Chapter 3

Statistical colocalisation of Primary Sclerosing Cholangitis risk loci with functional quantitative trait loci

3.1 Introduction

The majority of genetic variants associated with complex disease risk are located within non-coding regions of the genome. In the quest to unravel the function of non-coding risk variants, our next challenge is to identify the precise genes upon which they impact. It is now understood that many non-coding risk variants exert their influence via epigenetic gene regulatory mechanisms and exert a quantitative rather than a qualitative effect upon gene expression. Variation in gene expression is therefore an important mechanism underlying susceptibility to complex diseases. Expression quantitative trait loci (eQTL) are genetic variants that exert a quantitative effect upon gene expression, i.e. the abundance of a gene transcript is directly modified by a genetic polymorphism, usually within a regulatory element. In recent years eQTL mapping methods have been developed, which test the association between genetic polymorphisms and transcript abundance by assaying gene expression and genetic variation on a genome-wide scale, in a large number of individuals. Similar to any complex trait, the abundance of a gene transcript is a quantitative trait that can, with a sufficient sample size, be mapped with considerable power [116]. Variants associated with complex diseases are demonstrably enriched for eQTLs [117]. Nicolae *et al* have shown that SNPs associated with complex traits are significantly more likely to be eQTLs than MAF–matched SNPs chosen from high-throughput GWAS platforms that are not associated with complex traits. Investigating eQTLs in the functional study of genetic risk loci associated with complex diseases such as PSC therefore remains a priority.

In order to further investigate the mechanism via which non-coding genetic variants drive risk of complex disease, one challenge has been the integration of complex trait

association data with eQTL data to measure the plausibility of a shared causal variant between the two traits. Over the past decade, several methods have been developed to try and address this challenge. One of the first methods of assessing whether two traits shared a causal variant was by crude visual comparison of the overlap between two signals. A number of computational tools were developed to facilitate the visual comparison of trait-associated and gene-expression data [183]. For example, a study exploring eQTL data for a particularly gene-dense region on chromosome 17q23 strongly associated with susceptibility to asthma [184], found by visual comparison, that the same asthma-associated variants also had highly significant effects on the expression of *ORMDL3* [185]. However, observation of visual overlap cannot prove a causal relationship between, for example, *ORMDL3* and asthma because the abundance of eQTLs throughout the human genome make the chance finding of an overlap highly likely [186]. Indeed, inference about shared causality between two traits requires a more robust statistical assessment of colocalisation.

Plagnol *et al* proposed a 'proportionality-testing' method which tests a null hypothesis of proportionality of regression coefficients for any set of SNPs across two traits, with the assumption that where there are multiple causal variants, these are shared between both signals [187]. However it has been subsequently demonstrated that this method is biased as a result of having to specify a subset of SNPs on which to base the analysis [188]. Moreover these, and other methods, reliant on individual level genotype data have become impractical with the development of collaborative consortia facilitating the meta-analysis of GWAS data from increasingly large sample sizes. In 2014, Giambartolomei *et al* published *Coloc*, a method to test for colocalisation between two pairs of traits, which overcomes many of these shortcomings by using a Bayesian model with single-SNP summary statistics [189]. *Coloc*, discussed further in the following Methods section, assesses the plausibility of a single shared causal variant driving two traits, requiring densely-genotyped or well-imputed summary statistics that have undergone stringent QC. *Coloc* bases its analysis upon all SNPs within a locus, assuming each SNP is *a priori* equally likely to affect the traits under analysis. Furthermore it estimates the posterior probability (PP) for five different hypotheses ranging from no shared genetic variation between two traits within a region (PP0), to shared genetic variation with the same causal variant driving each signal (PP4). *Coloc* can be applied to any two pairs of traits, including disease traits or functional (epigenetic) traits such as eQTL, histone acetylation marks (histQTL) and methylation marks (methQTL). *Coloc*, and other methods using a similar statistical approach have become the singular method of analysis for performing colocalisation between genetic traits.

Gene expression is the subject of both global and local regulatory variation, i.e. there are eQTLs which act across multiple tissues, in addition to tissue-specific regulatory variation [125]. Colocalisation between disease-associated risk loci and functional traits,

therefore requires careful consideration of the tissue, cell-type or activation state in which the functional trait has been measured. However, identifying the relevant cell-type or stimulated state in which an eQTL is active remains challenging, as demonstrated by several studies, which have sought to address this through the mapping of eQTLS across multiple cell types challenged with multiple stimuli [129, 130]. Importantly, it has been demonstrated that eQTLs are enriched for disease-associated variants in disease-relevant cell- or tissue-types [190, 191]. For example, a recent IBD GWAS and colocalisation study found that a chromosome 2 IBD risk locus co-localised with an eQTL that increased expression of integrin α4 in stimulated monocytes, an eQTL that was not active in unstimulated monocytes [60]. Furthermore, this pathway is already the target of successful therapeutic blockade in IBD, by Vedolizumab, a monoclonal antibody to the α4β7 integrin which inhibits T-cell trafficking to the gut mucosa [83, 84]. Therefore, in order to unravel the molecular basis of disease-specific risk loci, the evidence supports the preferential use of eQTLs measured in disease-relevant tissues for colocalisation. However, paucity of published eQTL data means that colocalisation with eQTLs in mechanistically-related tissue/cell types may be limited by data availability. One interesting finding of a study combining RNA-seq with ATAC-seq (Assay for Transposase-Accessible Chromatin using sequencing) data, found that the majority of stimulus-specific eQTLs with a detectable effect upon chromatin accessibility also altered chromatin accessibility in the unstimulated state [134]. On this basis, colocalisation with other functional QTL, for example chromatin accessibility, histone modification or DNA methylation, may indicate the presence of an eQTL in another (unstudied) stimulation state, in addition to revealing the epigenetic mechanism via which disease-associated risk variants may influence gene expression. Therefore in order to fully understand the functional mechanisms underlying GWAS association signals using colocalisation, it is important to examine the relevant cell type, in the right state of activation, at the right time.

3.2 Chapter overview

Colocalisation is one means of identifying the mechanistic impact of non-coding disease risk loci, by examining whether the same non-coding variant is responsible for regulation of gene expression (i.e. is an eQTL). In this chapter I perform colocalisation between PSC risk loci and functional QTL in multiple immune cell- and gastrointestinal tissue-types. Genetic variation is often shared between several immune mediated diseases (IMDs), implicating the same genes and biological pathways as causal mechanisms for autoimmunity. I therefore perform colocalisation between PSC risk loci and other IMDs to identify risk loci that are PSC-specific and those that are shared. Genetic variants tend to exert a greater effect upon gene expression than upon risk of complex disease. For those risk loci that colocalise with

other regulatory or functional QTL traits, I harness this increased power by fine-mapping the colocalising functional QTL data, in an effort to further refine the fine-mapping results presented in the previous chapter.

3.3 Methods

3.3.1 Colocalisation analysis

To test the plausibility of a single shared causal variant between each of the 22 PSC risk loci, and the same regions in multiple functional QTL and IMD GWAS data sets, I implemented Bayesian tests of colocalisation using R package *Coloc* [189]. Specifically, I used the *coloc.abf* function as it implements approximate Bayes Factor colocalisation methods which can be applied to per-SNP summary statistics. I used full summary statistics for the largest PSC GWAS [42] and the datasets outlined in Table 3.1. The 22 genomic loci for colocalisation were defined as 1Mb regions of interest centred on the most associated or 'lead' PSC GWAS SNP for each locus.

Coloc requires per-SNP summary level data for each of the two input traits. This must consist of all variants within the locus, including those variants that did and did not reach the predetermined threshold for genome-wide significance or false discovery rate (FDR). Colocalisation can be conducted using different combinations of input data for each trait to approximate Bayes factors, depending upon the data available. The first combination of input data includes per-SNP p-values and MAF, sample size and ratio of cases:controls (if using a case-control trait). The second combination includes per-SNP regression coefficients (β) and the variance of these regression coefficients (SE^2), in addition to sample size and ratio of cases:controls. Where available, I used regression coefficients and their variance in preference to p-values and MAFs to approximate Bayes factors, as the former combination is more accurate when using imputed data. Where data availability meant that p-values and MAF were used to approximate Bayes factors, I preferentially used the MAF derived from the same dataset under investigation. Where study-specific MAF data was not available, I used the MAF derived from the UK10K reference panel, as all data-sets included only individuals with European ancestry and thus this was the reference panel that best represented the study population. To interpret the direction of effect of an eQTL on gene expression in the context of the PSC risk allele, I matched eQTL and GWAS reference alleles for all loci. To minimise the chance of combining the wrong alleles, I discarded all A/T and C/G variants that had MAF>0.45.

In this Bayesian method of colocalisation, binary vectors representing a sequence of SNPs by whether each individual SNP is causal (1) or not (0) are paired, with each binary vector representing one trait, and pre-assigned to one of five hypotheses (H0, H1, H2, H3,

H4);

H0: No SNP is associated with either trait.
H1: A SNP is associated with trait 1 (PSC), but no SNP is associated to trait 2 (IMD or eQTL)
H2: A SNP is associated with trait 2 (IMD or eQTL), but no SNP is associated to trait 1 (PSC).
H3: Both traits are associated with genetic variation in the region, but this is driven by different causal variants.
H4: Both traits are associated with genetic variation in the region and share the same causal variant.

