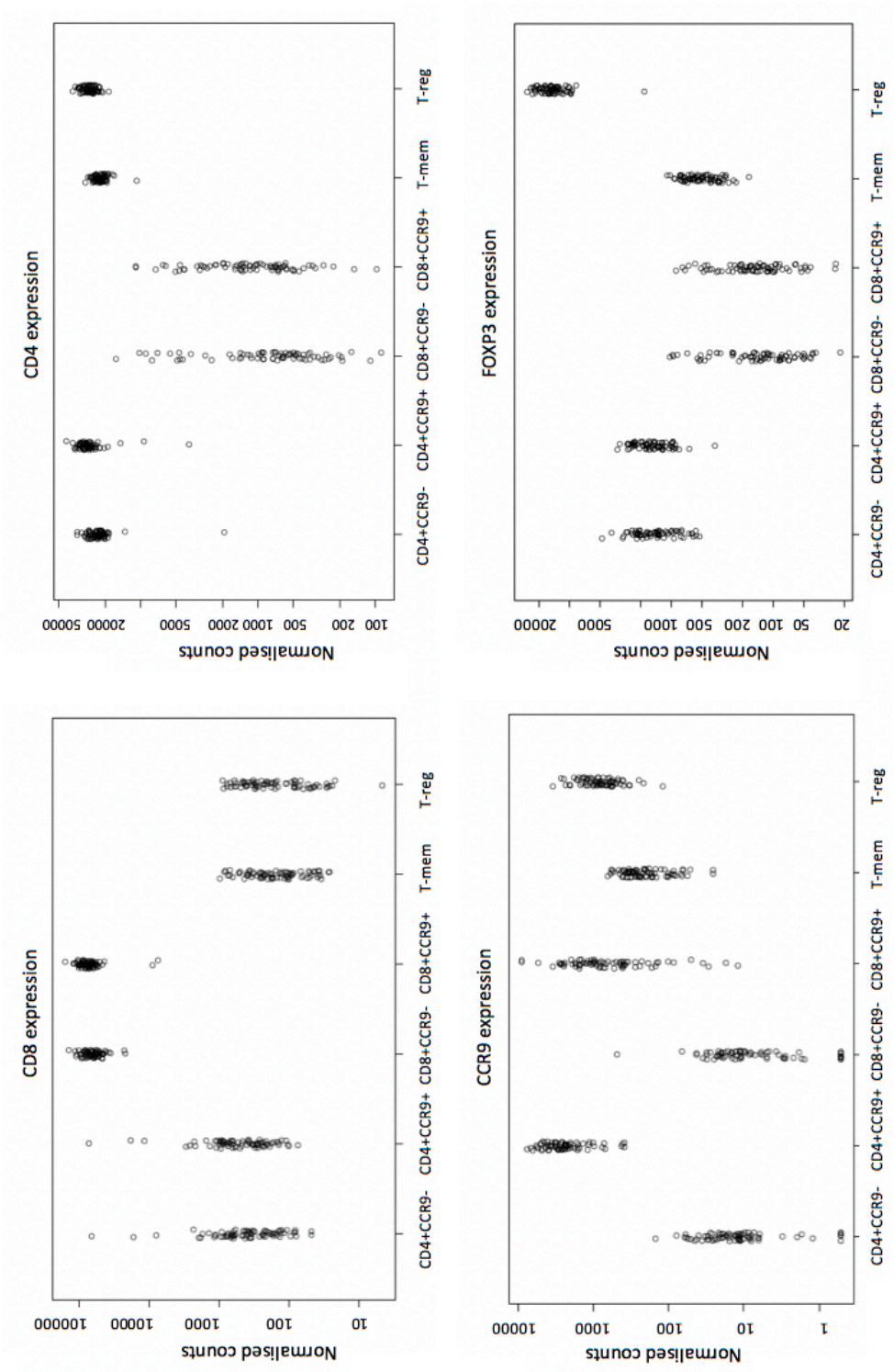

Figure 4.6: Expression of marker genes across all cell types.

4.3.5 Differential gene expression

As previously discussed, I recruited patients with both PSC-UC and lone UC for inclusion within this study, based upon evidence that the colonic inflammation in both PSC-UC

and UC patients results in increased numbers of CCR9+ effector-memory T-cells. Thus, I hypothesised that these cells would have a similar activated phenotype, with similar transcriptomic profiles in both disease groups. In order to prove that the cell types from the PSC-UC and UC groups had a similar transcriptomic profile, I performed differential gene expression (DGE) analysis between disease groups (PSC-UC, UC and HC) in each of the six T-cell subtypes.

For the analysis of differential gene expression I used *DESeq2* package version: 1.25.0. *DESeq2* is a tool for analysis of differential gene expression, using shrinkage estimation for dispersions and fold changes to improve stability and interpretability of estimates [230]. There are several similar methods available, including *edgeR* and *limma-voom*. I chose the *DESeq2* method over *limma-voom* because it offers a more stringent count normalisation method, based upon generalised linear modelling (GLM) or negative binomial modelling rather than linear modelling. This is especially important when dealing with very small sample sizes. For example, the HC sample group in my study contained only 5 individuals and *DESeq2* has been shown to have comparatively improved specificity and sensitivity as well as good control of false positive errors, even with small samples sizes [232]. In comparison with *DESeq2*, the *edgeR* method also uses a negative binomial distribution, with comparable specificity and sensitivity and I chose the former due to improved usability.

The input for *DESeq2* is the raw count matrix K (where 'count' refers to the number of sequencing reads unambiguously mapped to gene in a sample), including only those genes and samples taken forward following the aforementioned QC steps. Each row of the count matrix contains one gene *i*, and each column contains the number of counts for that gene in a sample *j*. *DESeq2* firstly normalises for sources of systemic variation between samples; library size and sequencing depth. This is important because not all samples have been sequenced to exactly the same depth and larger library sizes result in higher counts. It also normalises for two important sources of within-sample gene-specific effects. The first is related to gene length, because the total number of reads mapped to a given transcript is proportional to the expression level of the transcript multiplied by the length of the transcript [233]. The second is related to GC content which is heterogeneous across the genome and can affect the mapping of reads [234]. The method of normalisation used by *DESeq2* is called the 'median-of-ratios' method, which I have described in Figure 4.7. The output of this normalisation is a normalisation factor, S_{ij}, for each sample in the experiment [235]. *DESeq2* models the counts for K_{ij} as following a negative binomial distribution with mean μ_{ij} and dispersion α_i (dispersion estimation described more fully below). μ_{ij} is a quantity, q_{ij}, proportional to the concentration of cDNA fragments from the gene in the sample, scaled by the normalisation factor S for that sample;

$$\mu_{ij} = S_{ij} * q_{ij}$$

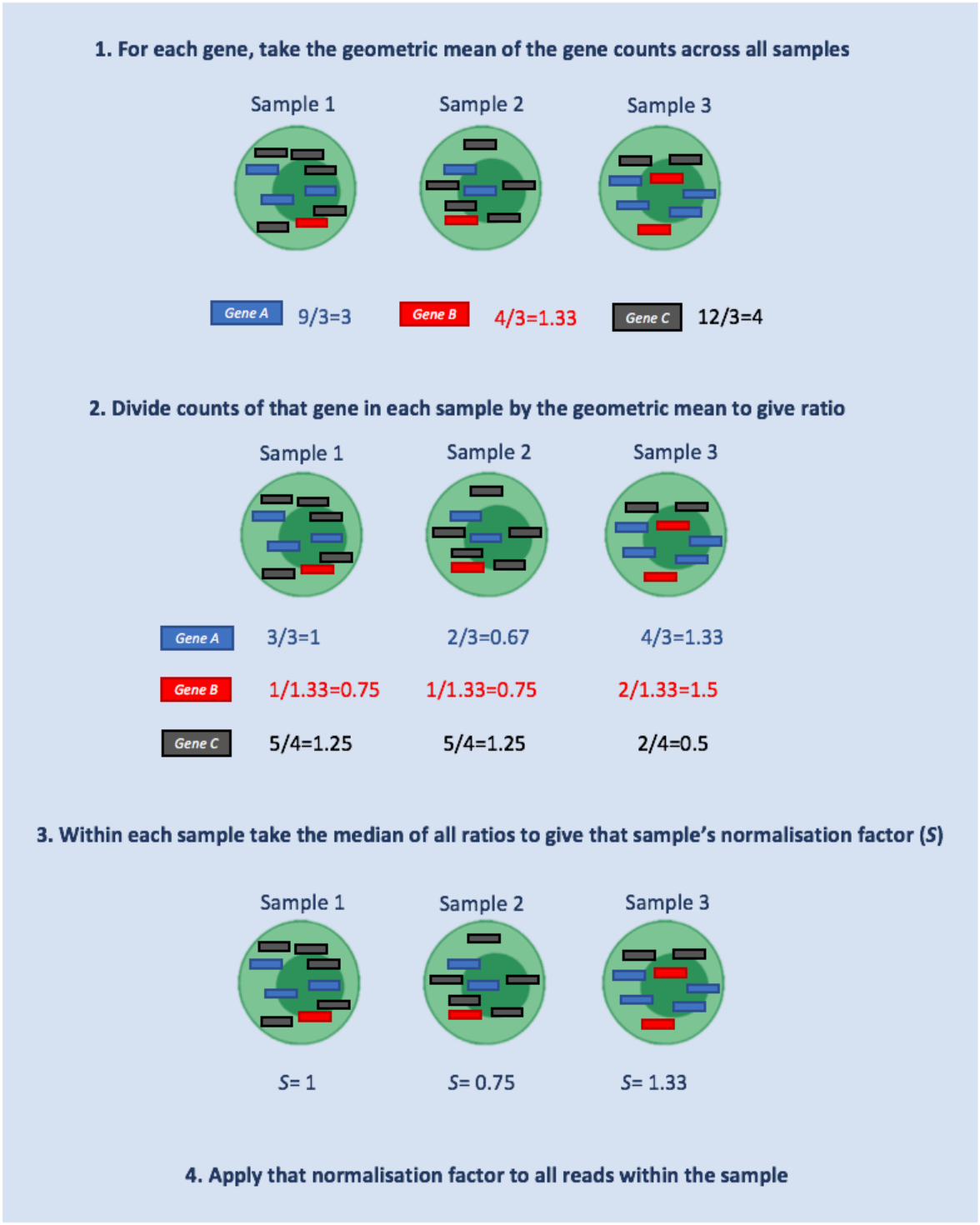

Figure 4.7: Schematic representation of the *DESeq2* method of normalisation.

To compare two groups (eg. PSC-UC versus UC), *DESeq2* fits a GLM with logarithmic link of the overall expression strength of a gene and the $\log_2$ fold change (LFC) between the two groups, as a combination of explanatory factors or covariates such as group, patient and sample;

$$log2q_{\mathrm{ij}} = a + (b * group) + (c * patient) + (d * sample) + e$$

where a is the intercept, b, c and d are parameters estimated from the data, and e is the error term. When comparing a gene's expression level between groups, *DESeq2* accounts for the within group variability of that gene's expression using dispersion estimation, α_{i} to model the variance of counts, Var K_{ij}.

$$VarK_{ij} = \mu_{\mathrm{ij}} + (\alpha_{\mathrm{i}} * \mu_{ij})$$

For the statistical inference of differential expression, it is important that estimation of the dispersion parameter, α_{i} is accurate. Because some RNAseq experiments, such as the HC group in my study, include only a few biological replicates, estimating the within group variability is difficult, especially because genes expressed at very low levels have much higher dispersion estimates. If used, these higher dispersion estimates would introduce noise and affect the accuracy of the differential expression analysis. To account for this *DESeq2* assumes that genes with a similar average expression have similar dispersion. It then estimates gene-wise dispersions (for each gene separately) using a maximum likelihood and shrinks dispersion estimates towards a fitted average dispersion curve, using an empirical Bayes approach. As sample size increases, the scale of shrinkage decreases.

When estimating log fold change (LFC), there is strong variance for genes expressed at very low levels. This is a result of working with count data, where even a small error in counting mapped reads causes a comparatively big change in LFC estimation for those genes expressed at very low levels. If unaccounted for, this would make the downstream estimation of effect sizes difficult to compare across the range of data. *DESeq2* deals with this by shrinking LFC estimates towards zero using an empirical Bayes method. This can be visualised on an MA plot, which shows the differences between measurements taken in samples, by transforming the data onto M (log ratio or LFC) and A (mean of normalised counts) scales, then plotting these values. Figure 4.8 shows two MA plots for all of the data in my differential gene expression (DGE) study, before and after shrinkage has been applied. This demonstrates that shrinkage is stronger when counts are low and dispersion is high, removing the problem of exaggerated LFCs for genes with low counts.

