Chapter 2

Fine-mapping of disease-associated risk loci in Primary Sclerosing Cholangitis

2.1 Introduction

GWAS have identified many thousands of regions of the genome associated with immune-mediated disease (IMD), which have been replicated both within and between diseases. One biological phenomenon facilitating the success of GWAS is that of linkage disequilibrium (LD), the non-random association of alleles between nearby genetic variants. LD blocks can consist of anywhere between two and thousands of highly correlated single-nucleotide polymorphisms (SNPs). The GWAS design uses LD to its advantage, utilising several hundred thousand 'tagging' SNPs to capture a large proportion of the common variation in the human genome, to accurately identify loci associated with disease. The most strongly associated variant with the smallest p-value within the disease-associated locus is referred to as the 'lead variant'. However, one important shortcoming of the GWAS design is that the lead variant identified by GWAS is likely to be in high LD with many other SNPs, resulting in a statistical association of approximately equivalent strength across all of the variants in high LD (defined as $r^2>0.8$), any one of which could be the true causal variant. Distinguishing the causal variant from the often hundreds of variants in LD is not possible from GWAS alone. Whilst conditional analysis is one means of identifying the number of independent association signals within a region, it cannot ascribe an individual probabilistic measure of causality for each individual variant within a locus. Identifying the true causal variant within each risk locus is an essential step in translating genetic associations into biological functions that explain disease processes. Knowledge of the precise location of the causal variant within, for example, a gene intron, exon, splice junction, promoter or enhancer region, may provide important clues about the mechanism via which the variant

exerts its effect on disease risk. With the development of CRISPR/Cas9 gene editing technology, which allows the introduction of targeted mutations within cellular DNA, knowledge of the causal variant can now greatly facilitate the design of functional assays investigating the mechanistic impact of disease-associated variants. Furthermore, it allows recall-by-genotype experiments, studying individuals with and without a particular causal variant.

Statistical approaches aimed at identifying the most likely causal variant within GWAS risk loci are known as fine-mapping methods. Fine-mapping aims to define a single variant or credible set of variants which contain the true causal variant with a high probability. Several conditions are important when conducting fine-mapping studies. Firstly, whilst GWAS requires only one variant in LD with the true causal variant to detect a signal for disease association, fine-mapping requires that all common SNPs within a region are genotyped or well-imputed [113]. This is because an important fine-mapping assumption is that the true causal variant is included within the data. Imputation using reference panels such as UK10K or the 1000 Genomes Project, allows the incorporation of variants that were not included within the original genotyping array [153, 154]. This aims to satisfy the assumption that when estimating the relative evidence for each variant being causal, the true causal variant is present within the analysis [113]. Secondly, fine-mapping utilises subtle differences in the strength of association between tightly correlated variants to infer causality. It is therefore especially sensitive to data quality and stringent quality control is essential to remove genotyping errors and batch effects. Large sample sizes are also necessary to achieve sufficient power to differentiate between SNPs in high LD. As shown in Figure 2.1, taken from Huang, Fang, Jostins *et al* [56], power to identify the causal variant in a correlated pair increases with the significance of the association, and therefore with sample size and effect size. Initially, the generation of larger sample sizes was achieved by the use of cheaper custom genotyping arrays, such as the Immunochip, at the expense of analyses restricted to only those regions of the genome previously associated with risk of those IMDs included within the Immunochip, of which PSC was not included. However more recently, the reducing cost of sequencing has allowed GWAS on an unprecedented scale, with the meta-analysis of pooled data through collaborative consortia.

The standard approach for refining association signals is via conditional analysis, a step-wise, iterative approach which conditions on the SNP with the lowest p-value for association and continues to add SNPs until no additional SNP reaches the p-value threshold, usually set at 5×10^{-8}. This process provides information about the number of complementary signals within a locus, but cannot assign an probabilistic measure of causality to each individual variant within the locus. Furthermore, p-values are not necessarily comparable between studies as they are heavily influenced by characteristics of the individual study design such as power and locus-specific factors such as minor

allele frequency (MAF) and effect size [113]. Currently, the most common approaches to fine-mapping therefore employ Bayesian methods in which the evidence for the association of each variant is tested using an approximate Bayes factor (ABF). These are then used to calculate the posterior probability (PP) for each variant being causal within a region. Each PP describes the ratio for that variant being causal, versus all other variants in the region, and thus Bayesian PPs are more comparable between variants of the same or different studies. Furthermore, the Bayesian approach enables the weighting of evidence for a particular variant being causal according to prior knowledge, known as the prior probability. Typical fine-mapping approaches in complex diseases aim to define the number of independent signals within a risk locus, and to identify a 'credible set' of variants, in which the sum of the PPs is >0.95, and thus the credible set is >95% likely to contain the true causal variant.