For each PSC risk locus tested, the probability of the data for each hypothesis is calculated and the aggregate support (probabilities) for each hypothesis combined with the prior probability, to obtain posterior probabilities for each hypothesis (PP0, PP1, PP2, PP3, PP4). The *Coloc* method uses approximate Bayes factors. Bayes factors are summary measures for the ranking of associations, similar to p-values and are defined as the ratio of the probability of the data under the null and alternative hypotheses [192]. Bayesian methods require the definition of prior probabilities for all five hypotheses. In line with recommendations made by the authors, for GWAS/eQTL analyses I set prior probabilities to 1×10^{-4} for individual trait associations and 1×10^{-6} for the probability of a SNP being associated with both QTL and PSC traits (denoted as the p^{12}). In a study of shared genetic variation between four IMDs (not including PSC or IBD) Fortune *et al* suggested that the selection of priors for colocalisation between two IMD traits should be set at a less stringent threshold between 1×10^{-5} and 1×10^{-6} for the prior probability of a SNP being associated with both traits (p^{12}) [193]. This is due to the expectation of more shared genetic variation between loci of IMDs. In their study, whilst the choice of p^{12} did not change which diseases were associated, the posterior odds for H3:H4 did vary with p^{12}. To inform the choice of priors for colocalisation between PSC and the other IMDs in this study, I tested how varying the prior may impact upon the results of colocalisation. I performed colocalisation between PSC and UC (the IMD expected to show the most genetic overlap with PSC), varying the p^{12} from 1×10^{-4} to 1×10^{-7} and examined the weights of the resulting PP3:PP4.

I performed colocalisation for each of the twenty-two PSC risk loci with the data-sets outlined in Table 3.1. I focused on loci for which the PP for the H4 hypothesis (PP4) was >80%, and subsequently refer to this as evidence of colocalisation when reporting results. I also noted regions for which the PP for the H3 hypothesis (PP3) was >80%, which suggests shared genetic variation between two traits, but a different causal variant driving each signal. Finally I noted regions for which PP4 did not reach the 80% threshold,

but where some of the PP had been attributed to PP0, PP1 or PP2, as this can, in the presence of a low PP3, indicate a loss of power to detect colocalisation.

Coloc makes a number of important assumptions. Firstly it assumes that the two traits undergoing colocalisation have been measured in two datasets of unrelated individuals. The method also assumes that the individuals in both datasets are of the same ethnicity and thus the MAF and LD structure are identical. Because the PSC GWAS data set is derived from individuals of European ancestry, only functional QTL and IMD GWAS data derived from European individuals could be included in this analysis. Resultantly, I excluded one large eQTL meta-analysis of whole blood from 32,000 individuals of many ethnicities [194]. A third *Coloc* assumption is that the true causal variant is included within each set of SNPs, requiring that the dataset for each trait is densely-genotyped or well-imputed. In situations where the true causal variant is not present within both datasets, this tends to result in a decrease in the resulting PP4 statistic. The final assumption of this method is that there is, at most, only one independent association for each trait within the region of interest. It is however not uncommon for genomic regions to contain more than one independent association signal. Indeed, fine-mapping of the PSC GWAS data from the previous chapter supported the presence of 19 independent signals across the 15 fine-mapped PSC risk loci. For those regions in which there is more than one independent signal, *Coloc* considers only the strongest of these distinct association signals.

3.3.2 Functional QTL data

Colocalisation of disease-associated risk loci with functional QTLs requires careful consideration of the choice of cell-type or tissue in which the functional QTL trait has been measured. Those tissues potentially relevant to PSC could be any whole-tissue or cell-type from the gastrointestinal or hepato-biliary systems, or any immune-cell type. To find published and un-published eQTL data for inclusion in my analysis, I performed a literature search of existing eQTL studies. From this, I gathered together 42 functional QTL data-sets covering five gastrointestinal whole tissues, six immune-cell types and five different functional traits including gene-expression (*cis*-eQTL), histone marks (histQTL), DNA methylation (methQTL) and splice site QTL (spliceQTL) data (Table 3.1). All data included for colocalisation in this analysis had been subject of prior QC conducted by the publishing authors.

Datasets used for colocalisation included functional QTL data from the Blueprint epigenome project phase 2 data release [195]. The Blueprint epigenome project is a large-scale research project which aims to generate at least 100 reference epigenomes for distinct haematopoietic cell-types in health and common autoimmune diseases (not including PSC or IBD). Blueprint have isolated CD14+CD16- monocytes, CD45+CD66b+CD16+ neutrophils and CD3+CD4+CD45RA+ naïve T-cells from the peripheral blood of between

Table 3.1: Characteristics of data-sets included in colocalisation analysis

data-set	Tissue type / GWAS	Trait	Condition	Sample size
GTEx v7	Liver	eQTL	unstimulated	153
	Transverse Colon	eQTL	unstimulated	246
	Sigmoid Colon	eQTL	unstimulated	203
	Terminal Ileum	eQTL	unstimulated	122
	Whole Blood	eQTL	unstimulated	369
	EBV-Transformed Lymphocytes	eQTL	unstimulated	117
Blueprint	Naïve T cells	eQTL	unstimulated	171
	(CD3+CD4+CD45RA+)	Methylation	unstimulated	133
		H3K4me1	unstimulated	104
		H3K27ac	unstimulated	142
		PSI	unstimulated	171
Blueprint	Neutrophils	eQTL	unstimulated	192
	(CD45+CD66b+CD16+)	Methylation	unstimulated	197
		H3K4me1	unstimulated	173
		H3K27ac	unstimulated	174
		PSI	unstimulated	192
Blueprint	Monocytes	eQTL	unstimulated	194
	(CD14+CD16-)	Methylation	unstimulated	196
		H3K4me1	unstimulated	172
		H3K27ac	unstimulated	162
		PSI	unstimulated	194
Glinos et al, unpub	T regulatory cells	eQTL	unstimulated	123
	(CD3+CD4+CD25highCD127-)	H3K4me3	unstimulated	73
		H3K27ac	unstimulated	91
		ATAC	unstimulated	88
Panousis et al, unpub	Macrophages	eQTL	CIL (6 and 24 hrs)	83
	(derived from iPS cells)	eQTL	Ctrl (6 and 24 hrs)	81
		eQTL	IFNB (6 and 24 hrs)	84
		eQTL	IFNG (6 and 24 hrs)	84
		eQTL	IL4 (6 and 24 hrs)	85
		eQTL	LIL10 (6 and 24 hrs)	75
		eQTL	MBP (6 and 24 hrs)	44
		eQTL	P3C (6 and 24 hrs)	86
		eQTL	PIC (6 and 24 hrs)	44
		eQTL	PIC (6 and 24 hrs)	45
		eQTL	Prec (Day 0 and 2)	42
		eQTL	R848 (6 and 24 hrs)	83
		eQTL	sLPS (6 and 24 hrs)	81
Kim-Hellmuth et al, 2017	Monocytes	eQTL	unstimulated	134
	(CD14+)	eQTL	LPS (90' and 6hrs)	134
		eQTL	RNA lipofectamine (90' and 6hrs)	134
		eQTL	MDP (90' and 6hrs)	134
Astle et al, 2016	Lymphocyte counts	GWAS		173,480
	Monocyte counts	GWAS		173,480
	Neutrophil Counts	GWAS		173,480
De Lange et al, 2017	Ulcerative colitis	GWAS		12,160
	Crohns Disease	GWAS		12,160
Cordell et al, 2015	Primary Biliary cirrhosis	GWAS		2,764
Bradfield et al, 2011	Type 1 Diabetes	GWAS		9,934
Trynka et al, 2011	Coeliac Disease	GWAS		12,041
Okada et al, 2012	Rheumatoid Arthritis	GWAS		29,880
Beecham et al, 2013	Multiple Sclerosis	GWAS		14,802
Bentham et al, 2015	Systemic Lupus Erythematosus	GWAS		7,219

100-200 healthy adults, followed by epigenomic analyses [196]. These include gene expression, CpG methylation, H3K4me1, a marker for active and poised enhancers, H3K27ac, a marker for active enhancers and active promoters and percentage splice index (PSI) which provides the inclusion level of each exon, indicating perturbation of a splice site. I also included published data from the Genotype-Tissue Expression (GTEx) Consortium v7 [197]. GTEx is an established data resource and tissue bank for the study of the relationship between genetic variation and gene expression in multiple human post-mortem tissues. Included within the GTEx database are whole tissue *cis*-eQTL maps for PSC-relevant tissues including liver, transverse and sigmoid colon, terminal ileum, whole blood and Epstein Barr Virus (EBV)-transformed lymphocytes (immortalised B-cells) isolated from between 100-400 individuals. To try and capture colocalisations with eQTLs only active in the stimulated state, I included published data from an eQTL study of CD14+ monocytes derived from 134 healthy individuals and stimulated with microbe-associated molecular patterns; lipopolysaccharide (LPS), RNA lipofectamine and muramyl dipeptide (MDP) [198]. Two sets of unpublished data, were also included for colocalisation. The first was an eQTL dataset measured in induced pluripotent stem cell (iPSC)-derived macrophages differentiated from the skin fibroblasts of up to 85 healthy donors, and exposed to 13 different states of stimulation. These included stimuli mimicking bacterial, viral and allergic response, and measured at 6 and 24 hour time-points (data kindly provided by Dr Nikolaus Panousis, Postdoctoral Fellow at the Wellcome Trust Sanger Institute). The second was data from an analysis of unstimulated T-regulatory cells (CD3+CD4+CD25highCD127-) derived from the peripheral blood of 70-125 healthy individuals and subject to RNAseq, ChIP-Seq and ATAC-seq (data kindly provided by Dr Daphne Glinos, former PhD student at the Wellcome Trust Sanger Institute). Finally, in order to identify PSC risk loci that colocalised with other IMDs, I downloaded summary statistics for the largest available GWAS study for each of eight IMDs from the GWAS catalogue (https://www.ebi.ac.uk/gwas/). These were UC and CD [60], Primary Biliary Cholangitis (PBC) [199], Type 1 Diabetes (TIDM) [200], Coeliac disease (CeD) [201], Rheumatoid arthritis (RhA) [148], multiple sclerosis (MS) [202] and systemic lupus erythematosus (SLE) [203]. I also conducted colocalisation between PSC risk loci and risk loci associated with lymphocyte, neutrophil and monocyte counts from a GWAS of human blood cell trait variation [204].