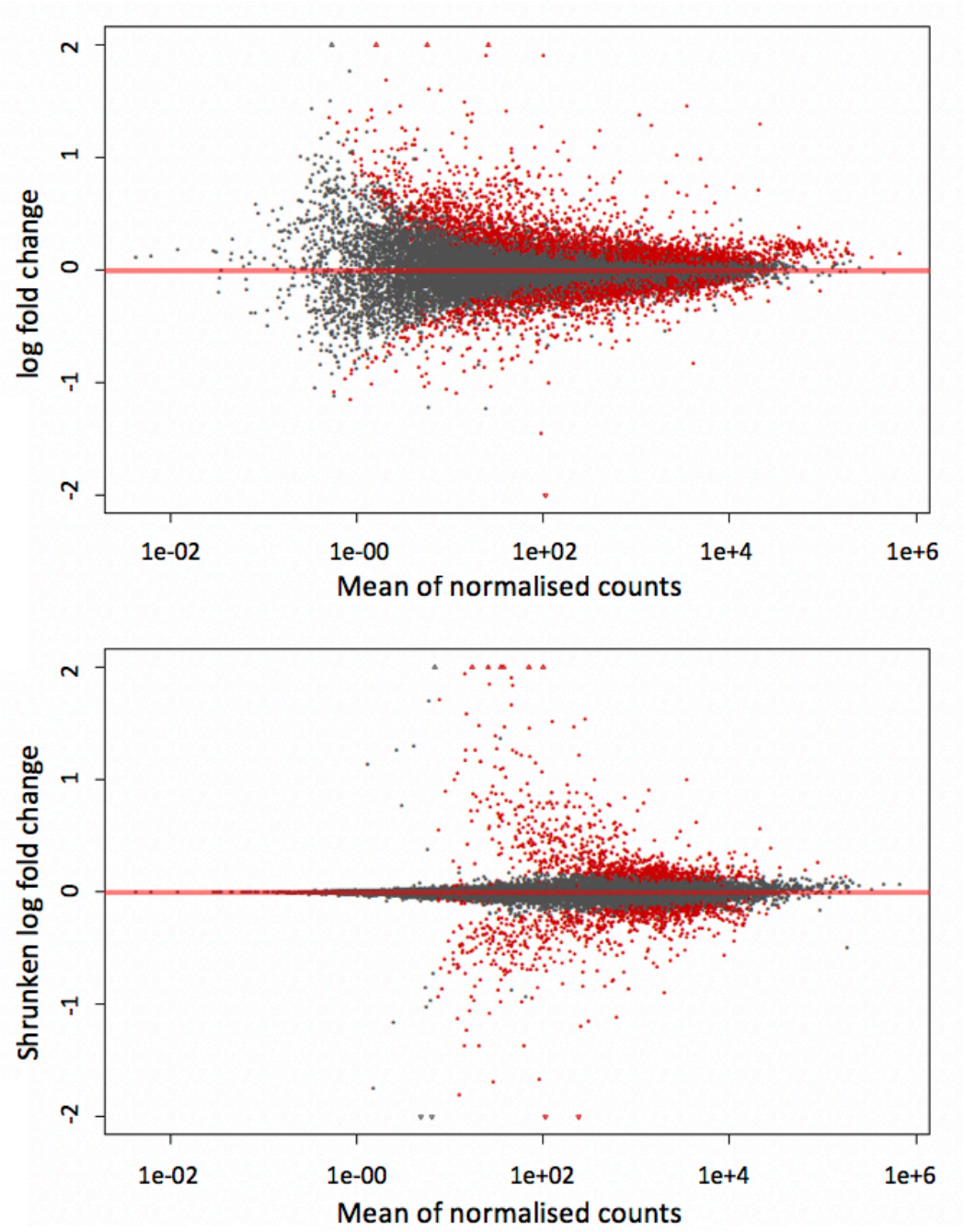

Figure 4.8: MA plots with and without shrinkage applied. Points are coloured red where the adjusted p-value is less than 0.05, and plotted as open triangles pointing either up or down if they fall outside of the window.

Having fit a GLM for each gene, the next stage is to test whether the coefficient for each model is significantly different from zero. *DESeq2* uses a Wald test for significance

where the shrunken estimate of LFC is divided by its standard error and the resulting Z statistic compared to a standard normal distribution with a resultant p-value. Because many thousands of genes are tested, it is possible to obtain some significant p-values just by chance (false positives), hence in the final stage of the analysis I corrected the p-values for multiple testing. I used the Benjamini-Hochberg (BH) correction method [164] to obtain adjusted p-values at a 5% false discovery rate (FDR).

I performed differential gene expression analysis between each of the three disease groups (PSC-UC, UC and HC), in a pair-wise fashion. I controlled for known covariates in the *DESeq2* model including patient age, sex, use of drugs including 5-aminosalicylates and azathioprine, and the sample sequencing run. I reported genes as differentially expressed if the adjusted p-value was <0.05. I performed gene ontology (GO) analysis of all DEGs in each group, using web-based GO platform, *g:Profiler* [236], to elucidate aspects of the underlying disease biology.

4.3.6 Genotype QC and imputation

Paired genotype and expression data is required for the mapping of eQTL. DNA samples from blood or saliva were available for 74 of the 76 patients. DNA extraction of all samples was performed by Dr Rebecca McIntyre, Senior Staff Scientist at the Wellcome Trust Sanger Institute. Extraction was performed using *Qiagen* DNeasy Blood and Tissue Kit and sequenced by the Wellcome Sanger Institute DNA pipelines, using the Illumina Omni2.5-8Exome BeadChip. I performed all QC on the raw genotype data, using the PLINK software version 1.9, following Anderson *et al*'s published standards for the QC of genotype data for genome-wide case-control association studies [237]. I considered all autosomal and chromosome X SNPs without insertions or deletions. The sequence of pre-imputation QC is shown in Figure 4.10 with further details on per-SNP and per-individual QC outlined below.

The removal of suboptimal SNPs is important for avoiding false-positive associations which reduce the ability to identify true associations correlated with disease risk. To remove individuals and SNPs with a particularly high error rate, but maximise the number of SNPs remaining within the study, I first removed individuals with a genotype call rate of $<95\%$ and SNPs with call rate of $<95\%$. SNPs with a very low frequency can be difficult to call using current genotype calling algorithms due to the small numbers of heterozygotes and homozygotes. Furthermore, power to detect association at rare variants is low, and thus I removed variants with a MAF <0.01.

Per-individual QC included the identification of individuals for whom information on sex was disconcordant between genotype and ascertained sex. This was done by calculating the homozygosity rate across all X chromosome SNPs for each individual within the sample and comparing this to the expected rate. Males are expected to have a homozygosity rate

around 1 (with some variation due to genotyping error), and females a homozygosity rate of around 0.2. This is because males have just one copy of the X-chromosome and thus cannot be heterozygous for any marker outside of the pseudo-autosomal Y chromosome region. There were no sex discrepancies between genotype and ascertained sex in my samples. In order to reduce the effect of population stratification, I next identified any individuals of ancestry divergent from the expected European ancestry. Excluding variants from regions of known high LD, I identified a pruned set of 62,805 independent variants from my dataset, all with an $r^2<0.2$ and MAF>0.01. I then identified the same subset of variants within the 1000 Genomes dataset. Using this pruned set of independent variants, I performed a PCA analysis of my individuals combined with the 1000 Genomes cohort. By plotting the first and second principal components of this combined dataset, I could visually identify that all of my samples were clustered with the known European individuals of the 1000 Genomes dataset (labelled 'PSC' in Figure 4.9). Notably, of all individuals passing QC and retained for downstream analysis, three were of Southern European/Iberian ethnicity, highlighted on Figure 4.9 and all remaining samples were of Northern European ethnicity. All samples from individuals of Northern and Southern European ethnicity were retained for further analysis.

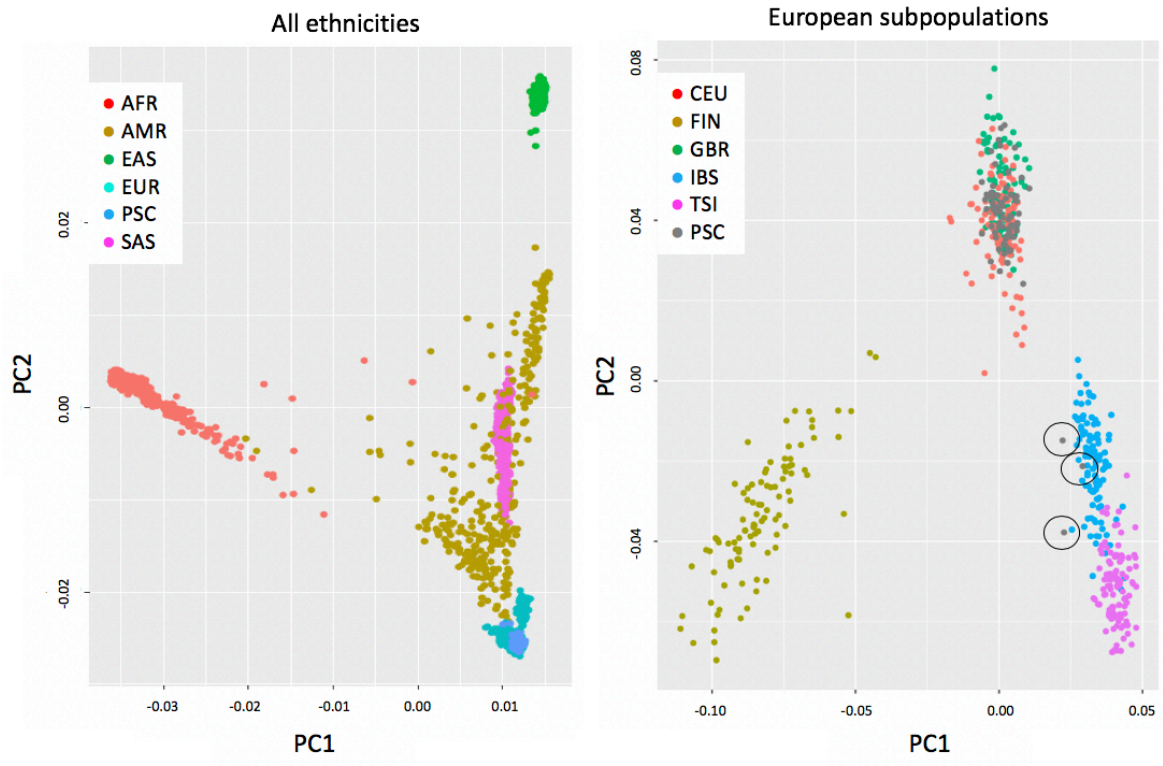

Figure 4.9: PCA of study samples compared to 1000 Genomes samples of known ethnicity using a pruned set of 62,805 independent variants with an $r^2<0.2$ and MAF>0.01.

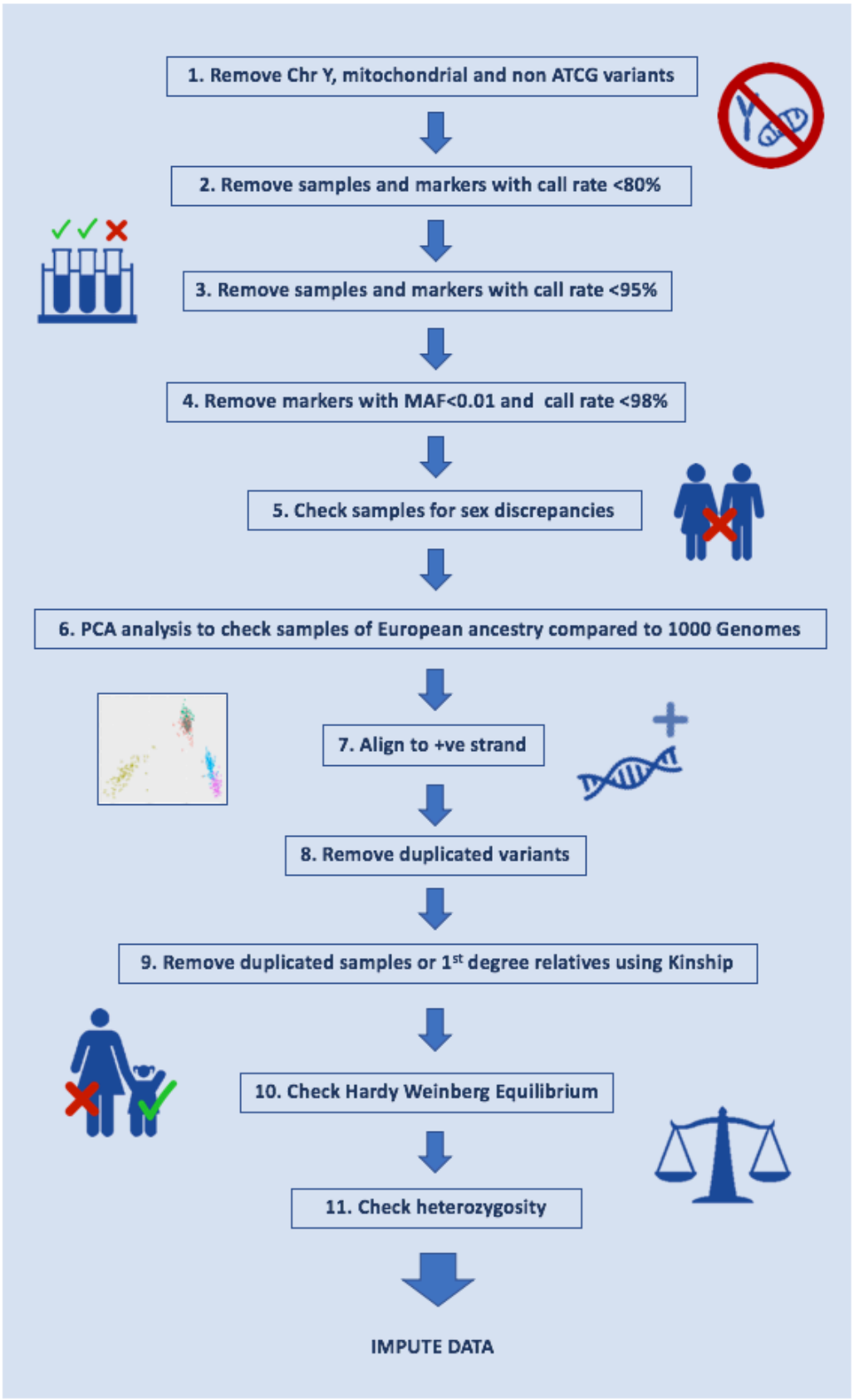

Figure 4.10: Outline of pre-imputation QC of genotype data.