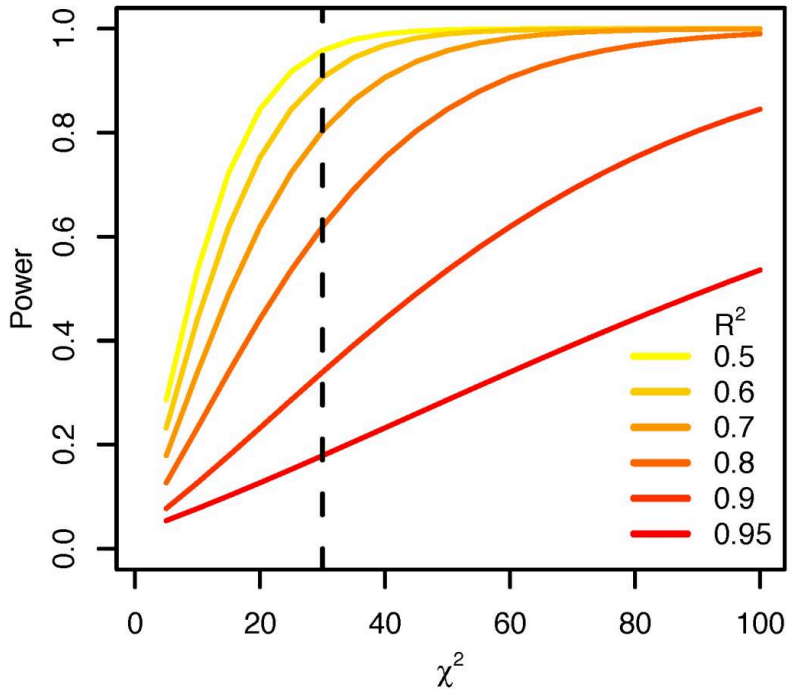

Figure 2.1: Power (y axis) to identify the causal variant in a correlated pair increases with the significance of the association (x axis), and therefore with sample size and effect size (vertical dashed line shows genome-wide significance level). Figure taken from Huang, Fang, Jostins *et al* [56].

Fine-mapping approaches have been applied to several IMDs to date and have been successful in resolving some disease risk loci down to single or small set of credible causal variants. For example Westra *et al* fine-mapped 76 rheumatoid arthritis (RhA) and type 1 diabetes (T1DM) risk loci, defining credible sets of ≤ 5 causal variants at 5 RhA and 10

T1DM loci [114]. IBD remains the most extensively fine-mapped IMD to date. Huang, Fang and Jostins *et al* fine-mapped 94 of the 240 known IBD risk loci and resolved 18 associations down to a single causal variant with >95% certainty, and 27 associations to a single variant with >50% certainty [56]. Importantly, of these 45 variants, 13 were found to be significantly enriched for protein-coding changes, 3 caused direct disruption of transcription-factor binding sites and 10 were tissue-specific epigenetic marks in specific immune cells. In addition de Lange and Moutsianas *et al* resolved an additional 7 IBD loci to a single credible variant with >50% PP of being causal [60]. To date, there have been no fine-mapping studies of the genetic risk loci associated with PSC.

2.2 Chapter overview

Twenty-three regions of the genome have been associated with susceptibility to PSC. The first step in translating these genetic associations into biological understanding of disease mechanisms is to accurately define those variants which are responsible for driving each risk locus. In this chapter I describe the first fine-mapping analysis of genetic risk loci associated with PSC susceptibility. I apply Bayesian fine-mapping approaches to PSC risk loci, with the aim of resolving each locus to a single causal variant or a small set of credible causal variants. To identify those credible variants in non-coding loci which overlap known functional regions of the genome, I perform annotation of the fine-mapped variants to define their functional effects.

2.3 Methods

2.3.1 Fine-mapping

There are multiple computational software programs which employ Bayesian approaches to fine-mapping. In order to develop a fine-mapping analysis pipeline that could be easily applied to data-sets for which full genotype data might not be available, I aimed to use a method of fine-mapping which could be applied to summary statistic data. At the time of conducting this study, several methods for fine-mapping using summary statistics and a SNP correlation matrix, were available. These included *Paintor* [155], *Caviar* [156], *CaviarBF* [157] and *FINEMAP* [158]. Of these four methods, the first three implement an exhaustive search through all possible causal SNP configurations and therefore become computationally slow when considering more than one independent causal variant within each region. I therefore opted to use *FINEMAP* v1.3, a computational software program for the fine-mapping of complex traits [159]. The *FINEMAP* model is made up of four components; the likelihood function, priors, likelihood evaluation and

search algorithm. It differs from the other three methods in that it employs a Bayesian approach to explore a set of the most likely causal configurations of variants within each region via a shotgun stochastic search algorithm [160]. By focusing the analysis on those variants with a non-negligible causal probability rather than searching through all possible causal configurations, *FINEMAP* avoids becoming computationally slow or intractable when considering several causal variants in a data-set with many thousands of variants within each region.