3.3.3 Fine-mapping of functional QTL loci

In the previous chapter I presented the results of fine-mapping the PSC risk loci. Fine-mapping is influenced by several factors including the sample size of the cohort, the effect size, the MAF and thus the strength of association of the variants within the locus. One of the challenges of studying a rare complex disease such as PSC is that amassing the GWAS samples sizes comparable to more common IMDs such as T1DM and IBD is not

feasible. Genetic variants tend to exert a greater effect upon gene expression than upon complex disease risk. Therefore, where colocalisation proves that a GWAS trait shares a causal variant with a functional trait, there will often be more power to fine-map within the functional QTL data and resolve the locus to a single causal variant, or small set of credible variants. With the aim of improving upon the fine-mapping of the PSC risk loci described in Chapter 2, I developed the following workflow pipeline (Figure 3.1). For each PSC risk locus I conducted fine-mapping in the PSC GWAS data (Chapter 2), followed by colocalisation with the multiple functional QTLs listed in Table 3.1. Where I observed a PSC-QTL colocalisation, I then fine-mapped the colocalising functional QTL data, using the same methods as described in Chapter 2. Fine-mapping requires an LD matrix, ideally calculated from the original genotype data rather than a reference panel [159], I therefore conducted fine-mapping in those functional QTL traits for which full genotype data was available for the calculation of SNP correlation matrices. Functional trait fine-mapping was therefore limited to the Blueprint data and Glinos *et al's* T-regulatory QTL data.

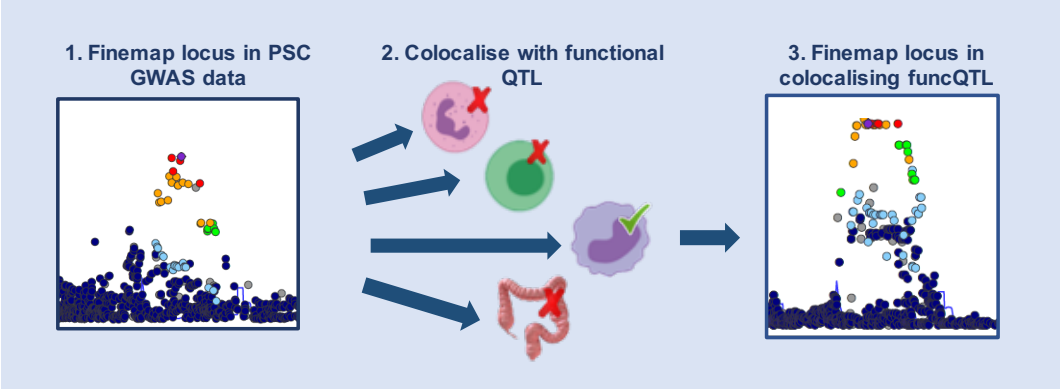

Figure 3.1: Schematic diagram of the GWAS fine-mapping - colocalisation - functional-trait fine-mapping pipeline to resolve the causal variants driving PSC risk loci, and the genes they perturb.

Throughout the analyses described in this chapter, all SNPs are referred to according to their RSID, and all base pair (bp) positions are reported according to Ensembl build 37. For ease of reference, all loci are referred to according to their chromosome and bp position (b37) of the most probable causal SNP from fine-mapping in Chapter 2 and where possible, the gene identified by colocalisation. Where a gene has not been identified by colocalisation with an eQTL, I use the GWAS candidate gene, stipulating where a causal association between a locus and a gene is proven and where it is not.

3.4 Results

I performed colocalisation analysis between the twenty-two non-HLA PSC risk loci and eight other IMDs (Table 3.2). To inform my choice of priors for this analysis, I first tested how varying the prior impacted upon the PP3 and PP4 weights. I performed colocalisation between PSC and UC, varying the p^{12} (prior probability for a SNP being associated with both traits) from 1×10^{-4} to 1×10^{-7}. For 7 PSC risk loci, Figure 3.2 demonstrates how varying the p^{12} changes the weights for PP3 and PP4. Fortune *et al* previously recommended a p^{12} threshold somewhere between 1×10^{-5} and 1×10^{-6}. Although the weights for PP3 and PP4 varied with a p^{12} of 1×10^{-5} and 1×10^{-6}, the results of colocalisation (number of loci for which PP4>80%) were the same. I therefore chose to retain the more stringent of the two p^{12} thresholds, which was set at 1×10^{-6} for all subsequent colocalisation analyses.

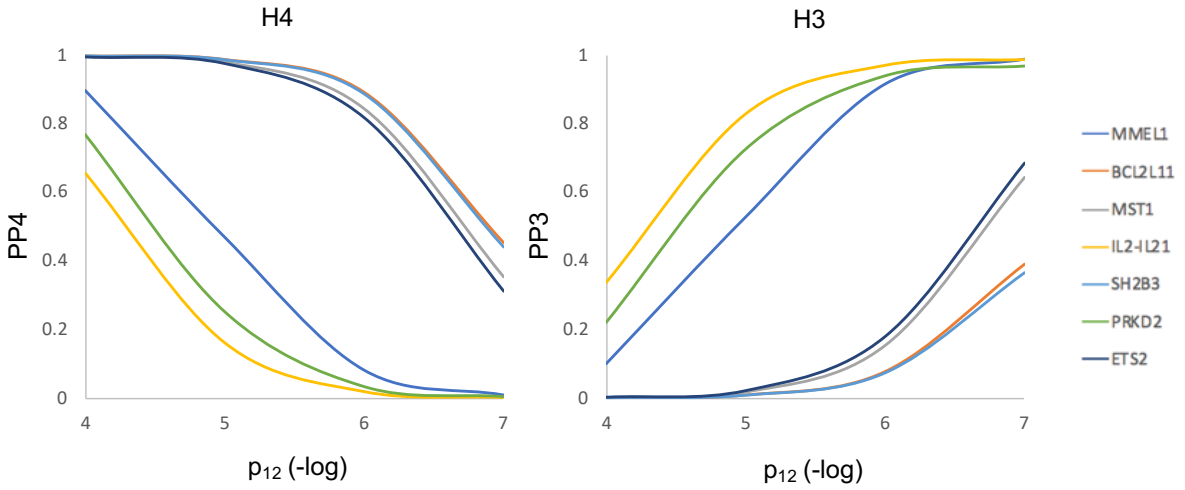

Figure 3.2: Colocalisation between seven PSC risk loci with UC and the evidence for PP4 and PP3 with varying p^{12}.

For those seven risk loci not reaching genome-wide significance in Ji *et al's* data, the results consistently supported no evidence of colocalisation and therefore the results for these seven loci are not subsequently shown. Supporting previous observations of shared genetic architecture between IMDs, eleven of the remaining fifteen PSC risk loci, colocalised with at least one other IMD with PP4>80%. I observed the largest number of colocalisations between PSC and UC, a finding that was expected due to the genetic overlap between PSC and IBD (particularly UC). Four loci colocalised between PSC and UC and two of these four were also shared with CD. Four loci also colocalised with loci for T1DM. There were several risk loci which could not be resolved to a single causal variant or small set of credible variants from Chapter 2's fine-mapping efforts. For these

Table 3.2: Colocalisation between PSC risk loci and immune-mediated diseases

Chr	Region	OR	p-value	UC H4	CD H4	PBC H4	T1DM H4	CeD H4	RhA H4	MS H4	SLE H4
1	*MMEL1*	1.20	5.12E-13	0.08	0.00	0.56	0.01	0.36	0.45	0.95	0.02
2	*BCL2L11*	1.29	2.18E-11	0.89	0.05	0.73	0.00		0.23	0.00	0.08
2	*CD28*	1.25	4.12E-16	0.06	0.01	0.00	0.00		0.00	0.00	0.07
3	*MST1*	1.33	5.25E-26	0.85	0.74	0.01	0.00	0.08	0.00	0.00	0.00
3	*FOXP1*	1.44	2.80E-15	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.00
4	*IL2-IL21*	1.28	8.25E-14	0.02	0.06	0.56	0.00		0.04	0.00	0.01
6	*BACH2*	1.21	1.09E-09	0.01	0.07	0.04	0.18	0.49	0.87	0.00	0.02
10	*IL2RA*	1.22	1.44E-16	0.05	0.00	0.00	0.00		0.95		0.00
11	*CCDC88B*	1.20	1.81E-13	0.00	0.56	0.03	0.82	0.00	0.29	0.00	0.04
12	*SH2B3*	1.18	3.86E-13	0.89	0.84	0.94	1.00	1.00	0.20	0.00	0.73
16	*CLEC16A*	1.20	5.22E-13	0.00	0.00	0.57	0.61		0.00	0.00	0.05
18	*CD226*	1.19	5.87E-12	0.00	0.03	0.88	0.76		0.02	0.00	0.25
19	*PRKD2*	1.28	2.12E-12	0.03	0.00	0.00	0.96	0.01	0.00		0.00
21	*ETS2*	1.23	3.40E-13	0.82	0.79	0.00	0.00	0.00	0.00		0.00
21	*UBASH3A*	1.22	2.42E-12	0.05	0.00	0.00	0.82	1.00	0.42	0.00	0.00