The heterozygosity rate per individual can be used as a measure of DNA sample quality. Considering only autosomal chromosomes, I examined the distribution of the heterozygosity rate, excluding any samples with a heterozygosity rate more than two standard deviations from the mean. The mean heterozygosity rate was 0.274 and one sample fell outside the two standard deviations threshold resulting in its removal from the analysis. To avoid the bias of over-represented genotypes introduced by first-or second degree relatives, the next stage of per-individual QC was to identify any duplicated or related individual, to ensure the maximum relatedness between any pair of individual was less than second-degree relatives. I used KING software v2.2 and a set of 2,513,131 variants with MAF>0.01, and call rate >98% to infer close relatives based on the estimated kinship coefficients. I identified two first degree relatives (kinship coefficient range 0.177 to 0.354), one of which was removed from subsequent analysis.

SNPs with extensive deviation from Hardy-Weinberg Equilibrium (HWE) may indicate selection, occurring at loci associated with disease, but can often be indicative of genotype calling error. As part of the per-marker QC, I removed variants with a HWE p-value of $<1 \times 10^{-8}$. Following the above QC steps a dataset including 71 individuals and 1,590,593 variants remained, and were put forward for imputation. I imputed a further ∼5.5 million variants against the UK10K, 1000 Genomes phase 3 and Haplotype Reference Consortium reference panels, using the Wellcome Sanger Imputation and Phasing Service pipeline, IMPUTE2 [238]. IMPUTE2 provides an 'info' score related to the quality of the imputation for each SNP. Post-imputation QC consisted of removing any SNPs with a low info score <0.3. This threshold was decided by plotting an info score frequency curve and assigning the threshold at the inflexion point [239]. The final post-imputation QC step was to re-check the HWE as described above. The resultant post-imputation, post-QC dataset consisted of 7,027,506 SNPs. Mapping of eQTLs requires the addition of known covariates within the model, including principal components (PCs) from the genotype data. Therefore, using the final QC'd and imputed genotype dataset, I performed a PCA using the PLINK (v1.9) *PCA* function with the aforementioned pruned set of 62,805 independent variants from low LD regions. I retained the resulting genotype PCs for inclusion as covariates in the downstream eQTL analysis.

I processed all genotype and imputed data in ensembl build 37, but for further downstream processing performed a genome coordinates conversion or 'lift-over' to ensembl build 38 using CrossMap v0.3.5 which supports the conversion of variant call format (VCF) files between different genome assemblies [240].

4.3.7 eQTL mapping

I conducted all eQTL analysis and mapping using *QTLtools* v1.1 9, which provides a complete toolset for molecular QTL discovery and analysis [241]. The analysis outlined

below was performed using a normalised gene expression matrix, which had undergone prior QC (as described in the RNA sequencing and sample QC section above), and the previously QC'd and imputed genotype data (as described in the Genotype QC and Imputation section above).

4.3.7.1 Identifying sample mismatches and amplification bias

To ensure that the genotype and gene expression data for each individual in the study was a true match, I used the MBV (Match BAM to VCF) module of *QTLtools* [231]. MBV identifies sample mislabelling, cross-sample contamination and PCR amplification bias. The input files for MBV were the VCF file containing the genotype data for all 71 individuals within my study, and the BAM file for the mapped RNA reads for each individual at a time. For each SNP site in the VCF file, MBV aggregates the sequencing reads and discards those SNPs not reaching a minimal coverage parameter threshold. For each individual within the VCF file, it calculates the proportion of heterozygous and homozygous genotypes for which both alleles have been captured by the sequencing reads and reports the two concordance measures for each individual. Where both measures are close to 100% concordance, this describes a match between genotype and gene expression datasets. Where there is decreased heterozygous concordance with no change in homozygous concordance this is described as 'no match' between genotype and gene expression, but in fact represents a match but with amplification bias effect (Figure 4.11). Twenty-three percent of samples demonstrated heterozygosity concordance of less than 0.66 with no change in homozygous concordance. In order to account for the effect of such amplification bias, the fraction of heterozygosity concordance for each sample was taken forward as a covariate for inclusion in the eQTL analysis.

There were no instances of sample contamination within this dataset, which can be detected by a reduction in the fraction concordance at homozygous compared to heterozygous sites. I detected four cases of 'unexpected matches', two from the same male recruit and two from the same female recruit (Figure 4.12). These were the same four samples detected to be outliers on PCA according to sex, as previously described in the RNA sequencing and sample QC section above. This was the result of an accidental direct swap of two RNA sample labels (CD8+CCR9- and CD8+CCR9+ samples) from one male individual with two RNA sample labels for the same two cell types from one female individual. Following this stage of QC, these four samples could be re-assigned to the correct individual and therefore retained for eQTL mapping.

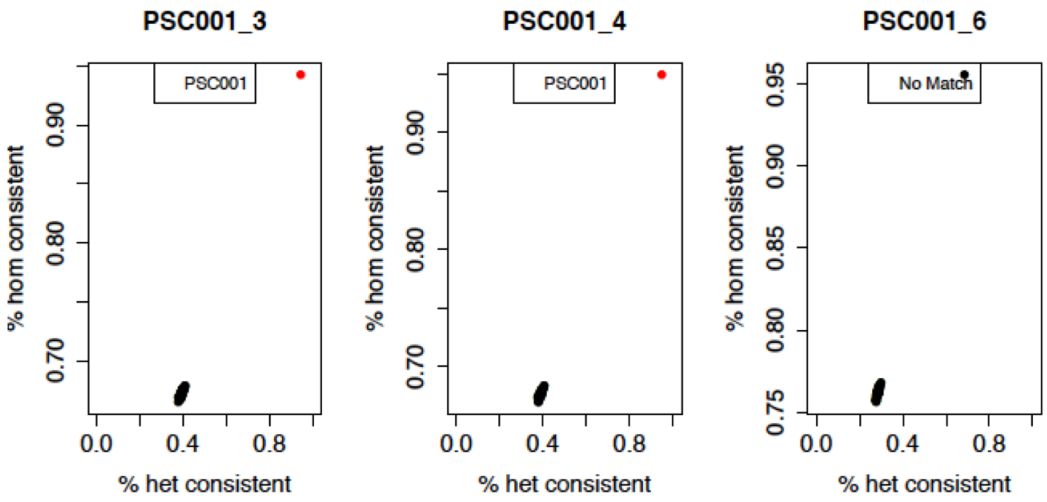

Figure 4.11: Concordance at heterozygous genotypes (x-axis) versus concordance at homozygous genotypes (y-axis), for each individual genotype sample (black dots). A match between genotype (box at top) and gene expression data (plot title) is coloured red (two left hand examples). A mismatch or amplification bias is coloured black (right hand example).

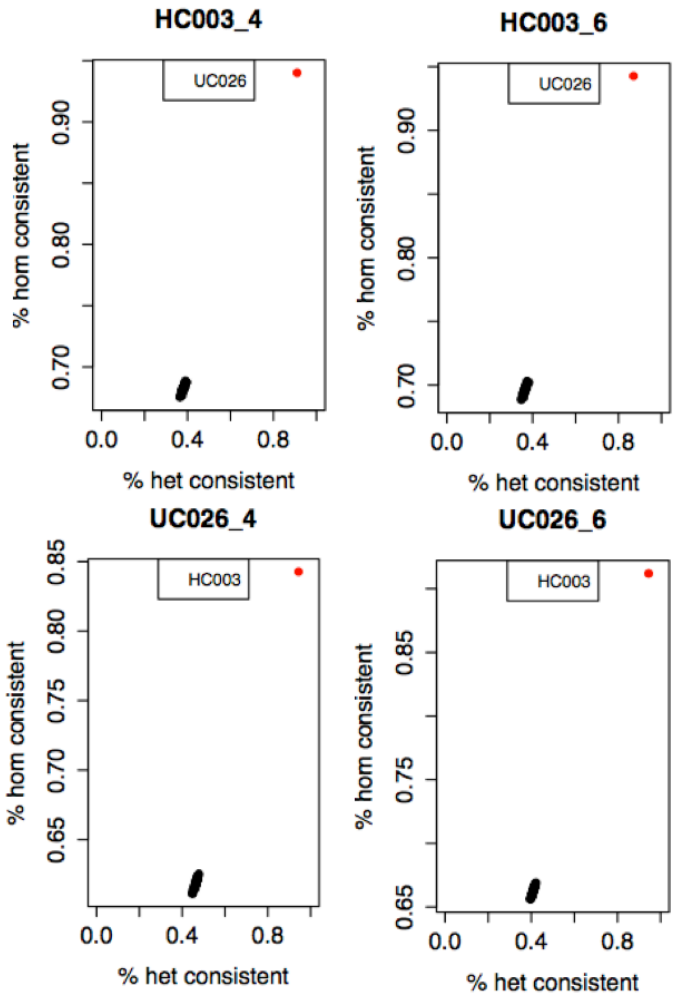

Figure 4.12: Concordance at heterozygous genotypes (X-axis) versus concordance at homozygous genotypes (Y-axis), for each individual genotype sample (black dots). An sample mismatch is shown by a match between a different genotype (in box at top) and gene expression data (plot title) in all four examples.

4.3.7.2 Identifying *cis*-eQTLs

For the identification and mapping of *cis*-eQTLs in each of my T-cell subsets, I used *QTLtools* [241]. Mapping eQTLs involves the testing of association between gene expression (phenotype of interest) and all the genetic variants within a window upstream and downstream of the transcription start site (TSS) of the gene, with millions of tests performed genome-wide. A linear regression model is fitted between the genotypes and gene expression, including multiple covariates to correct for batch and other effects, in order to find the best nominal associated variant per gene. Analysis of all gene-variant pairs requires millions of association tests, each producing a nominal p-value. Whilst adjustment of nominal p-values to correct for multiple testing and avoid false positives

must be performed, the presence of linkage disequilibrium (LD) means that the tests are not entirely independent, calling for a less stringent correction than the Bonferonni method. To deal with this issue *QTLtools* uses permutations to derive adjusted p-values per phenotype/gene. *QTLtools* uses a beta approximation permutation scheme based on Ongen *et al*'s *FastQTL* beta approximation permutation scheme, to correct for the testing of multiple variants per gene [242]. This scheme creates multiple permuted datasets by keeping the genotypes static (thus preserving correlation structure between variants) and permuting the gene expression data for every gene. For every permutation the best nominal association is retained to form a distribution of p-values expected under the null hypothesis of no association. Next, an adjusted p-value is calculated based on how likely it is that an observed association obtained in the nominal pass, originates from the null. *QTLtools* models this distribution of p-values expected under the null hypothesis of no associations, using a beta distribution. It approximates the tail of the null distribution to estimate adjusted p-values at any significance threshold, with no lower bounds.

The input files for *QTLtools* are the zipped and indexed VCF file (which had been previously QC'd and imputed as described above), the indexed and zipped gene expression BED files (reporting normalised expression in TPM and QC'd as previously described) and the covariate files. In my analysis I included age, sex, the first three genotype principle components (described in Genotype QC and imputation section above), fraction of heterozygosity concordance (described in Sample mismatch and amplification bias section) and a variable number of gene expression-derived principal components (PCs) as covariates. I calculated the gene expression PCs in the same way as the genotype PCs, using PLINK v1.9's *PCA* function.

To map eQTLs I ran *QTLtools* for each of the six T-cell datasets, using 1,000 permutations and a *cis*-window of 1Mb. To maximise the number of eQTL discoveries, I optimised the QTL mapping by performing multiple runs of the analysis including an increasing number of gene expression-derived principal components from zero to 50. To account for the thousands of genes tested genome-wide, I performed an FDR correction on the set of adjusted p-values obtained by the permutation analysis for every gene, using the R package, *qvalue* [243]. In contrast to the p-value, which measures significance in terms of the false positive rate, the q value is a measure of significance in terms of the false discovery rate. An FDR threshold of 5% therefore means that on average, 5% of the eQTLs called significant are truly null [243]. In order to find the optimal number of gene expression PCs required to detect the maximum number of eQTLs, I plotted the number of eQTLs against the number of expression PC's included within the linear regression model. For each cell type, I settled on the number of PC's that maximised the number of eQTLs, and included this number of PCs in the covariate model which was taken forward for subsequent analyses as described below [215, 241].

Further analysis of eQTL data, for example for meta-analysis or colocalisation, requires all the nominal associations (including those that do not reach statistical significance). To generate this data, I used the *QTLtools nominal pass* function, the same gene expression BED and genotype VCF files as described above and the covariate files containing the same number of gene expression PCs for detecting the maximum number of significant eQTLs.

4.3.8 Identifying shared and tissue-specific eQTL

Having mapped eQTLs for six individual cell types, an important question is to identify those eQTLs which are shared across cell types, and those that are cell-type specific. By allowing for the correlations of effect sizes among cell types using a form of meta-analysis, this can increase power by improving estimation of effect sizes and allow for more accurate comparison of effect sizes between tissues. Several statistical methods for analysing shared eQTL associations have been published which learn the patterns of eQTL sharing from the data using a hierarchical model [244–246]. However, each has its own limitations, for example the model by Flutre *et al* is limited by the assumption that correlations are non-negative and equal, such that it does not allow for genetic variants leading to an increased effect in one trait and a decrease in another [244]. Furthermore, Flutre *et al*'s methods provides flexibility at the cost of becoming computationally intractable when considering even moderate numbers of tissues or cell types and thus the authors sought to solve this by restricting effects to either a single cell type, or shared across all cell types. Another method published by Wei *et al* allows for all patterns of sharing, but is limited by the assumption that nonzero effects are uncorrelated among conditions, and thus focuses only on testing for significant effects and not on estimating effect sizes [245]. *MashR* (multivariate adaptive shrinkage in R) is a method that addresses these limitations, allowing for shared, condition-specific and arbitrary patterns of correlation among conditions, as well as providing measurements of significance and effect size estimates [246].

I used *mashR*, implemented in R, for further analysis of my eQTL data. The input data for *mashR* are the nominal pass of the individual cell-type analysis performed with *QTLtools* as described above. These include the effect size estimates (β's) and corresponding standard errors (SE) for all eQTL/Gene pairs in each cell type with no significance threshold. These measurements are the input for *mashR*'s two-step empirical Bayes procedure, which firstly learns the patterns of sparsity, sharing and correlations among effects from the individual cell-type results and secondly, combines these learned patterns to produce improved estimates of effect and their corresponding significance. For the first step, *mashR* requires a subset of 'strong' tests, corresponding to the strongest effects in the individual cell-type analysis. I identified this subset of 'strong' tests by taking the most significant eQTL per gene across all six cell types, from all significant eGenes

from the individual cell-type analysis. This produced a strong subset of 5,487 eGenes. Next, *mashR* requires a 'random' subset of all tests, which is an unbiased representation including null and non-null tests. I created a 'random' subset of 200,000 tests using the R function, *set.seed*, which is a reproducible random number generator. The random subset is used by *mashR* to estimate the correlation structure between tests, via a PCA-like approach and the strong subset is used to define the data-driven covariance matrices. The *mashR* model is then fitted to the random tests using both the data-driven covariance and *mashR*'s in-built canonical covariances. I then used the resultant *mashR* model to compute posterior summaries for all of my data. For each eQTL/Gene test in each of the six cell-types, the output includes the posterior β, SE, *lfsr* (local false sign rate, analogous to an FDR) and $\log_{10}$Bayes factor (a measure of the overall significance for a non-zero effect in any condition).

The final stage of the analysis is to call cell-type specific and shared eQTL from the *mashR* posterior summaries. From the posterior summaries for all of my data, I identified the subset of eQTL/Gene pairs significant in at least one cell type at *lfsr*<0.05. From this subset I extracted data for the most significant eQTL per gene, defined by the eQTL/gene pair with the smallest *lfsr* across any of the six cell types as described by Kim-Hellmuth *et al* in the analysis of cell-type specific eQTLs in the GTEx data [247].

4.3.9 Colocalisation

I performed colocalisation with the eQTL data derived for each individual cell type using the output data from *QTLtools'* nominal pass and permutation pass. I performed colocalisation at the fifteen PSC risk loci reported by Ji *et al* [42] with GWAS summary statistics from the same study using the same methods for colocalisation as previously described in Chapter 3. Where the PP4 for colocalisation of a PSC risk locus with a T-cell eQTL was >0.8 for at least one cell type, I explored whether this same locus also colocalised with the same eQTL in other cell types using the *mashR* eQTL data. I took the posterior results of *mashR* analysis for posterior standard deviation (standard error), lfsr (analogous to an FDR) and posterior mean (β) for each cell type, and performed colocalisation at those PSC risk loci, visualising the results on regional association plots. Finally, given that the majority of the study cohort were patients with UC and that genetic architecture is shared across many IMDs, I conducted colocalisation with other IMDs. I performed colocalisation with 240 IBD, 100 RhA and 45 T1DM risk loci, using their associated GWAS summary statistics [60, 148, 200] and nominal pass eQTL data for each T-cell subset (derived from the *QTLtools* individual cell-type eQTL analysis).

4.4 Results

4.4.1 Differential gene expression

I tested 20,547 genes for differential expression between each of the three disease groups (PSC-UC, UC and HC). Characteristics of the study cohort, according to disease group are shown in Table 4.2. I controlled for covariates including patient age, sex, use of 5-aminosalicylates or azathioprine and the sample sequencing run. The results of this analysis showed no significant differences is gene expression across all six T-cell subtypes in the PSC-UC group compared to the UC group (Table 4.3). Given that both groups share the UC phenotype, this finding is not unexpected. Furthermore, the results supported no significant changes in gene expression between both the PSC-UC and UC groups versus HC, in T-regs, CD4+CCR9-, CD4+CCR9+ and CD8+CCR9+ cells. Whilst there were a few DEG's (≤ 7) between the above comparator groups, genes are reported at a 5% FDR, therefore a false positive rate of 5% is expected, limiting any interpretation where such low number of DEGs are reported. Further visualisation of normalised counts for these few genes in each disease group confirmed that most genes reported as differentially expressed, were false positives.

Differential gene expression was observed between both PSC-UC and UC groups compared to HCs in two cell populations; T-memory and CD8+CCR9- T-cells. Using *gProfiler*, I performed GO analysis of all genes differentially expressed between these disease groups. GO term analysis of 367 DEGs in the T-memory cells of UC compared to HCs demonstrated enrichment of pathways involved in cellular metabolic activity ($p=1.1 \times 10^{-12}$) (Figure 4.13). The finding of a more metabolically active phenotype in the T-memory cells of patients with UC versus HCs may support a role for these cells in the disease pathogenesis. GO analysis was unable to find any more specific pathway enrichment based upon these DEGs. There were 101 DEGs between PSC-UC and HCs in T-memory cells. However GO analysis did not find any significant pathway enrichment, likely a result of the relatively low numbers of DEGs between these two groups.

The second cell type demonstrating significant numbers of DEGs in the PSC-UC and UC groups compared to the HC group, were the CD8+CCR9- T-cells. Here, 94 and 34 genes were differentially expressed in PSC-UC and UC groups compared to HCs respectively. GO analysis did not find any specific pathway enrichment for any of the DEGs, again, likely a result of the low numbers of DEGs. However the finding of a difference between the transcriptomes of CD8+CCR9- cells of PSC-UC and UC patients versus HCs, in the absence of any difference in the transcriptomes of CD4+CCR9- in the same groups, is interesting given existing evidence in IBD, of a CD8+ T-cell signature of immune-cell exhaustion, driving a more severe disease course in IBD [248]. Indeed, it has been reported that elevated expression of genes involved in antigen-dependent T-cell

Table 4.2: Characteristics of the study cohort according to disease group.

	PSC-UC n=42	UC n=29	HC n=5
Gender (% male)	78	69	60
Mean Age (Range)	50 (17-86)	52 (48-56)	44 (28-51)
UDCA use (%)	71	0	0
5-ASA use (%)	67	90	0
Azathioprine use (%)	57	21	0

responses, including IL-7 signaling and TCR ligation, specific to CD8+ T-cells and absent in CD4+ T-cells, can predict a more severe disease phenotype in IBD patients [248, 249]. Importantly, I found that several genes involved in TCR antigen recognition were up-regulated in CD8+CCR9- T-cells of PSC-UC groups versus HC, including *TRAV38-2DV8* (T cell receptor alpha variable) and *TRBV25-1* (T cell receptor beta variable), genes which encode the variable domain of T cell receptor (TCR) α and β chains respectively. In the CD8+CCR9- T-cells of PSC-UC versus HC, there was increased expression of *BTLA* (B- and T-lymphocyte attenuator), a gene induced during activation of T cells, and decreased expression of *IL-15*, a cytokine which prevents apoptosis and maintains memory T cells in the absence of antigen. Thus it appears that the CD8+CCR9- memory T-cells in PSC patients express genes consistent with a more active phenotype with reduced repression of apoptosis, compared to HCs.

Table 4.3: Comparison of differentially expressed genes for each of six T-cell subtypes according to disease group, reported at 5% FDR

Samples	Group 1	Group 2	Total no. of DEGs	No. up-regulated	No. down-regulated
T-reg	PSC-UC	UC	3	1	2
T-reg	PSC-UC	HC	5	3	2
T-reg	UC	HC	3	1	2
T-mem	PSC-UC	UC	0	0	0
T-mem	PSC-UC	HC	101	32	69
T-mem	UC	HC	367	143	224
CD4+CCR9-	PSC-UC	UC	1	0	1
CD4+CCR9-	PSC-UC	HC	7	1	6
CD4+CCR9-	UC	HC	4	1	3
CD8+CCR9-	PSC-UC	UC	0	0	0
CD8+CCR9-	PSC-UC	HC	94	47	47
CD8+CCR9-	UC	HC	33	27	6
CD4+CCR9+	PSC-UC	UC	1	1	0
CD4+CCR9+	PSC-UC	HC	1	0	1
CD4+CCR9+	UC	HC	2	1	1
CD8+CCR9+	PSC-UC	UC	2	2	0
CD8+CCR9+	PSC-UC	HC	6	2	4
CD8+CCR9+	UC	HC	3	0	3

Numbers of differentially expressed genes (DEGs) are for group 1 versus group 2

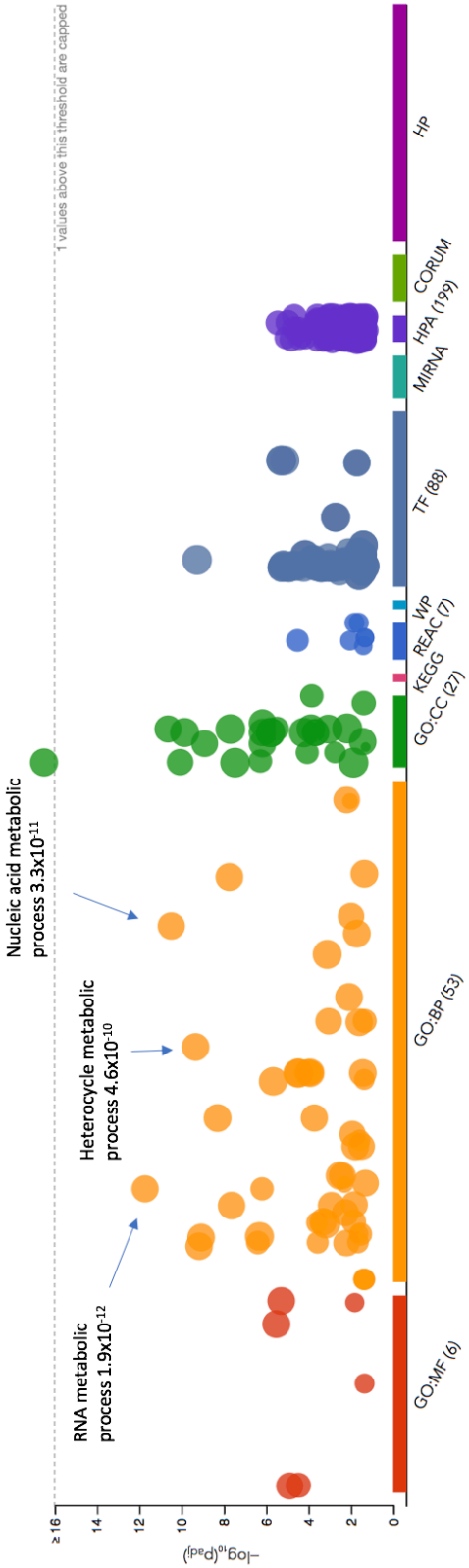

Figure 4.13: Gene ontology pathway analysis for DEGs in T-memory cells of UC compared to HC. Figure generated using g:profiler [236], 20/12/2019.

4.4.2 eQTL mapping

I mapped *cis*-eQTLs for six T-cell subtypes. For four of the T-cell subtypes (T-regulatory, T-memory, CD4+CCR9- and CD8+CCR9- T-cells), the optimal number of expression-derived PCs for detecting maximum number of significant eQTLs at 5% FDR was nine, for T-regs and CD8+CCR9- T-cells this number was eight (Figure 4.14). After extracting all significant eQTL/gene pairs, I detected a median of 1,337 eQTLs per cell type (5% FDR). The largest number of eQTLs (2,804) were detected in T-memory cells and the fewest (901) in CD8+CCR9+ cells (Figure 4.14). This is likely to reflect that T-memory cells were the most abundant cell type, and CCR9+ cells the least abundant, thus influencing the power to detect eQTLs for each cell type. Whilst data for the initial numbers of cells per sample was not available, the lesser-abundant CCR9+ cells underwent more amplification bias compared to the other cell types, as represented by the heterozygosity concordance rate was included within the covariate model. For each cell type, I plotted the position of each eQTL in relation to the gene transcription start sites (TSS), demonstrating that the majority of significant eQTLs were within 100,000 bp of the TSS (Figure 4.15). This is in keeping with the findings of several previous studies that most *cis*-eQTLs occur in close proximity to gene TSS [120, 250, 251].