OR; odds ratio for lead GWAS SNP risk allele, p-value; for lead GWAS SNP
PP H4>0.8 highlighted in green, evidence for PP H3>0.8 highlighted in red

loci, in Chapter 2 I had examined fine-mapping studies of the same loci in other IMDs to define the most likely causal variants. Reassuringly, for all loci where this was the case, colocalisation supported a shared causal variant between PSC and the other IMD locus. For example, fine-mapping of PSC region 8 (Chr10:6108139) resolved this locus to a five variant credible set with a 46% PP of causality supporting rs4147359 as the most probable causal variant. Westra *et al* have previously fine-mapped this locus in RhA to a three variant credible set in which rs706778 has 89% PP of causality [114]. Here, I show that this same locus colocalised in PSC and RhA with 95% PP4 supporting a common causal variant between the two IMDs. Thus rs706778 is the most probable causal variant for both PSC and RhA.

When trying to identify the gene perturbed by a PSC risk locus, colocalisation with an eQTL is the most useful functional trait, as it enables us to identify not only the gene affected, but whether PSC risk is conferred by increased or decreased expression of that gene. The gene quantitatively affected by an eQTL, or the eQTL-gene pair is called an eGene. I conducted colocalisation analysis between the twenty-two PSC risk loci and 42 functional QTL datasets covering five gastrointestinal whole-tissue types, six immune cell-types and five different functional traits including gene-expression (eQTL), histone marks (histQTLs), DNA methylation (methQTLs) and splice site QTL (spliceQTLs) (Table 3.1). For those seven risk loci not reaching genome-wide significance in Ji *et al's* data, the results consistently supported no evidence of colocalisations with any functional traits, thus the results for these seven loci are not shown. I found colocalisations with eQTL for four of the remaining fifteen PSC risk loci. Of these four loci, three colocalised

with one eGene and one colocalised with two eGenes. Where a disease risk locus colocalises with an eQTL, further colocalisation of the same region with another functional QTL such as a histQTL or methQTL helps us to idenitfy the epigenetic mechanism via which that eQTL affects gene expression. For example an eQTL may decrease expression of gene X by impeding transcription factor binding, evidenced by colocalisation of the same locus with an eQTL of gene X and a H3K27ac mark (histQTL). Of the four loci that colocalised with one or more eGenes, I found evidence that all four also colocalised with another functional QTL; two with methQTLs, one with a histQTL and one with a spliceQTL.

Where colocalisation for a risk locus identifies the same single eGene in more than one cell-type or tissue, particularly those mechanistically related to PSC, this lends further weight to a causal role for this gene in disease pathogenesis. This was the case for three of the fifteen PSC risk loci; Chr19:47205707 *PRKD2*, Chr21:40466744 *ETS2* and Chr21:43855067 *UBASH3A*. For each of these three loci, I found colocalisations with one eGene across several cell-types and tissues. Followed by functional trait fine-mapping, for these three loci this allowed me to identify a perturbed gene, a direction of effect, a set of relevant cell-types, a single or small set of credible causal variants and the mechanism via which the causal variant potentially dysregulated the quantitative expression of that gene. The colocalisation and functional trait fine-mapping results for these three loci are discussed in more detail below. This is followed by the discussion of two other loci of interest; Chr12:11184608 *SH2B3* and Chr18:67543688 *CD226*.

Table 3.3: Colocalisation of PSC risk loci with functional QTLs

GWAS			Colocalisation				PSC GWAS			Functional Trait		
Chr	**Lead SNP**	**eGene**	**QTL Type**	**Tissue type**	**Tissue State**	**H4 PP**	**Risk allele**	**Beta**	**p-value**	**Risk allele**	**Beta**	**p-value**
3	rs3197999	n/a	methQTL cg06313718	Monocytes	Unst.	0.81	A	0.26	2.60E-13	A	-0.61	2.43E-08
		n/a	methQTL cg06313718	Neutrophils	Unst.	0.85				A	-1.20	2.37E-37
11	rs663743	*AP003774.1*	eQTL	Whole blood	Unst.	0.97	G	0.17	8.42E-08	G	0.82	1.45E-35
		AP003774.1	eQTL	EBV-transformed lymphocytes	Unst.	0.96				G	0.96	3.73E-18
		CCDC88B	eQTL	Monocytes	Unst.	0.85				G	0.57	7.73E-08
		n/a	methQTL cg07884764	Neutophils	Unst.	0.96				G	-0.71	4.29E-12
16	rs725613	n/a	methQTL cg04616529	CD4+ T cells	Unst.	0.96	T	0.20	5.50E-10	T	0.93	3.49E-18
		n/a	methQTL cg00121339	Neutrophils	Unst.	0.92				T	-0.85	4.58E-19
19	rs60652743	*PRKD2*	eQTL	Transverse Colon	Unst.	0.94	A	0.26	1.01E-07	A	-0.21	1.16E-06
		PRKD2	eQTL	Sigmoid Colon	Unst.	0.94				A	-0.39	8.14E-08
		PRKD2	eQTL	Monocytes	Unst.	0.94				A	-1.14	2.30E-25
		n/a	methQTL cg00838415	Monocytes	Unst.	0.94				A	-0.77	6.26E-11
		n/a	methQTL cg00838415	Neutrophils	Unst.	0.94				A	-0.74	4.15E-10
		n/a	methQTL cg08634012	Monocytes	Unst.	0.94				A	-0.97	2.17E-17
		n/a	methQTL cg08634012	Neutrophils	Unst.	0.95				A	-0.89	1.53E-14
21	rs2836883	*ETS2*	eQTL	Whole blood	Unst.	0.83	G	0.30	5.40E-14	G	0.15	2.50E-08
		ETS2	eQTL	Monocytes	Unst.	0.87				G	0.69	4.38E-11
		ETS2	eQTL	Macrophages	IL-4 at 6hrs	0.84				G	0.91	1.91E-09
		n/a	H3K7acQTL	Monocytes	Unst.	0.90				G	1.04	4.77E-23
		n/a	H3K7acQTL	Neutrophils	Unst.	0.92				G	0.62	2.16E-08
21	rs1893592	*UBASH3A*	eQTL	Whole blood	Unst.	0.99	A	0.20	1.90E-07	A	-0.21	8.07E-16
		UBASH3A	eQTL	Transverse Colon	Unst.	0.95				A	-0.20	9.44E-07
		UBASH3A	eQTL	CD4+ T cells	Unst.	0.99				A	-1.25	9.71E-39
		UBASH3A	eQTL	T regs	Unst.	1.00				A	-0.25	1.37E-13
		n/a	spliceQTL	CD4+ T cells	Unst.	0.99				A	1.29	6.19E-43

3.4.1 The *PRKD2* locus

The Chr19:47205707 risk locus colocalised with an eQTL for *PRKD2* in monocytes with 94% PP of causality. Notably, the PSC risk increasing allele was associated with decreased expression of *PRKD2*. This locus also colocalised with two CpG methylation sites (cg00838415 and cg08634012) in both monocytes and neutrophils, suggesting that the causal variant for this locus may exert its repressive effect upon gene expression via hypermethylation. Interestingly, although this region colocalised with an eQTL decreasing expression of *PRKD2* in transverse and sigmoid colonic tissue (PP4=94%) and the 1Mb region surrounding this PSC risk locus also contains a significant IBD risk locus, the evidence supported a different causal variant driving the IBD signal (PP3 for colocalisation with UC and CD of 94% and 97% respectively). However, co-localisation with other IMDs demonstrated that the causal variant for this region was shared between PSC and T1DM. Furthermore, in T1DM this locus has been reported as an eQTL for *PRKD2* in monocytes [182], a finding I was able to replicate by conducting colocalisation between T1DM and the Blueprint monocyte eQTL data (PP4=96%). Thus, these results support that PSC risk, T1DM risk and expression of *PRKD2* in monocytes are all likely driven by the same causal variant. The most probable causal variant was identified by fine-mapping this locus in the PSC GWAS data which resolved the region to a fourteen variant credible set with the majority of the PP attached to rs313839 (PP=23%), followed by rs112445263 (PP=20%) (see Chapter 2). This finding was replicated by fine-mapping the same region in the monocyte *PRKD2* eQTL data, resulting in an eight variant credible set led by rs313839 (PP=14%), and rs112445263 (PP=14%) two variants in high LD (r^2=0.98) with one another (Table 3.4 and Figure 3.3). The remaining PP was split evenly across a further 6 variants in high LD, all with r^2>0.8). Fine-mapping supported a second independent signal in the *PRKD2* eQTL data with 55% PP of causality. This was supported by the finding that the most probable causal configuration contained two uncorrelated SNPs; rs313839 and rs314675.