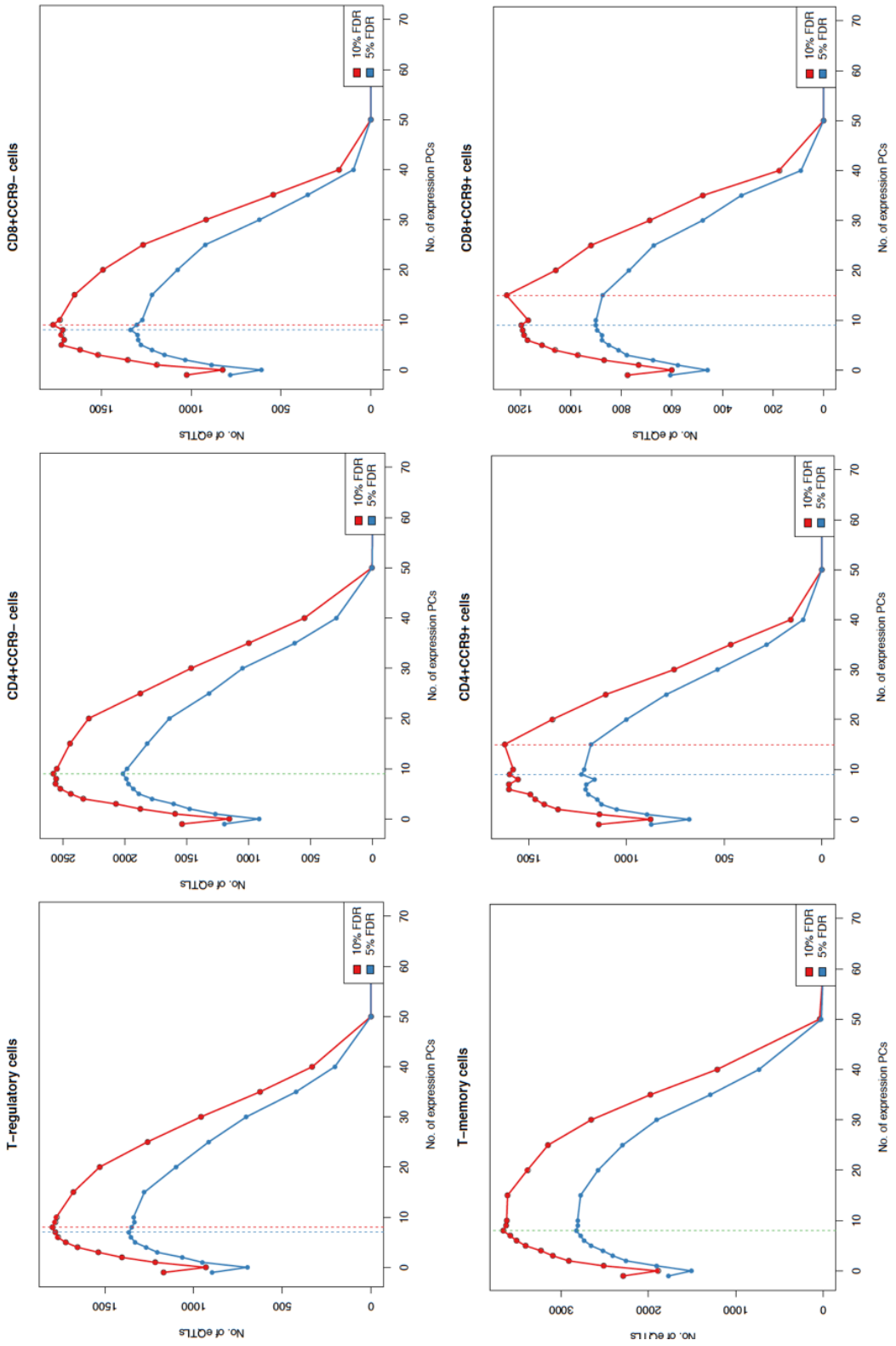

Figure 4.14: Number of significant eQTLs (y-axis) mapped for each individual cell type at 5% (blue line) and 10% FDR (red line), using covariate models with different numbers of gene-expression derived PCs from zero to fifty (x-axis).

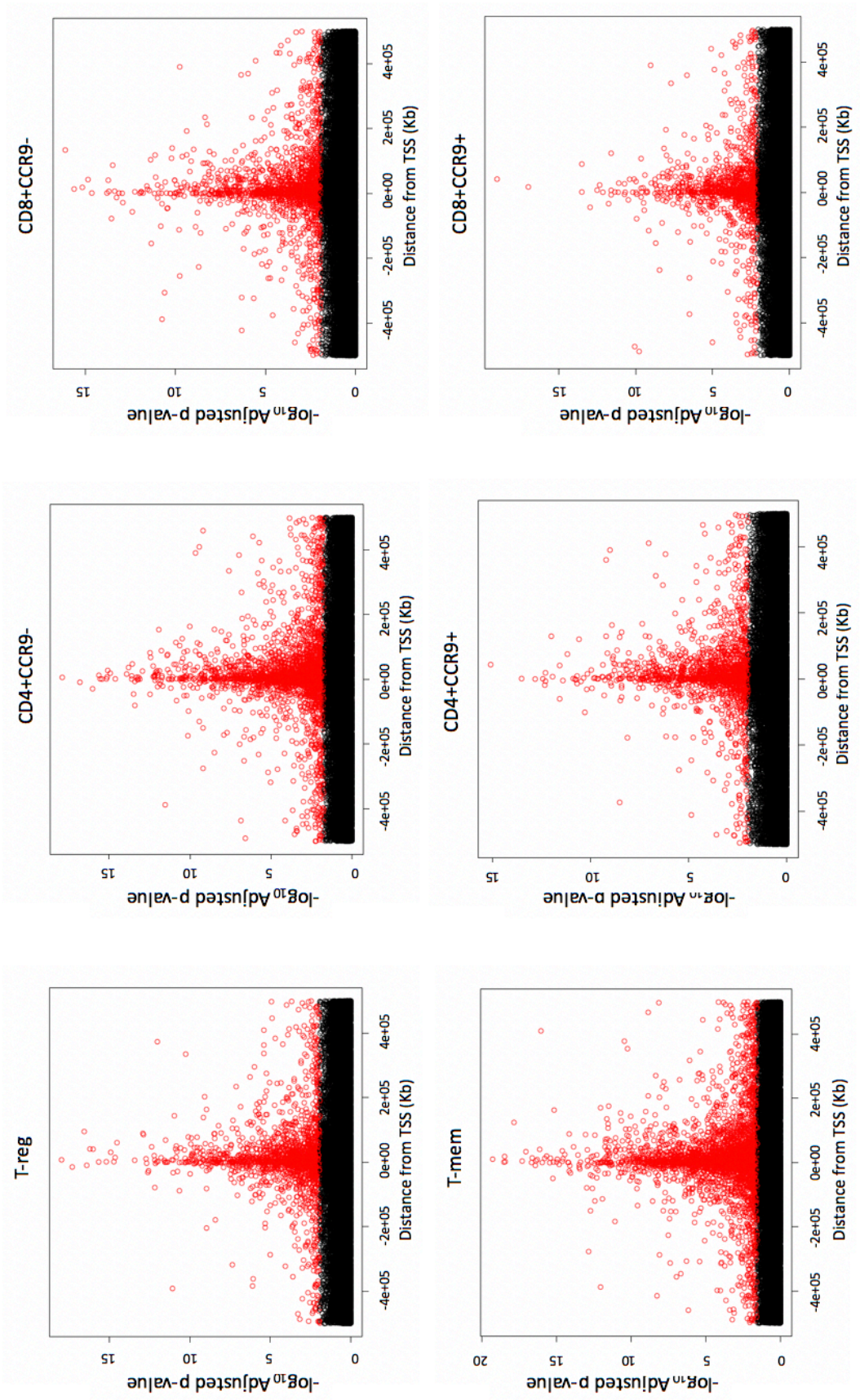

Figure 4.15: Distance from transcription start site (TSS) for each significant eQTL (coloured red for those less than 5% FDR) per cell type.

4.4.3 Shared and tissue-specific eQTLs

With *mashR*, I identified a set of 10,459 significant unique eGenes (5% FDR). This number was more than three times the sum of all significant, unique eGenes detected in the individual cell-type analysis, demonstrating the enormous boost in power provided by the aggregation of measurements across the six cell types to improve the estimates of the β/SE's. Of these 10,459 unique eGenes, 87% (9,176) were shared across all 6 cell types, 4.7% (489) were specific to a single cell type. The distribution of eQTL-sharing across the six cell types is shown in Figure 4.16. These data suggest that the vast majority of eQTLs are shared across all six T-cell subtypes, with very few cell-type specific eQTLs. This finding is not unexpected given that all six of these cell types are subsets of peripheral blood T-cells subject to similar disease conditions. GO analysis of the eGenes using g:profiler [236] did not highlight any gene sets or pathways enriched for cell-type specific or shared eQTLs.

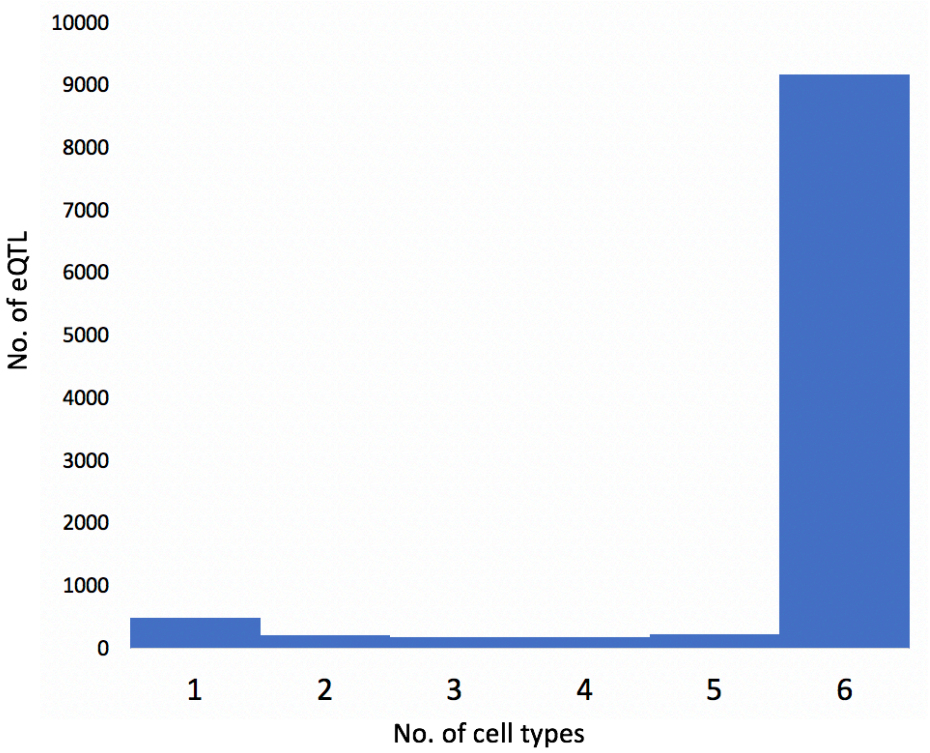

Figure 4.16: Number of cell-type specific and shared QTLs.

4.4.4 Colocalisation of disease-risk loci with eQTL

To identify eQTLs with a causative role in PSC pathogenesis, I performed colocalisation of the fifteen PSC risk loci reported by Ji *et al* [42], with the PSC GWAS summary statistics and eQTL data for each individual T-cell subtype. Two of the fifteen risk loci colocalised (PP4≥0.8) with eQTLs in one or more T-cell subtypes (Table 4.4).

Table 4.4: Colocalisation of PSC risk loci with eQTLs mapped in individual cell-types and eQTLs mapped with *mashR*

					Individual cell type analysis			*Mashr* analysis		
Chr	GWAS SNP	Condition	eGene	Cell type	PP4	eQTL Beta	eQTL p-value	PP4	eQTL post mean (beta)	eQTL lfsr (p-value)
11	rs663743	PSC	*AP003774.1*	T-reg	0.99	0.98	7.27E-06	0.99	1.00	1.90E-18
				T-mem	0.95	1.07	1.35E-07	0.95	0.82	1.71E-14
				CD4+CCR9-	0.95	0.96	1.83E-05	0.99	0.93	3.15E-15
				CD4+CCR9+	0.44	0.86	8.95E-04	0.72	0.81	3.41E-11
				CD8+CCR9-	0.70	0.88	4.38E-04	0.96	0.83	4.44E-12
				CD8+CCR9+	0.21	0.75	2.10E-03	0.00	0.74	5.36E-10
21	rs1893592	PSC	*UBASH3A*	T-reg	0.09	-1.32	1.06E-02	0.98	0.67	4.81E-08
				T-mem	0.91	0.93	4.83E-04	1.00	0.83	1.29E-11
				CD4+CCR9-	0.29	0.79	3.37E-02	0.99	0.77	4.42E-10
				CD4+CCR9+	0.47	0.88	6.69E-03	1.00	0.67	3.87E-10
				CD8+CCR9-	0.23	0.77	5.26E-02	0.99	0.68	9.10E-09
				CD8+CCR9+	0.02	0.82	4.71E-01	0.00	0.60	1.00E-06

Colocalisation of the Chromosome 21 rs1893592 PSC risk locus demonstrated that this locus was an eQTL of *UBASH3A* in T-memory cells. Whilst this is in keeping with my previous finding of colocalisation of this locus with an eQTL for *UBASH3A* in both T-regs and CD4+ naive T-cells in Chapter 3, there was no evidence from the individual cell type analysis to support colocalisation with this eQTL in the other T-cell subsets (all PP4<0.5) (Table 4.4). To identify if this GWAS risk locus was an eQTL of *UBASH3A* across all T-cell subsets, I conducted colocalisation with the eQTL data from the *mashR* analysis. Colocalisation with the *mashR* data supported that this risk locus colocalised with an eQTL for *UBASH3A* across five of the six cell types, including T-regs, T-mems, CD4+CCR9-, CD4+CCR9+ and CD8+CCR9- T-cells, with PP4≥ 0.98 (Figures 4.17 and 4.18). This supports the finding that the Chromosome 21 rs1893592 SNP is an eQTL of *UBASH3A* across most T-cell sub types. Furthermore, plotting of the *UBASH3A* eQTL at this SNP confirmed that the PSC risk increasing rs1893592*A allele reduced expression of *UBASH3A* across all T-cell subtypes (Figure 4.19), in keeping with my previous findings for this locus in Chapter 3.

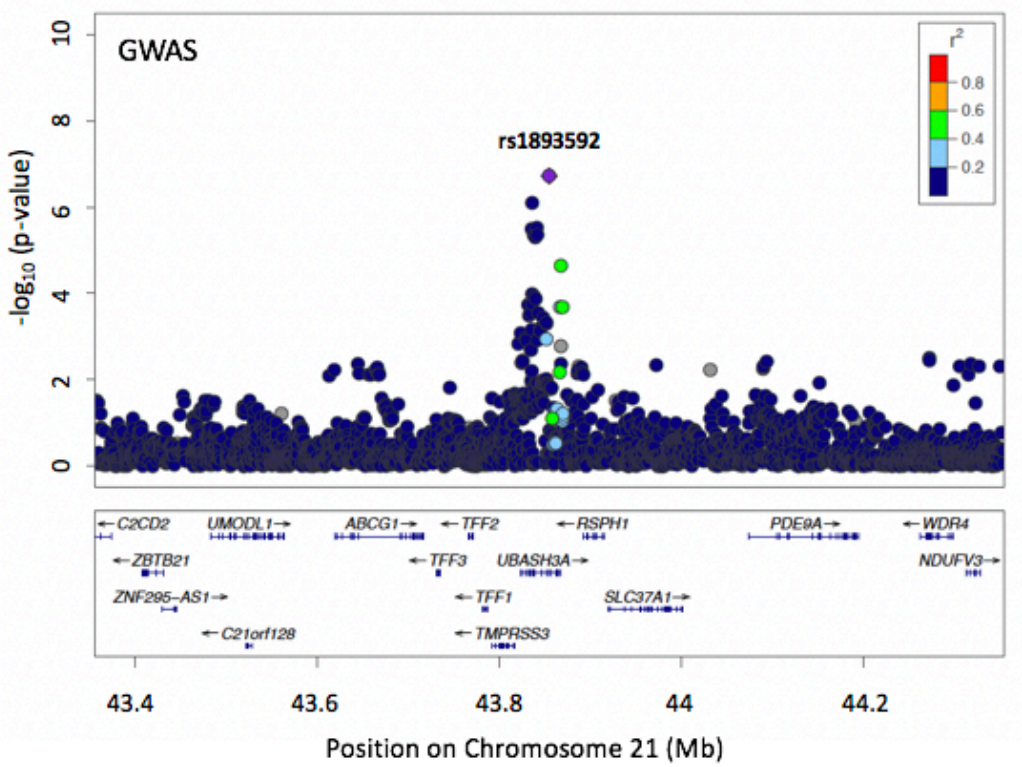

Figure 4.17: Regional association plot for the Chromosome 21 rs1893592 risk locus in PSC GWAS data.

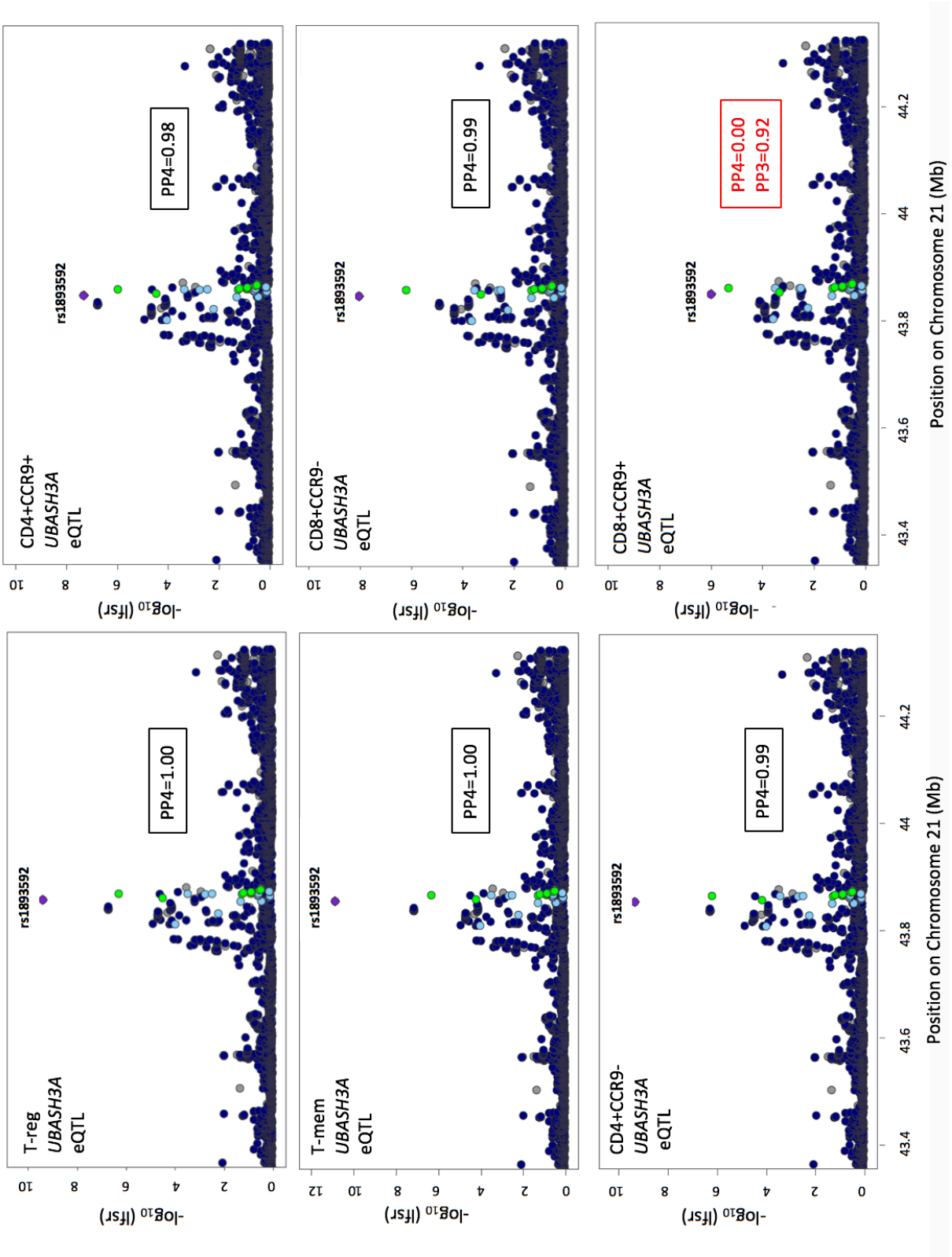

Figure 4.18: Regional association plots for colocalisation between PSC GWAS and eQTLs for *UBASH3A* in T-cells at Chromosome 21 rs1893592 risk locus, using *mashR* eQTL data.

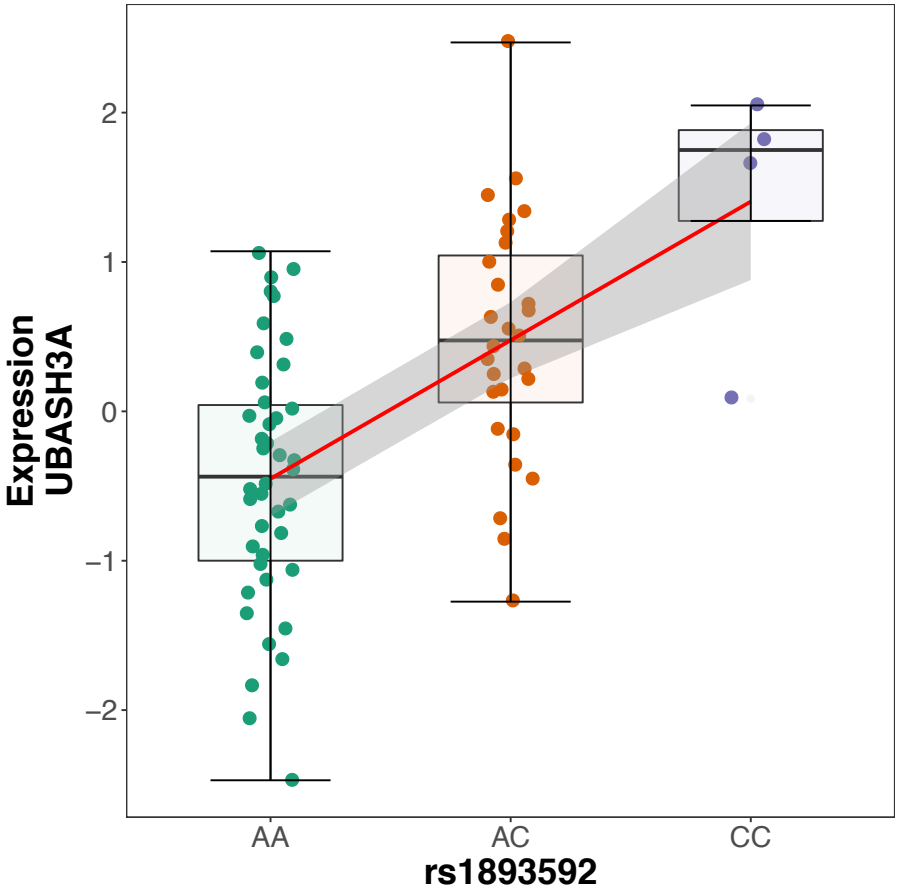

Figure 4.19: Expression of *UBASH3A* according to Chromosome 21 rs1893592 genotype in T-memory cells.

The second PSC risk locus which colocalised with an eQTL in one or more T-cell subsets was the Chromosome 11 rs663743 PSC risk locus. This locus colocalised with an eQTL for *AP003774.1* in three of the six T-cell subtypes; T-regs, T-mems and CD4+CCR9- T-cells with $\geq$ 95% PP (PP4) of causality (Figure 4.20). In addition, there was some evidence to support colocalisation of this locus with an eQTL for *AP003774.1* in CD8+CCR9- T-cells with PP4 of 0.70. Following *mashR* analysis, the strength of the association for this eQTL increased across all six cell types (Table 4.4). Subsequent colocalisation of this locus within the *mashR* eQTL data supported the finding that this PSC risk locus colocalised with an eQTL for *AP003774.1* in four of the six cell types including T-mem, T-reg, CD4+CCR9- and CD8+CCR9- T-cells (PP4 of $\geq$ 0.95) with some additional evidence (PP4=0.72) to support colocalisation with CD4+CCR9+ T-cells (4.21). Plotting of the *AP003774.1* eQTL at rs663743 confirmed that the PSC risk increasing rs663743*G allele reduced expression of *AP003774.1*, with a consistent direction of effect across all cell types (Figure 4.22).

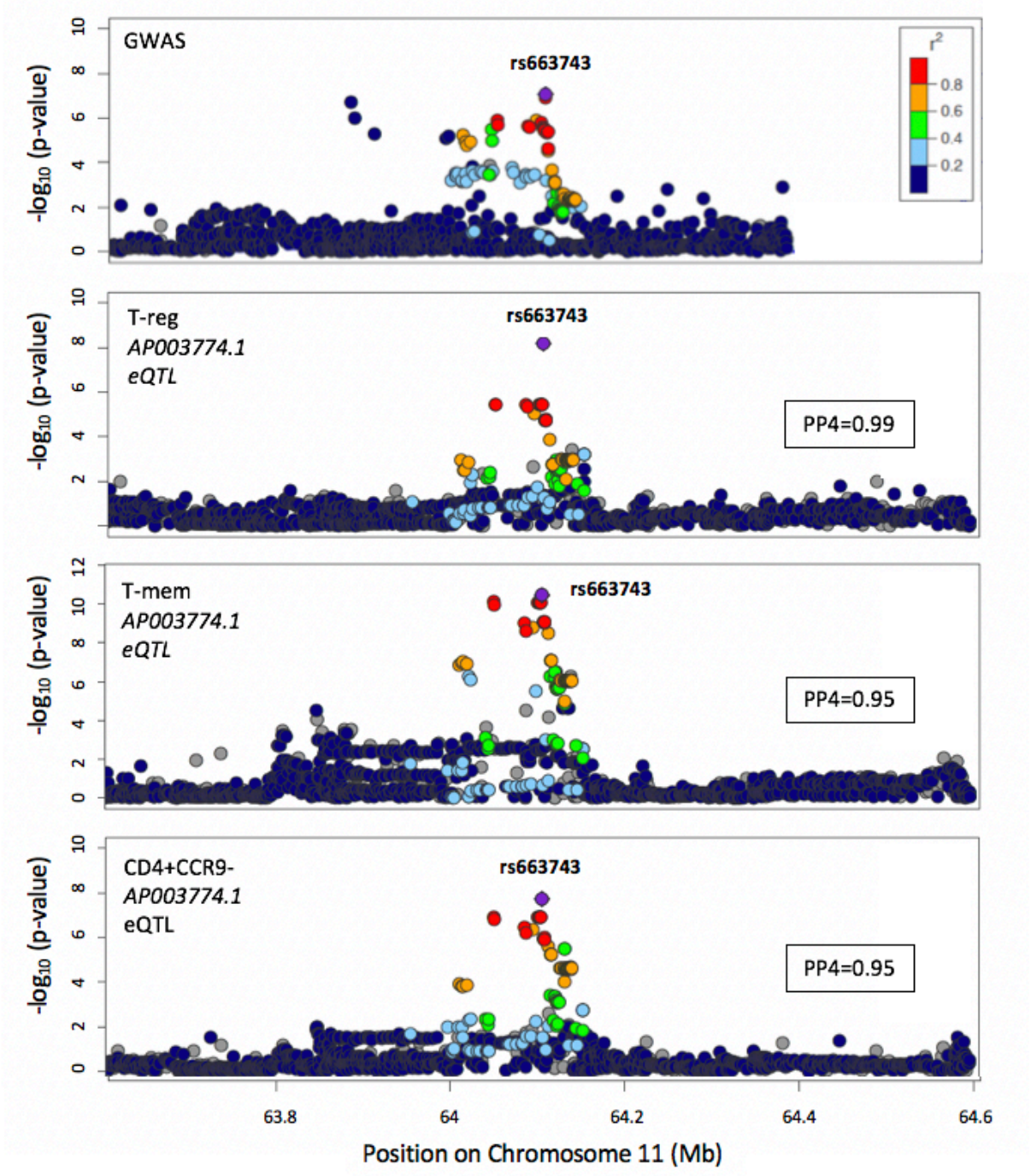

Figure 4.20: Colocalisation between PSC GWAS and *AP003774.1* eQTL data from the individual cell-type analysis, at the chromosome 11 rs663743 PSC risk locus.