Confirming the fine-mapping assumption that all potential causal variants have been included in the analysis, a search of the 1000 Genomes and UK10K reference panels found there were no SNPs in high LD (r^2>0.8) with rs313839, missing from the eQTL data. The most probable credible variant for this locus, rs313839, lies within an intron. Colocalisation with two methQTLs suggests that this variant alters two CpG methylation sites in monocytes and neutrophils. Furthermore, rs313839*C>G (where rs313839*G is the PSC risk increasing allele) has also been associated with the disruption of many transcription factor binding motifs though ChIPseq studies, as previously discussed in Chapter 2. This suggests several plausible mechanisms via which rs313839*G may exert its repressive effect upon *PRKD2* expression and subsequent effect upon PSC risk. However, the location of other variants in the credible set within gene regulatory elements may

also be important in driving the observed molecular QTL trait. For example, rs402072 at Chr19:47219122, is the only variant that is in high LD with rs313839 and also lies within several hundred base pairs of the transcription start site (TSS) within the promoter region.

PRKD2 (Protein kinase D2) is a member of the serine/threonine protein kinase family and is known to be highly expressed in PSC-relevant tissues including whole blood, small intestine, colon and liver [176]. PRKD2 has known roles in monocyte migration and adhesion. In THP-1 cells (a widely used experimental model of monocytes) expression of a dominant-negative form of PRKD2 resulted in decreased monocyte migration in response to stimulus [179]. Knockdown of *PRKD2* was shown to reduce adhesion of THP-1 cells to endothelial cells in culture, whereas activation of PRKD2 through phosphorylation at Ser 744/748 was shown to increase adhesion to endothelial cells [180]. Monocytes and their macrophage progenitors play an important role in immune-regulation and tissue-repair. Therefore genetic variants that result in decreased expression of *PRKD2* may impair monocyte migration into tissues and subsequent tissue regeneration. *PRKD2* is however not only active in monocytes. The importance of *PRKD2* in T-cells has been demonstrated *in vivo* through T-cell-mediated immune responses in mice expressing PRKD2 variants that cannot be phosphorylated by protein kinase C [205]. While PRKD2 catalytic activity is not essential for the development of mature peripheral T- and B-lymphocytes [206], PRKD2-mutant mice show a striking reduction in the ability of the T-cell receptor (TCR) to induce production of pro-inflammatory cytokines such as interleukin 2 (IL-2) and interferon-γ (IFN-γ), which are important for optimal T-cell-dependent antibody responses [205]. In response to TCR stimulation in Jurkat cells (a model of peripheral T-cells), PRKD2 was activated and translocated from the cytoplasm to the nucleus, to allow IL-2 and IFN-γ promoter up-regulation [207]. Furthermore, in T-cell specific PRKD2-deficient mice, the generation of CD4+ thymocytes is abrogated. This defect is likely to be caused by attenuated TCR signalling during positive selection and incomplete CD4+ lineage specification. The role of PRKD2 in activated T-cells/thymocytes may explain the absence of an observed effect in the naïve CD4+ T-cells studied in my colocalisation analysis. This suggests that the generation of eQTL maps in other T-cell subsets, in different states of activation, may be useful in the further investigation of *PRKD2* in immune-mediated disease risk.

Table 3.4: Fine-mapping of PSC risk loci in functional QTL data

	GWAS Finemapping			Colocalisation			MolQTL Finemapping		
Chr	SNP	PP	CS	QTL type	Cell type	Gene	SNP	PP	CS
11	rs663743*	0.41	2	eQTL	Monocyte	*CCDC88B*	rs663743	0.03	245
19	rs313839	0.23	14	eQTL	Monocyte	*PRKD2*	rs112445263	0.14	8
21	rs4817988	0.58	10	eQTL	Monocyte	*ETS2*	rs4817987	0.07	47
				H3K27ac	Monocyte	N/A	rs2836878	0.13	11
21	rs1893592	0.61	5	eQTL	CD4+ T-cell	*UBASH3A*	rs1893592	1.00	1
				SpliceQTL	CD4+ T-cell	*UBASH3A*	rs1893592	1.00	1

*2nd signal in region, CS; Credible set size, PP; Maximum posterior probability

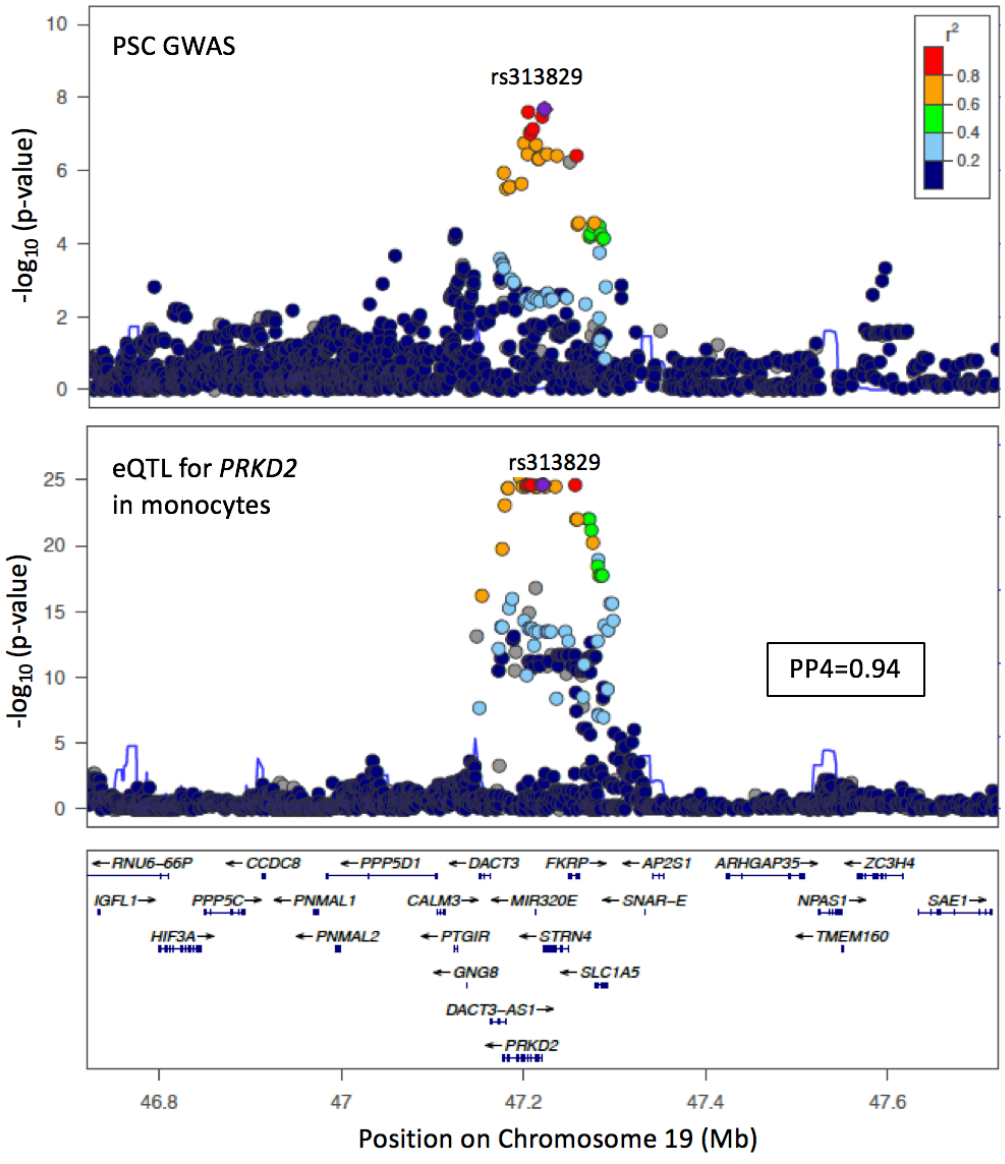

Figure 3.3: Chr19:47205707 regional association plots for most probable fine-mapped SNP, rs313839, in PSC GWAS data and colocalising eQTL data for *PRKD2* in monocytes.

3.4.2 The *ETS2* locus

The Chr21:40466744 locus colocalised with a eQTL for *ETS2* in three different tissues; whole blood, monocytes and IL-4 stimulated macrophages. GWAS studies in both IBD and PSC have consistently proposed that the most likely candidate gene for the Chr21:40466744 locus is *PSMG1* [42, 73]. This was based upon the paucity of genes in this region, and a study of colonic biopsies from paediatric-onset IBD patients, which demonstrated a 'modest' increase in the colonic expression of *PSMG1* in IBD cases compared to controls [175]. Indeed, *PSMG1* which encodes proteasome assembly chaperone 1, has a biologically plausible role in IBD, as part of the ubiquitin-protesosome system. The ubiquitin-protesosome system regulates the generation of peptide antigen presented to MHC class I [208] and TCRs, in addition to regulating co-stimulatory signaling [209]. However, the results of my analysis instead support *ETS2* as the gene dysregulated by this locus. In each tissue, the PSC risk increasing allele was associated with increased expression of *ETS2*. This locus also co-localised with a histQTL for H3K7ac, a marker associated with higher activation of transcription, in both unstimulated monocytes and neutrophils. This suggests that the mechanism by which the causal variant increases expression of *ETS2* in monocytes may, for example, be via increasing the affinity of transcription factor binding. Where a risk locus does not colocalise with an eGene in a particular cell-type, colocalisation with a functional QTL may suggest the presence of an eQTL in a different activation state [134]. It is therefore possible that this locus may be an eQTL of *ETS2*, if investigated in stimulated or activated neutrophils.