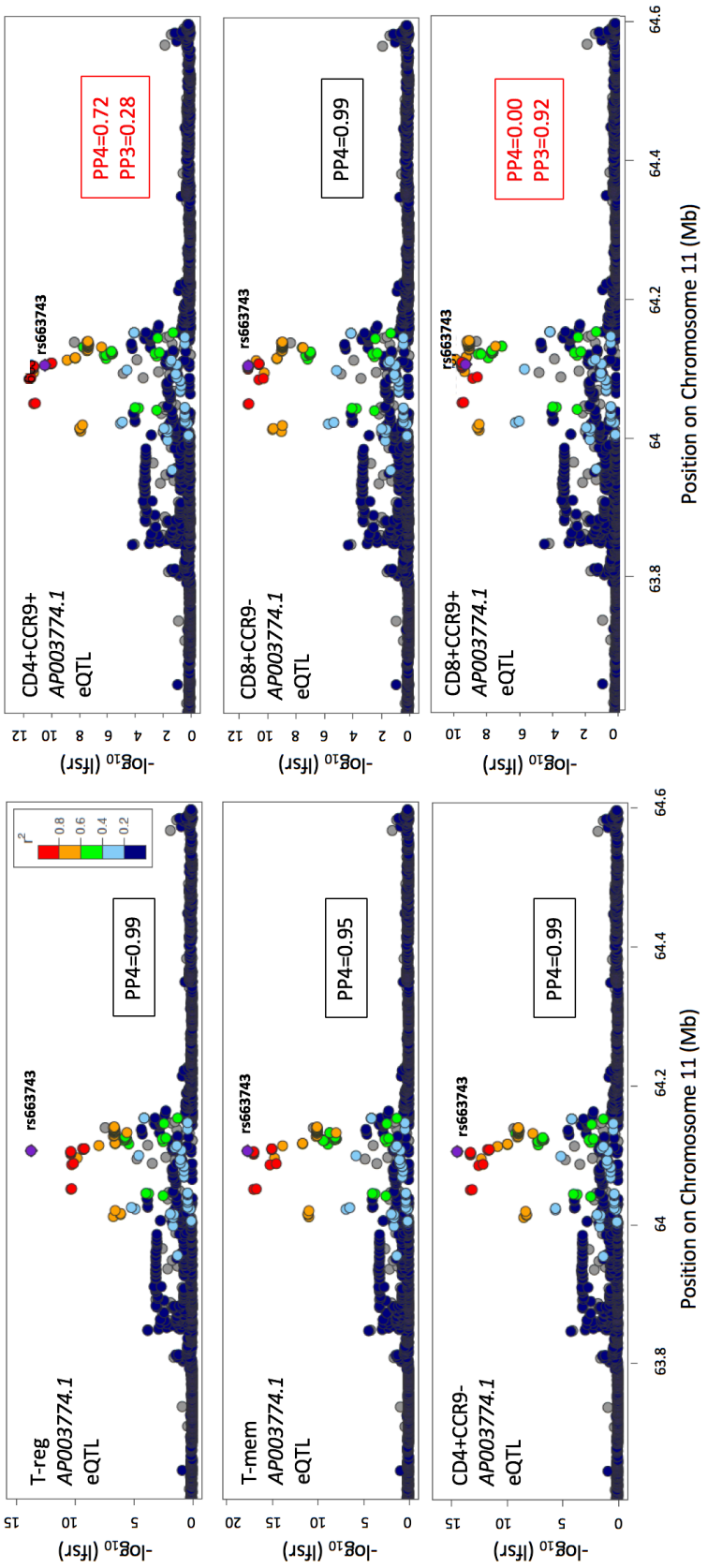

Figure 4.21: Colocalisation between PSC GWAS and*AP003774.1* eQTL data from the *mashR* analysis, at the chromosome 11 rs663743 PSC risk locus.

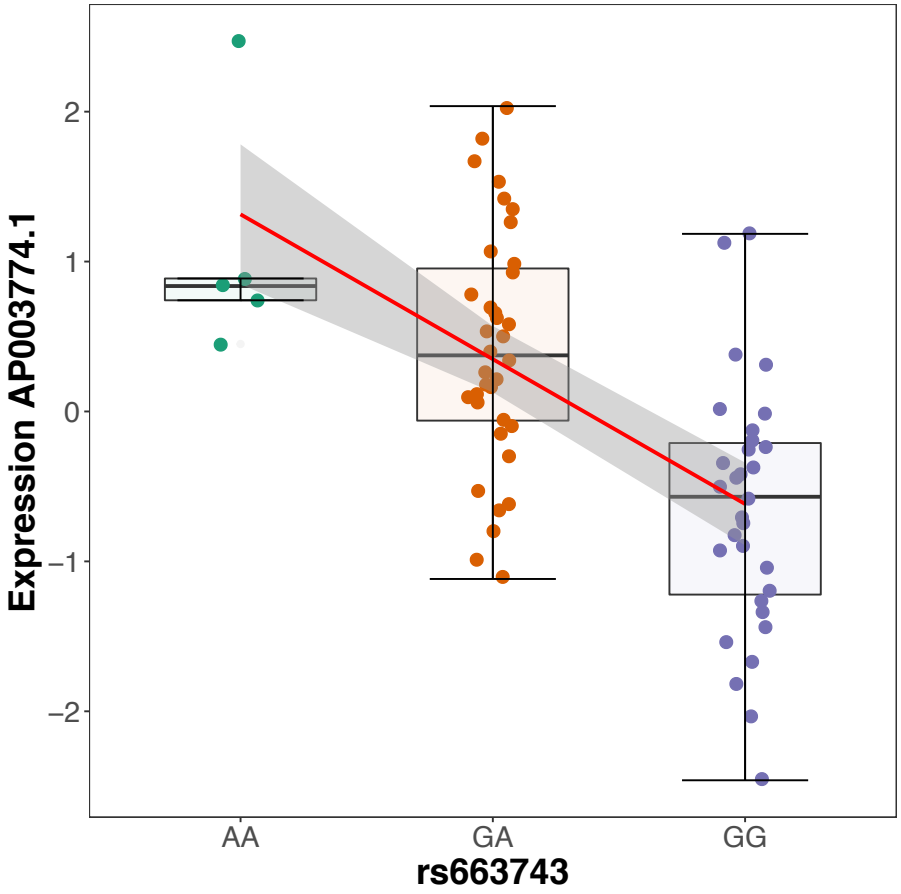

Figure 4.22: Expression of *AP003774.1* according to Chromosome 11 rs663743 genotype in T-regulatory cells.

AP003774.1 is a long non-coding RNA or lncRNA. LncRNA's are defined as transcripts with lengths exceeding 200 nucleotides that are not translated into protein. Whilst the function of the majority of lncRNAs are unknown, it has been shown that lncRNAs are themselves important regulators of gene expression, via interactions with transcription factors or epigenetic modifiers [252, 253]. LncRNAs thus provide a link between non-coding variants and protein-coding genes. Moreover, there is accumulating evidence that lncRNAs are important regulators of both immune cell differentiation and the innate and adaptive immune responses [254–256]. They have also been implicated in the pathogenesis of several IMDs, including (but not limited to) SLE, RhA, T1DM and MS [257–259]. Indeed, one study that mapped *cis*-eQTLs at 460 IMD-associated SNPs found that >10% affected the expression of a lncRNA [260]. Whilst little is known about *AP003774.1*, according to GTEx, this lncRNA is highly expressed in PSC-relevant tissues including colon, small intestine and whole blood (Figure 4.23) [176]. In addition, a search of the database for immune cell eQTL expression epigenomics (DICE) demonstrated that amongst immune cells, *AP003774.1* is most highly expressed in T-cells and NK cells, with lower expression in monocytes [261]. In Chapter 2, I demonstrated that this same region overlaps both

promoter and enhancer elements in multiple PSC-relevant tissues, suggesting plausible mechanisms via which this eQTL for *AP003774.1* may interact with epigenetic modifiers to regulate expression of other genes in the region. More specifically, this locus overlaps H3K27me3, a marker of an inactive or silenced regulatory region, in keeping with the PSC risk increasing allele reducing expression of *AP003774.1* (Figure 4.22). Interestingly, Ricano-Ponce *et al* demonstrated that expression of *AP003774.1* is also linked to another IMD, MS, where the lead GWAS SNP for the MS risk locus (rs694739 at Chr11:64097233, build 37) has been shown to decrease the expression level of *AP003774.1* in PBMCs [260]. Whilst this region has not been fine-mapped in MS, the MS lead SNP, rs694739, lies close to the fine-mapped SNP for this locus in PSC (rs663743 at Chr11:64107735) and both SNPs are in high LD with one another (r^2=0.74). In previous chapters I show that this same rs663743 risk locus in PSC colocalises with a monocyte eQTL for another gene, *CCDC88B*, which is not expressed in T-cells. It is therefore of particular note that Ricano-Ponce *et al* similarly observed that this same MS SNP also affected the expression of *CCDC88B* in PBMCs and that many SNPs associated with IMDs can affect the expression of more than one gene within a 500Kb region. It is therefore likely that this PSC risk locus functions as an eQTL for two different genes in two different cell types; *AP003774.1* in T-cells and *CCDC88B* in monocytes.

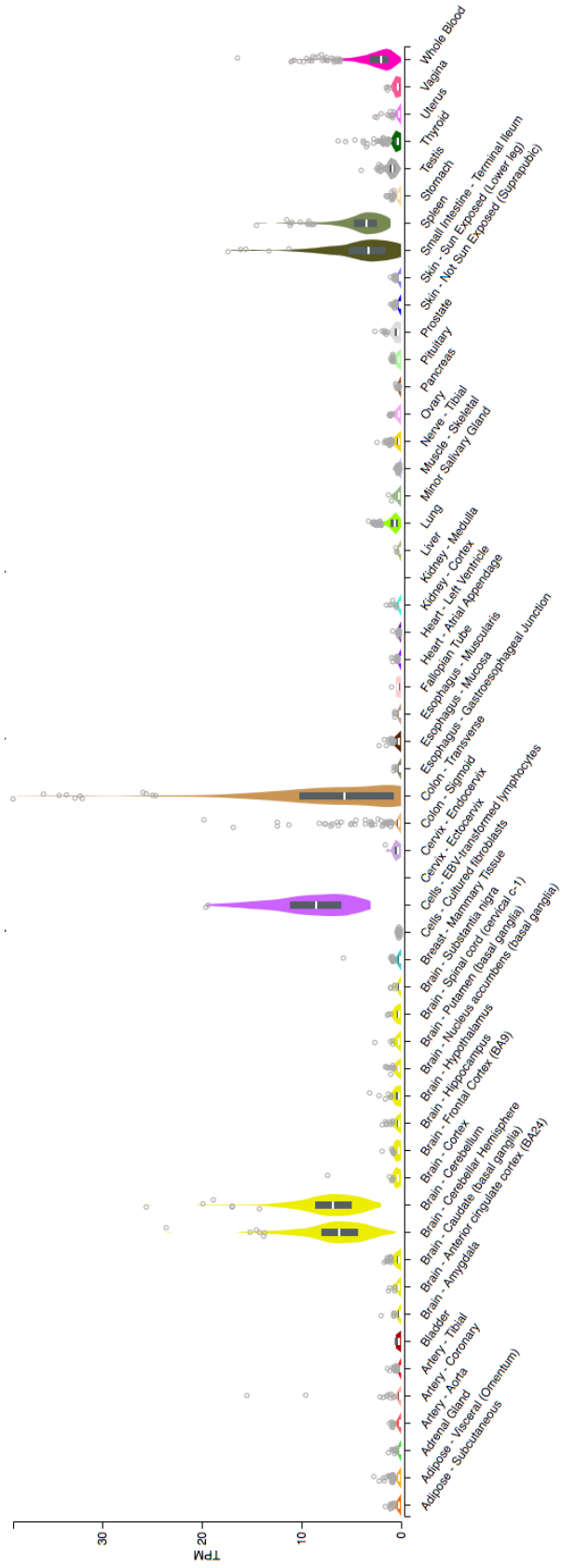

Figure 4.23: Expression of *AP003774.4* across multiple human tissues (figure generated by GTEx portal, 25/02/20 [176]).