Colocalisation of this locus with an eQTL for *ETS2* in iPSC-derived macrophages, six hours following stimulation with IL-4, is particularly interesting as this is a stimulus that mimics the allergic response. Of note, there was no evidence for colocalisation with a macrophage eQTL in either the resting state or the multiple other stimulation states outlined in Table 3.1. This is particularly notable because the vast majority of eQTLs in these data are shared widely across stimulation states. Not only does this highlight the importance of studying cells in the correct state of activation on our ability to identify eQTLs, but also supports a role for *ETS2* in the autoimmune response. The *ETS2* locus also colocalised with a GWAS locus for neutrophil counts (PP4=83%), where the PSC risk increasing allele was associated with a reduction in neutrophil counts. This is biologically plausible given the role of *ETS2* in inducing expression of pro-inflammatory cytokines in macrophages, and the close interactions between macrophages and neutrophils in the inflammatory response.

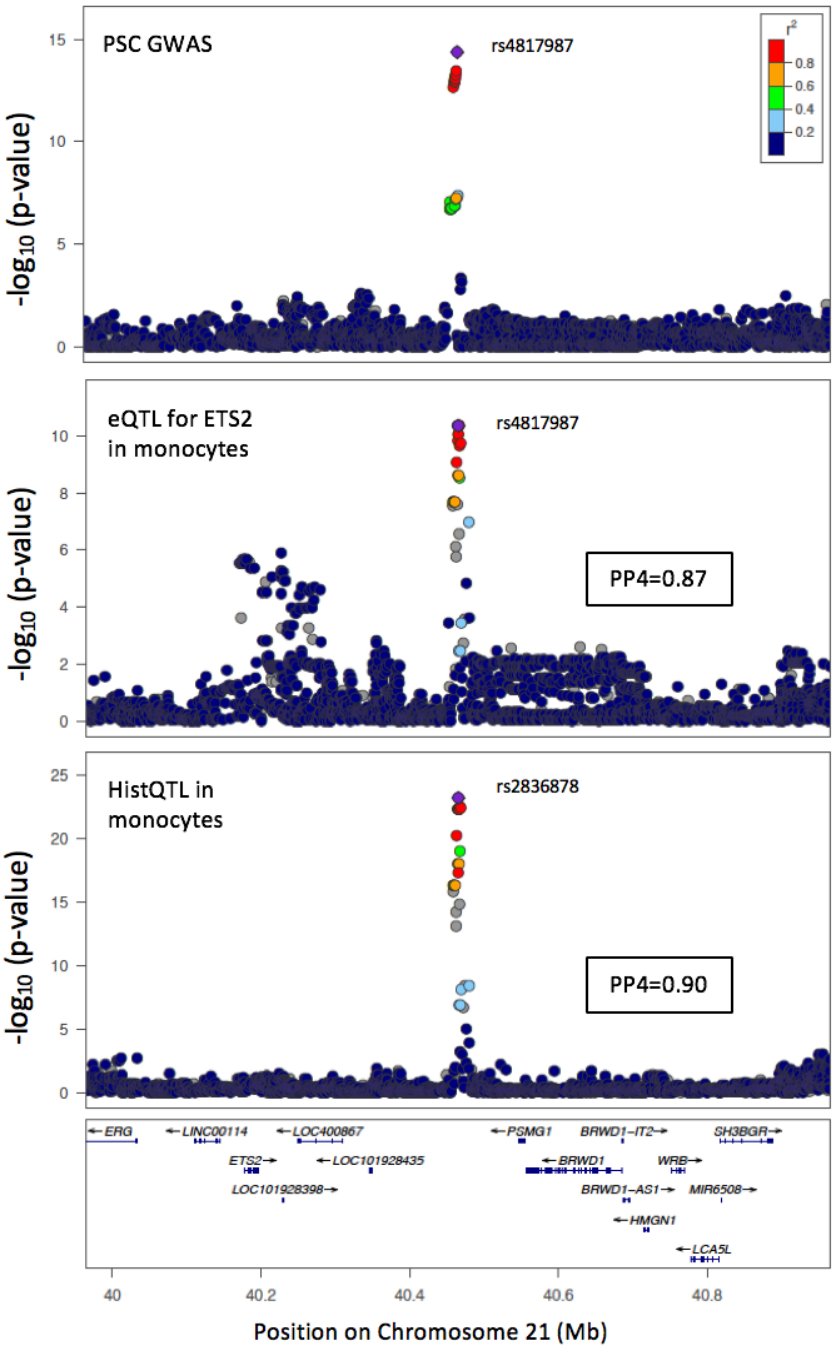

Figure 3.4: Chr21:40466744 regional association plot showing the most probable fine-mapped SNP for PSC GWAS (rs4817987) and colocalising eQTL data for *ETS2* in monocytes (fine-mapped to rs4817987) and for a H3K27ac histQTL in monocytes (fine-mapped to rs2836878).

The Chr21:40466744 locus colocalised with UC and CD (PP4 of 84% and 80% respectively), with no evidence supporting colocalisation with any other IMD (Table 3.2).

However, *ETS2* is a ubiquitously expressed transcription factor, with a well-defined role in macrophages. Cytokine-dependent phosphorylation of Ets-2 results in Ets-2 directly binding the promoters of matrix metaloproteinases *MMP-1* and *MMP-9*, and the induction of other pro-inflammatory target genes including *TNFa, IL-1β*, and chemokines, *CCL2/MCP-1* and *CCL3/MIP-1α* [210]. In mice with severe macrophage-induced pneumonitits, the prevention of Ets-2 phosphorylation on Thr(72) by the Ets-2(A72) mutant allele results in decreased tissue macrophage infiltration [211]. Thus, activated Ets-2 has an important role in the persistent inflammatory response. It is therefore biological plausible that increased expression of *ETS2* could contribute to driving the aberrant inflammatory response observed in PSC. Although I did not observe any colocalisation of this locus with functional QTLs in T-regulatory or CD4+ T-cells, *ETS2* also has a role in IL-2 regulation, the first cytokine produced when naïve T-helper (Th) cells are activated and differentiate into dividing pre-Th0 proliferating precursors. A study by Panagoulias *et al* has demonstrated that Ets-2 binds to the IL-2 promoter which allows transition of naïve Th cells to Th0 cells upon stimulation with antigen, and that Ets-2 silencing allows for constitutive IL-2 expression in unstimulated T-cells [212]. Indeed, they hypothesise that disturbance of this pathway could cause deranged Th cell plasticity and resultant autoimmune disease. Further analysis of eQTL maps in different T-cell subsets and activation states would be necessary to evaluate any effect of the *ETS2* risk locus on *ETS2* expression in T-cells.

Fine-mapping of the Chr21:40466744 *ETS2* locus within the functional QTL data did not prove useful in resolving the causal variant(s) driving this locus. Fine-mapping in the monocyte eQTL data resulted in a credible set of forty-seven variants compared to ten variants in the GWAS data fine-mapping (presented in Chapter 2). This larger credible set is partially attributable to the higher numbers of variants directly genotyped in the whole-genome sequenced eQTL data. Resultantly, the PP was more evenly split between a larger number of very highly correlated variants (Figure 3.4). Furthermore, the failure of eQTL fine-mapping to improve upon the GWAS fine-mapping is a consequence of the reduced strength of association between the lead variant in the eQTL signal compared to the GWAS signal, reducing the power to fine-map. Fine-mapping in the histQTL data resulted in a similar sized credible set of eleven variants compared to ten in the PSC GWAS data. Whilst all ten variants in the GWAS credible set overlapped with the histQTL credible set, the evidence supported rs2836878 as the most probable causal variant in the histQTL signal (PP=13%), compared to rs4817988 in the GWAS data (PP=58%).

3.4.3 The *UBASH3A* locus

Of all PSC risk loci, the Chr21:43855067 locus was the most extensively investigated prior to this study. This locus was already a known eQTL of *UBASH3A* from two whole-blood

and one B-cell only analysis [172, 213, 214] and a likely shared risk locus with CeD and RhA [148, 201]. I confirmed, with colocalisation, that this PSC locus shared a causal variant with CeD (PP4=100%), as well as T1DM (PP4=82%). However colocalisation resulted in almost equivocal evidence supporting a shared (PP4=42%) or different causal variant (PP3=54%) driving the signal in RhA.

Colocalisation confirmed that this locus is an eQTL of *UBASH3A* in T-regulatory cells (PP4=100%) and naïve CD4+ T cells (PP4=99%). In both T-cell types, the PSC risk increasing rs1893592*A allele, which is also the major allele at this locus, was associated with decreased expression of *UBASH3A*. Interestingly, although there was no evidence supporting shared genetic variation with UC or CD, the Chr21:43855067 rs1893592 locus also colocalised with a eQTL of *UBASH3A* in transverse colon tissue (PP4=95%), but not sigmoid colon, a pattern of colonic involvement reminiscent of the PSC-associated IBD phenotype. Fine-mapping of this locus in PSC GWAS and CD4+ T-cell eQTL data supported rs1893592 as the most probable causal variant. As a result of the higher strengths of association in the functional QTL data increasing power to fine-map the signal, fine-mapping in the eQTL data attributed 99% of the PP4 to rs1893592 compared to 61% in the GWAS data. The rs1893592 variant is thought to alter the conserved 5' splice donor sequence. The predicted consequence of the PSC protective rs1893592*C allele is to increase expression of the downstream intron, causing intron 10 to be retained in the *UBASH3A* mRNA [42, 215]. This was supported by the finding of a colocalisation with a spliceQTL in CD4+ T-cells (PP4=99%), which was also fine-mapped to the same causal variant, rs1893592, with 100% PP4 of causality.

UBASH3A has a described role in human T-cells where it has been shown to attenuate the NF-$\varkappa$B signalling pathway upon TCR stimulation, by specifically suppressing activation of the I$\varkappa$B kinase complex, through a ubiquitin-dependent mechanism [216]. In the T-cell eQTL data used for colocalisation in this analysis, the PSC protective rs1893592*C allele was associated with increased *UBASH3A* expression. It has been previously demonstrated in human primary CD4+ T cells that following TCR stimulation, the PSC-protective rs1893592*C allele is associated with a significant reduction in the overall mRNA levels of UBASH3A, but an increase in the proportion of a normally occurring, but low-abundant UBASH3A transcript that retains intron-9 sequences and cannot produce full-length UBASH3A protein [217]. The reduction in UBASH3A mRNA subsequently results in increased secretion of IL-2, a key cytokine in T-cell function and activation. This therefore provides important insights into how dysregulation of *UBASH3A* splicing and expression may be causal in the pathogenesis of PSC.

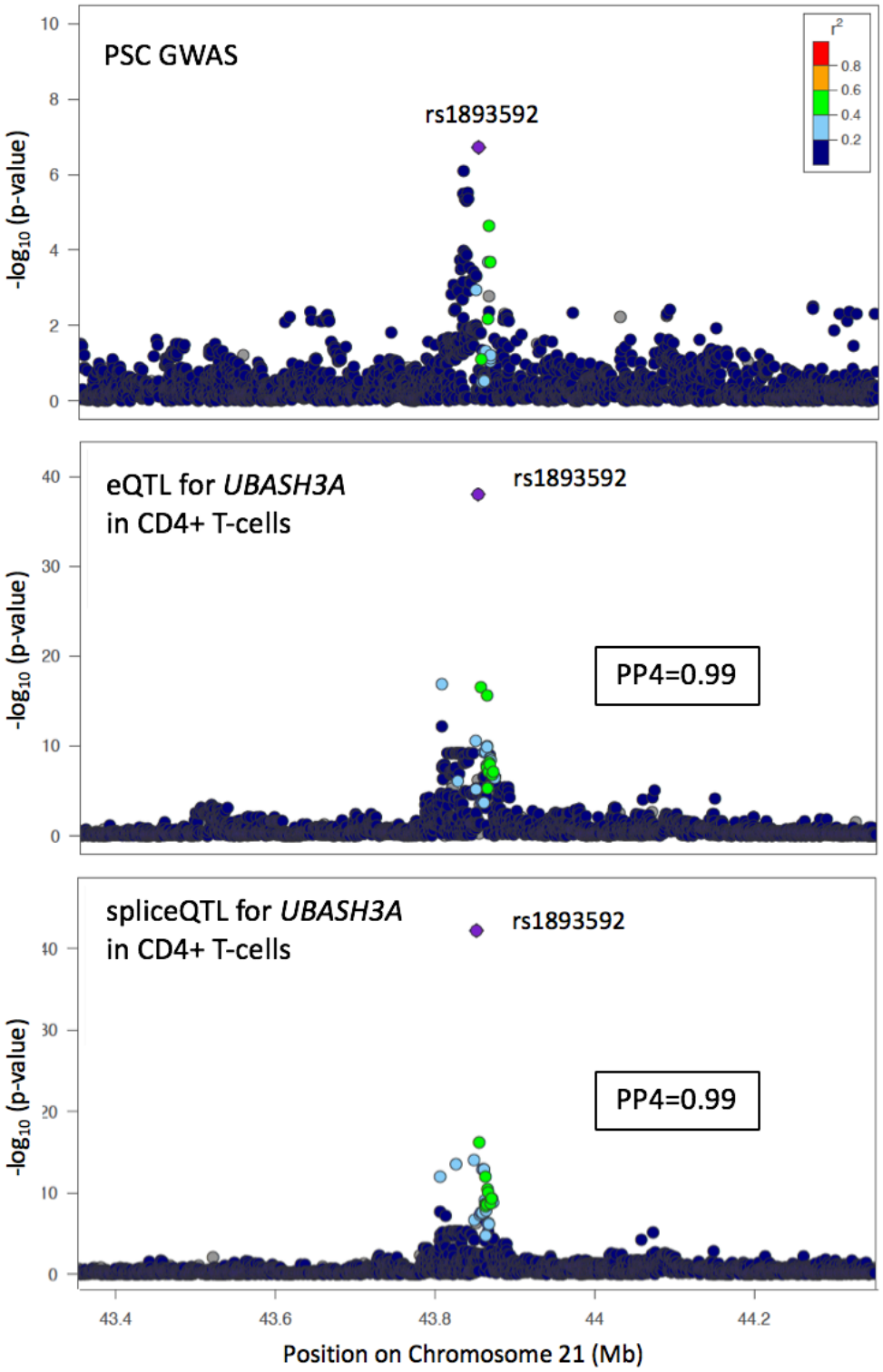

Figure 3.5: Chr21:43855067 regional association plots for fine-mapped SNP, rs1893593, in PSC GWAS and colocalising eQTL data for *UBASH3A* and spliceQTL data for *UBASH3A*.

3.4.4 The *SH2B3* locus

The Chr12:11184608 *SH2B3* PSC risk locus colocalised with five other IMDs; UC, CD, PBC, T1DM and CeD, supporting a single-shared causal variant driving all six diseases. Fine-mapping in the PSC GWAS data (Chapter 2), highlighted rs3184504, a missense variant within the 3rd exon of *SH2B3*, as the most probable causal variant (PP=99%). Whilst this locus has not been fine-mapped in UC, CD, PBC or CeD, a fine-mapping study of T1DM resolved this locus to a credible set including two variants; rs653178 (PP4=66%) and rs3184504 (PP4=33%) [114]. Notably, both of these variants were included within the PSC fine-mapping analysis.

SH2B3 is ubiquitously expressed across many cell and tissue types, with a role in the regulation of signalling pathways involved in cell migration, differentiation, inflammation and haematopoiesis [218]. It was therefore unsurprising that this locus colocalised with GWAS traits for leucocyte, monocyte and neutrophil counts. The PSC risk increasing rs3184504*T allele is associated with an increase in all myeloid and lymphoid cell counts, compared to the reference rs3184504*C allele. *SH2B3* is a negative regulator of T-cell activation, TNF production, and Janus kinase (JAK)-2 and -3 signalling. Another eQTL study has shown that the autoimmune hepatitis (AIH) rs3184504*A risk allele, which results in the same protein coding change as rs3184504*T, is associated with increased expression of genes involved in IFNγ production [172]. This suggests a mechanism via which this locus might contribute to increased immune cell counts and aberrant immune- and inflammatory-response.

3.4.5 The Chr18:67543688 locus

PBC is an immune-mediated inflammatory condition affecting the small bile ducts that is often considered a sister condition to PSC. Somewhat surprisingly, colocalisation of PSC risk loci with PBC identified only two loci for which there was evidence of a single shared causal variant between both traits. The first was the Chr12:11184608 *SH2B3* locus discussed above, and the second was the Chr18:67543688 locus. The Chr18:67543688 locus could not be well fine-mapped in the PSC GWAS data with a credible set containing 44 variants. Furthermore, this locus did not colocalise with any other IMD, other than PBC, in this analysis. It is therefore possible that the genes and pathways affected by this locus are perhaps the most likely candidates for bile-duct specific effects. This 1Mb region of the genome contains only four genes (see Figure 2.7, Chapter 2); *DOK6, CD226, RTTN* and *SOCS6*, however I did not find subsequent evidence for colocalisation with any functional QTLs or eQTLs to support a causal role for one of these four genes. Two of these four candidate genes, *CD226* and *SOCS6*, have important roles in relevant immune cell pathways. The first gene, *CD226*, is expressed on the surface of natural killer

cells, T-cell subsets, platelets and monocytes. CD226 mediates cellular adhesion to other cells expressing its ligands, CD112 and CD155. The second gene, *SOCS6* (suppressor of cytokine signalling 6), is a cytokine-inducible negative regulator of cytokine signaling. The third gene in this region, *DOK6*, is a less likely candidate for involvement in PSC or PBC pathogenesis as it is expressed mainly in the central nervous system where it is involved in the receptor tyrosine kinase signalling cascade [219]. Whilst little is known about the final candidate gene in this locus, *RTTN*, it encodes Rotatin, an intracellular protein thought to play a role in the maintenance of normal ciliary structure [220]. Cholangiocytes are ciliated cells which have a role in expediting bile flow, and disturbance of the normal structure or function of cholangiocyte cilia is likely to contribute to several cholangiopathies [221]. Thus *RTTN* is the only gene within the Chr18:67543688 PSC-PBC risk locus with a potential role in bile duct homeostasis, highlighting it as a potential candidate gene for further investigation.