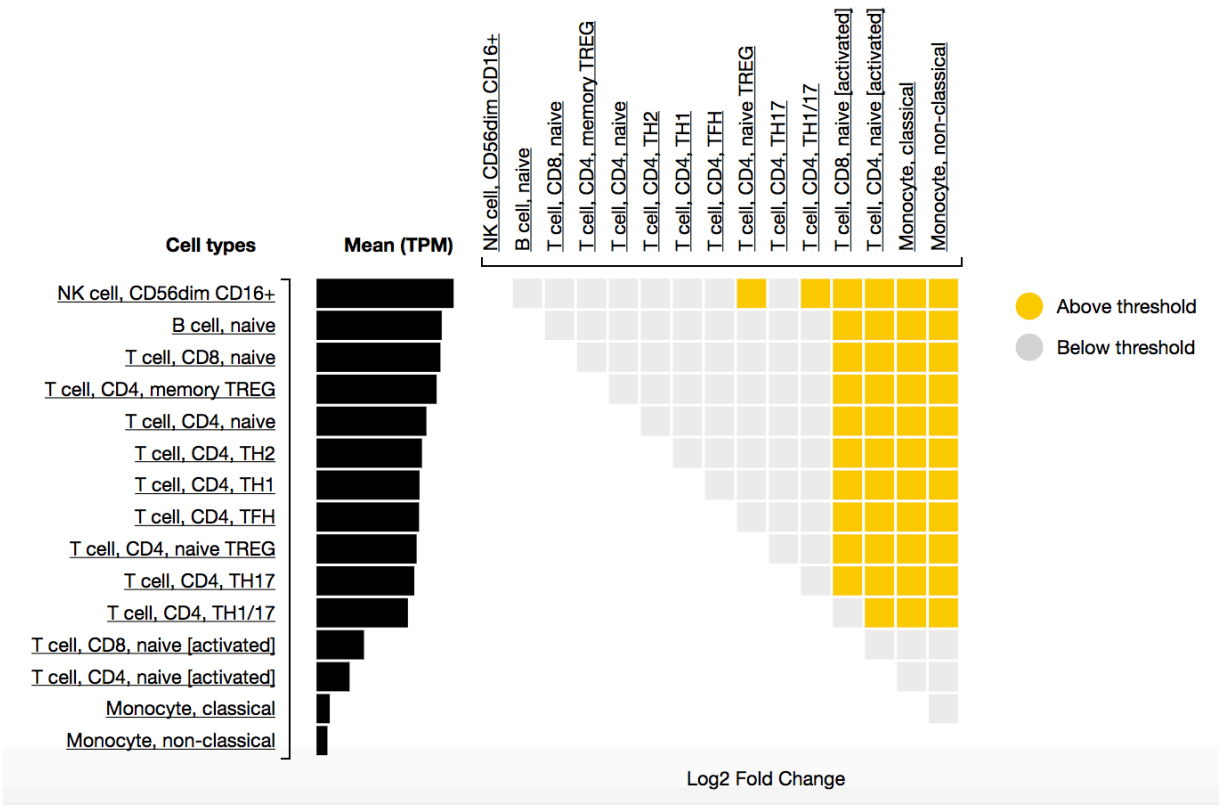

Figure 4.24: Expression of *AP003774.4* across multiple immune cell types (figure generated by the Database of immune cell eQTL expression [261], 26/02/2020).

Many genetic risk loci are known to be shared across multiple IMDs and similar eQTL studies have performed colocalisation of eQTLs with a range of IMDs. The majority of samples for this study were derived from patients with UC. In an effort to identify other IMD risk loci that function as eQTLs, I performed colocalisation of T-cell eQTLs with UC, CD and two other IMDs; RhA and T1DM. I identified ten IMD risk loci that colocalised with eQTLs for one or more genes. The results of colocalisation with all IMDs are shown in Table 4.5, however given that the focus of this thesis is PSC, only those IBD risk loci that colocalised with T-cell eQTLs are discussed further.

This analysis identified two UC risk loci and one CD risk locus that colocalised with T-cell eQTLs, thus identifying several genes involved in inflammatory or immune pathways with a potential causal role in IBD. Of note, the UC Chromosome 7 rs4728142 risk locus colocalised with an eQTL for *IRF5* in T-memory cells. *IRF5* is a transcription factor which forms one of the major inflammatory pathways, crucial for activation of the

Table 4.5: Colocalisation of non-HLA GWAS risk loci for immune-mediated diseases and T-cell eQTL

Chr	GWAS SNP	Disease	eGene	Cell type	PP4	eQTL Beta	eQTL p-val
1	rs3180018	UC	*GBAP1*	T-reg	0.91	-0.74	3.88E-04
				T-mem	0.98	-1.01	1.30E-10
				CD4+CCR9-	0.98	-0.92	2.01E-07
				CD4+CCR9+	0.98	-0.91	7.66E-07
		UC	*THBS3*	CD4+CCR9-	0.98	0.88	5.59E-07
1	rs2317230	RhA	*FCRL3*	CD8+CCR9-	0.93	0.82	3.29E-04
5	rs7731626	RhA	*IL6ST*	T-reg	0.97	-0.86	2.08E-04
				T-mem	0.90	-0.81	6.16E-04
		RhA	*ANKRD55*	T-mem	1.00	-1.00	5.94E-07
				CD4+CCR9-	1.00	-0.95	9.63E-06
				CD4+CCR9+	0.86	-0.86	1.20E-03
7	rs4728142	UC	*IRF5*	T-mem	0.86	0.77	4.71E-05
11	rs663743	PSC	*AP003774.1*	T-reg	0.99	0.98	7.27E-06
				T-mem	0.95	1.07	1.35E-07
				CD4+CCR9-	0.95	0.96	1.83E-05
11	rs968567	RhA	*FADS1*	T-reg	0.89	1.51	1.38E-07
		RhA	*FADS2*	T-reg	0.98	1.60	2.01E-09
				T-mem	1.00	1.58	2.23E-09
				CD4+CCR9-	0.99	1.58	5.40E-09
				CD4+CCR9+	0.98	1.47	2.82E-07
				CD8+CCR9-	0.96	1.56	9.58E-09
				CD8+CCR9+	0.95	1.46	4.22E-07
12	rs4760341	T1DM	*SUOX*	T-reg	0.80	-0.71	1.06E-03
14	rs941576	T1DM	*WARS*	T-reg	0.85	1.07	3.90E-05
				T-mem	0.93	1.19	1.46E-07
				CD4+CCR9-	0.80	1.35	2.41E-08
				CD8+CCR9-	0.97	-1.03	1.37E-05
19	rs4802307	CD	*PPP5C*	T-mem	0.85	-0.92	1.63E-06
				CD4+CCR9-	0.86	-0.99	2.38E-08
				CD8+CCR9-	0.83	-0.82	1.48E-04
21	rs1893592	PSC	*UBASH3A*	T-mem	0.91	0.93	4.83E-04
22	rs909685	RhA	*SYNGR1*	CD8+CCR9+	0.98	1.14	7.15E-04

UC; Ulcerative colitis, CD; Crohn's Disease, RhA; Rheumatoid arthritis, T1DM; Type 1 diabetes mellitus

pro-inflammatory cytokines IL-6, IL-12 and TNF-α [262, 263]. Its expression is induced in lymphocytes by activation of the Toll-like receptor (TLR) 7 and 9 pathways and polymorphisms within this gene have been associated with SLE, RhA, MS, Sjogren's syndrome, psoriasis and IBD [56, 264]. Although there are no existing drugs targeting this gene, it is widely considered to be a promising future target [265]. Furthermore, the Chromosome 1 rs3180018 UC risk locus colocalised with an eQTL for *GBAP1* in T-reg, T-memory, CD4+CCR9- and CD4+CCR9+ T-cells (PP4≥ 0.91). Interestingly this differs from the previously reported candidate genes for this UC locus, *SCAMP3* and *MUC1*. However a causal role for *GBAP1* in IBD has been further supported by the fact that this same variant has been shown to increase expression of *GBAP1* in a peripheral blood eQTL study of patients with CD, resistant to anti-TNF treatment [266]. *GBAP1* is an expressed pseudogene which is known to regulate *GBA* levels, a gene encoding lysosomal glucocerebrosidase and the major predisposing gene involved in Parkinson's disease (PD) pathogenesis. It functions as a competing-endogenous RNA (ceRNA), acting as a microRNA (miRNA) sponge, resulting in subsequent *GBA* degradation [267]. To date, there have been no studies investigating the potential role of *GBAP1* in relation to IBD pathogenesis, however given the emergence of therapies modulating glucocerebrosidase activity in PD, further investigation of this pathway outside of the central nervous system and in the context of UC pathogenesis may be warranted [268].

4.5 Discussion

In this study, I develop the first eQTL maps of peripheral blood T-cell subsets in patients with PSC. Using recently published methods to estimate patterns of similarity across cell-types and thus improve estimates of effect, I was able to identify >10,000 unique eQTLs in at least one or more of the six T-cell subsets. Furthermore, by performing colocalisation of disease risk loci with eQTLs in PSC-specific T-cell subsets, I was able to identify the genes perturbed by two PSC risk loci, in addition to three IBD, four RhA and two T1DM risk loci.

An important finding from this work is the identification of a lncRNA with a potentially important role in PSC causal pathogenesis. The Chromosome 11 rs663743 PSC risk locus functions as an eQTL of *AP003774.1*, which is highly expressed in PSC-relevant tissues including colon, small intestine and whole blood, as well as T-cells and NK-cells. Indeed, expression of this lncRNA has also been linked to MS, where an MS risk locus in this region has also been shown to decrease the expression level of *AP003774.1* in PBMCs [260]. Further work to fine-map the causal variant for this signal in MS is needed to establish if the same causal variant is responsible for the effects seen in PSC and MS. Nevertheless, these findings suggest that *AP003774.1* may have an important role in the

immune-regulatory pathways of T-cells and further study is warranted to establish how reduced expression of this lncRNA might potentiate increased risk of IMD. One means of identifying other genes within the same biological pathway, affected by this risk locus would be to map *trans*-eQTL in the same cell types. Due to their smaller effect sizes and the large numbers of tests required with all genes across the genome, *trans*-eQTL mapping requires much larger sample sizes than available in this study (although this is less of an issue with a targeted trans-eQTL study). However, the finding of more distant genes affected by this same risk variant may identify the relevant biological pathway for further functional investigation.

The findings of this study confirm *UBASH3A* as an important gene in the causal pathogenesis of PSC, a finding that appears, from the analyses outlined in this thesis, to be specific to T-cells. The PSC risk increasing variant results in a reduction of *UBASH3A* expression at the Chromosome 21 rs1893592 PSC risk locus in T-cells. This same risk locus has been associated with several other IMDs including T1DM, CeD and RhA [169, 216]. Furthermore, an RNA sequencing study has demonstrated reduced expression of *UBASH3A* in the PBMCs of patients with SLE [269]. UBASH3A functions to attenuate the signal transduction of NF-κB upon TCR stimulation, by suppressing the activation of the I-κB kinase complex, lending biological plausibility to its role in IMD pathogenesis [216]. There are, to date, no known drugs targeting the *UBASH3A* gene, however there are several therapeutic options for targeting the NF-κB/I-KKβ pathway. Proteasome inhibitors, such as bortezomib and carfilzomib are known modulators of targets in the NF-κB/I-KKβ pathway. In addition, the widely available drug, acetylsalicylic acid or Aspirin, is an inhibitor of IKKβ [270]. Notably, whilst there have been no randomised controlled trials of Aspirin use in PSC, there are case-control data to support a chemoprotective role for Aspirin in the development of de-novo cholangiocarcinoma, which is one of the serious complications of PSC [271, 272]. Further study of the potential therapeutic effects of Aspirin and other modifiers of the NF-κB/I-KKβ pathway in PSC are therefore warranted.

One of the most important limitations of this, and indeed many eQTL studies, is the sample size. This study included ∼450 samples from ∼75 individuals, which was at the lower limit to powerfully detect a significant number of eQTLs. Using DGE, I demonstrated transcriptional equivalence between T-cell subsets in the PSC-UC and lone UC groups, supporting the amalgamation of disease groups to improve subsequent power to detect eQTLs. One important analysis, which was not possible due to the small sample size in each individual disease group, would be to examine the effects of disease-specific eQTLs. For example, identifying those eQTLs which are active in PSC-UC, but not UC may point to important causal biological pathways for PSC. Despite sample size limitations, the use of stringent quality control measures enabled me to robustly identify a total of ∼3,000 unique eGenes across all T-cell subtypes from the individual cell-type analysis, increasing

this number to ∼10,000 with *mashR*. For those PSC loci which colocalised with an eQTL in one or more T-cell subtype, analysis of the *mashR* eQTL data enabled me to identify PSC risk loci that colocalisaed with the same eQTL across multiple T-cell subtypes. Of the ∼10,000 eQTLs identified in this study, >85% were shared across all six T-cell subtypes. The finding that the majority of eQTLs in this study were shared across all six T-cell subtypes is likely to be explained by the relative similarity between the T-cell phenotypes studied in this analysis; all six cellular subtypes were CD3+ T-cells and four were CD4+. During the design of this study, it was hypothesised that the acquisition and analysis of six different T-cell subsets from each donor would allow the detection of cell-type specific eQTLs. However, the resultant benefit of multiple T-cell subtypes from each donor was, in fact, to enable the estimation of patterns of similarity across conditions or cell-types using *mashR*, to improve accuracy of effect estimates and thus identify greater numbers of eQTLs. In a rare diseases such as PSC, where patient recruitment for sample donation is limited by the number of sample donors, this may be an useful future mechanism to improve eQTL mapping. The vast majority of samples in this study were from patients with active PSC and UC, or UC alone. Whilst it is likely that mapping of eQTLs in cell-types that have been subject to the active disease state may have uncovered some additional eQTLs not active in the healthy state, this may only be evident in studies that are well-powered enough to detect those effects. Although this study focused on deriving samples from PSC and UC patients, I also identified eQTLs that colocalised with RhA and T1DM, suggesting that a more fruitful approach might be to study large cohorts of individuals with RNA-seq data, whether or not they have the disease phenotype.

An important future analysis of this T-cell eQTL data would be to conduct fine-mapping of those colocalising risk loci within the eQTL data. Whereas the previous fine-mapping analysis in Chapter 3 had resolved the Chromosome 21 rs1893592 PSC risk locus to this single causal variant, the Chromosome 11 PSC risk locus was fine-mapped to two potential causal variants. Given that the strengths of association between rs663743 on Chromosome 11 and *AP003774.1* expression are greater than with PSC risk, there is likely to be greater power to fine-map the eQTL data and thus attribute a greater PP of causality to a single causal variant. This would pave the way for future biological studies to analyse the impact of the true causal variant perhaps through CRISPR analysis, or recall by genotype experiments.

The rigorous analysis outlined in this chapter has resulted in the generation of a robust set of eQTL maps for six T-cell subtypes, several of which have not previously been the subject of eQTL mapping efforts, and none of which have been previously mapped in patients with PSC. As demonstrated by the finding of eQTLs that colocalise with other IMD risk loci, the results of these analyses can be relevant and important to variety of IMDs outside of PSC and UC. These eQTL maps, which have revealed important findings

for our understanding of PSC, will also provide a public resource available for further scientific study.