3.5 Discussion

In this chapter I describe the first investigation of PSC risk loci using colocalisation with multiple traits including IMD risk, cell count indices, eQTLs and functional QTLs across a variety of PSC-relevant cell-types and tissues. By combining colocalisation to identify the genes impacted by PSC risk loci and the epigenetic mechanisms underlying the gene perturbation, with fine-mapping in the colocalising functional traits, I identify the genes, cell-types and causal variants affected by several PSC risk loci. For four of the fifteen PSC risk loci, this was successful in identifying the genes perturbed and for three of these five loci, it was successful in identifying a single causal variant, or small set of credible variants. Perhaps most notably, these analyses determine that the most probable causal variant driving the Chr19:47205707 PSC and T1DM risk locus, rs313839, results in a reduction of *PRKD2* expression in monocytes and colonic tissue, possibly mediated by hyper-methylation. Similarly, I have fine-mapped the shared PSC-IBD Chr21:40466744 risk locus to a set of ten credible variants, of which the true causal variant increases expression of *ETS2* by activating transcription in monocytes and macrophages subject to allergic stimulus. Thus the results of this study can guide further functional follow-up of these loci in terms of causal variants, direction of effect upon gene expression and relevant cell types in which the effects are mediated. Furthermore, they advocate the combination of colocalisation and functional trait fine-mapping as an alternative approach to resolving the causal variants driving complex trait loci in rare diseases, in which amassing the large sample sizes required to improve upon GWAS trait fine-mapping is unlikely to be feasible. However, for some non-coding PSC risk loci, this pipeline was not effective in determining either causal variants or genes. For example the Chr2:111933001 (*BCL211*),

Chr6:91030441 (*BACH2*), Chr10:6108439 (*IL2RA*) and Chr18:67543688 (*CD226*) loci did not colocalise with any functional QTL in the tissues and cell types included within this analysis. Each of these loci did, importantly, colocalise with at least one other IMD (Table 3.2), all of which are more common diseases with larger GWAS sample sizes, meaning that colocalisation and fine-mapping in those other colocalising IMD traits may be an alternative route to resolving these PSC risk loci.

This study focused on colocalisation with functional traits in cells and tissue types relevant to PSC. This was based upon previous studies demonstrating that some eQTLs are only active in particular cell types or activation states [130], and that eQTLs are enriched for disease-associated variants in disease-relevant tissue-types [190, 191]. The choice of disease-relevant tissues in this study was however limited by two factors. Firstly, designating a cell-type 'relevant' in a disease such as PSC, in which we have limited understanding of disease pathogenesis, is challenging. Secondly, the limited availability of functional QTL data with publicly accessible full summary statistics in these 'relevant' cell types further impairs this choice. However the results from this analysis serve to highlight the importance of conducting colocalisation with eQTLs measured in the relevant cell-types. For example, analysis of the Chr21:40466744 locus supported *ETS2* as the most likely gene perturbed by this risk locus, with colocalisations observed in monocytes and IL-4 stimulated macrophages. Whilst *ETS2* has a described role in the induction of pro-inflammatory cytokine release from macrophages, *ETS2* also has a role in IL-2 regulation in naïve Th transitioning to Th0 cells upon antigenic stimulation. Given this role in naïve Th cells, it is unsurprising that we did not find any colocalisation with eQTLs for *ETS2* in the available CD4+ or T-regulatory cell datasets. However, it is plausible that if examined in the right T-cell subtype or activation state, the Chr21:40466744 locus may also be an eQTL of *ETS2* in some T-cell subtypes. Similarly, investigation of the Chr19:47205707 risk locus found it colocalised with an eQTL for *PRKD2* in monocytes, a gene with a role in the adhesion of monocytes to endothelial cells. Whilst this gene also has a role in negative selection of T-cells, I did not find any colocalisation with a *PRKD2* eQTL in the available CD4+ T-cell and T-regulatory cell data. Furthermore, there are no published and publicly available eQTL datasets for T-cells in the activated or stimulated state, again introducing the possibility that the correct cell type has not been examined. Future work could focus upon conducting combined colocalisation and fine-mapping in functional QTL data from all available cell-types and tissues, with the added risk of introducing noise by examining traits across multiple tissue types and the difficulty of interpreting colocalisations with genes in tissues such as brain or muscle which are seemingly remote from PSC pathogenesis. Another solution would be to use the current hypotheses of disease pathogenesis in PSC to select those cell types of most potential mechanistic relevance to PSC and to build eQTL maps in those PSC-specific

cell types. This is an analysis presented in the following chapter.

Several properties of *Coloc* are likely to have influenced the results presented in this chapter. Firstly, Bayesian colocalisation analysis is strongly influenced by the choice of priors. Indeed, as the p^{12} threshold is increased (e.g. from 10^{-6} to 10^{-5}), there is more certainty that the data supports a shared causal variant between both traits. This can be especially important in regions where there are extended patterns of strong LD and thus uncertainty as to whether the data supports the H3 or H4 hypothesis, because it is in keeping with both scenarios. For these loci, the choice of prior becomes the determinant of the colocalisation. An example of this is the Chr4:123499745 locus near the candidate gene *IL2-IL21*, for which there was no evidence supporting colocalisation with any functional QTLs or IMDs at $p^{12}=10^{-6}$, but with evidence supporting shared genetic variation with several other IMDs driven by a different causal variant (PP3>80%) (Table 3.2). However, the evidence supporting colocalisation (PP4) increases as the p^{12} threshold is increased (Figure 3.2). Whilst this may favour a higher p^{12} for the detection of more colocalising IMD traits, it is known that variants associated with complex traits are more likely to be eQTLs than MAF-matched variants from GWAS analyses chosen at random, thus supporting the more stringent choice of priors used in this analysis [117, 222]. Secondly, *Coloc* makes the assumption that each risk locus contains only one independent signal. For those regions in which there were more than one independent signal, *Coloc* considers only the strongest of these distinct association signals. Where each of the association signals explains a similar proportion of the variance of the trait, the PP4 will drop and PP3 proportionately increase [189]. Fine-mapping of the PSC risk loci described in Chapter 2 supported the presence of two independent signals in four of the 15 loci. For those four PSC risk loci containing two independent signals, there was evidence for colocalisation with functional QTLs for only one of these four risk loci. This was the Chr11:64107735 locus, which colocalised with an eQTL for *CCDC88B* in monocytes and an eQTL for *AP003774.1* in whole blood and EBV-transformed lymphocytes. A future means of investigating these multi-signal loci is to include a step-wise conditional regression [223] to identify additional independent signals within a locus, and to perform colocalisation on the resultant conditional p-values, as a means to accounting for multiple independent signals [189].

Colocalisation with eQTLs, functional QTLs and other IMDs allows us to ascribe a gene, the direction of effect on gene expression associated with PSC risk, the epigenetic mechanism dysregulating that genes expression as well as the other IMDs impacted via the same gene and epigenetic mechanism. With the example of the Chr19:47205707 risk locus, colocalisation identified that the causal PSC risk increasing allele for this locus correlated with an eQTL reducing expression of *PRKD2*, via hypermethylation, and that the same causal variant also conferred risk of T1DM. In order however, to unequivocally prove that T1DM risk at this locus is also mediated by perturbation in *PRKD2* expression,

I needed to performed further colocalisation between T1DM and monocyte eQTL data. More recently, Giambartolomei *et al* have published methods to quantify the evidence supporting a common causal variant in a particular region across multiple traits from summary statistics [224]. This method, *Moloc*, was published in 2018 after the analysis presented in this chapter was largely complete. Similar to *Coloc*, *Moloc* uses a Bayesian framework to integrate GWAS and functional QTL data, with the same three assumptions pertaining to the inclusion of the true causal variant within the data, a maximum of one independent association per region and that samples are drawn from the same ethnic population and thus share LD structure. The future use of such a method would be advantageous in providing a quantification of evidence for a shared causal variant between all traits tested for one locus, avoiding the need for the multiple rounds of pair-wise colocalisation conducted in this analysis. Such an approach could also be useful in the fine-mapping of PSC risk loci. An important part of this analysis was to conduct fine-mapping of loci within functional traits, in an effort to identify the causal variant driving these colocalising traits. Whilst data availability meant that this approach could only be applied to four of the PSC risk loci, it was successful in improving fine-mapping resolution for two of these loci. An potentially fruitful future analysis might focus upon boosting power for fine-mapping by combining multiple colocalising datasets for a single locus into one meta-dataset using a model that allows for mixed effect sizes, followed by fine-mapping of the meta-dataset. Methods based upon similar approaches have been published by Wallace *et al* [225] and will form part of my future follow-up work, not presented in this thesis.

Using a combination of colocalisation and fine-mapping across multiple traits, I have been able to identify the genes, causal variants and epigenetic mechanisms implicated by five PSC risk loci. In addition, my work highlights some of the cell-types in which these aforementioned genes and mechanisms are especially relevant. However, several loci remain unresolved, and future work should focus upon using current knowledge of PSC pathogenesis to build eQTL maps in the most PSC-relevant cell types, followed by similar colocalisation and fine-mapping analyses. This analysis, presented in the following chapter is a means to further understanding the causal biology of PSC.