The first step of Bayesian fine-mapping requires a set of summary statistics and risk loci. I used summary statistics from the largest available PSC GWAS dataset, published by Ji *et al* [42]. Ji *et al* identified 15 PSC risk loci outside of the HLA, associated with PSC susceptibility. In addition, a further seven risk loci have been previously reported as associated with PSC from other studies (*CCL20, CPR35, NFKB1, SIK2, HDAC7, RFX4, TCF4*), however these did not reach genome-wide significance in Ji *et al's* study. I therefore excluded these seven loci, focusing fine-mapping efforts on the 15 significant PSC risk loci (shown in Table 2.1). *FINEMAP* assumes that each region to be fine-mapped includes at least one causal SNP, and that all causal SNPs are included within the data (either directly genotyped or imputed). Genotyping of the PSC GWAS data-set had been previously conducted on three different genotyping arrays; the Illumina Omni 2.5-8 and Omni 2.5-4 and the Affymetrix Affy 6 with imputation using a combined reference panel of the 1000 Genomes Phase 1 integrated version 1 and the UK10K cohort [153]. Quality control had been previously conducted by Ji *et al* using strict standards for genetic association analysis [42]. For the purposes of fine-mapping, I defined each of the 15 PSC risk loci as 1Mb regions centred upon the lead GWAS SNP. GWAS summary statistics required for the analysis were the variant RSID, the chromosome and base pair (bp) position of each SNP, all reported according to Ensembl build 37, the major and minor alleles along with the MAF, the estimated effect size (β) and standard error (SE) of the effect size.

The second step of Bayesian fine-mapping requires the calculation of an LD matrix with the estimation of LD between variants within each risk locus using Pearson correlations. Recent studies support the use of original genotype data, where available, for the calculation of LD structure over the use of reference panels [159]. This is not only because the LD matrix will then match exactly the study population, but because the size of the reference panel for calculation of the LD matrix needs to scale with the GWAS sample to maintain optimal fine-mapping performance. Fine-mapping with smaller reference panels (e.g. 100 individuals) misleadingly results in smaller credible sets, with much lower coverage over variants than the larger reference panels or original genotype data. Benner *et al* demonstrated that a reference panel of 1,000 individuals is sufficient when summary statistics originate from a GWAS with 5-10,000 individuals. For this reason, I used full original genotype data from Ji *et al* [42] to calculate the SNP correlation matrix, using

computational software program *LD Store* v1.1 [159]. Importantly, this ensured that the ancestry of the LD cohort matched exactly the ancestry of the GWAS cohort. For any two variants, LD information was only extracted if absolute Pearson correlation was above zero. One of the drawbacks of fine-mapping methods based on summary statistics is that they are more sensitive to the choice of data-set used to calculate the LD matrix. I therefore also conducted fine-mapping analyses using LD structure derived from the UK10K project reference panel, to assess any differences dependent upon the choice of LD matrix.

FINEMAP assumes that each SNP is causal with prior probability of 1 / number of SNPs in the genomic region. I left prior probabilities for the number of independent causal SNPs in the genomic region unspecified, however repeated the analyses with iterations assuming between one and five independent causal variants per region. For each of the 15 risk loci, the analysis output included model-averaged posterior summaries for each SNP, posterior summaries for each causal configuration, posterior summaries for the number of independent signals per region and the 95% credible sets for each causal signal conditional on other causal signals in the genomic region. To declare a locus fine-mapped to single causal variant, I defined that the PP of causality for that single variant had to be $\geq 95\%$. Evidence for additional independent signals within a risk locus was taken as a PP of >50% in support of two or more independent signals within a risk locus. In order to check the fine-mapping assumption that each locus contained a true causal variant and all potential causal variants had been included within the analysis, I searched the UK10K and 1000 Genomes reference panels for any SNPs in moderate or high LD (defined as an r^2>0.5 and >0.8 respectively) with the most probable fine-mapped variant, noting any that were not included within the analysis.

2.3.2 Functional annotation

The majority of common disease-associated variants are located within non-coding regions of the genome. These non-coding variants are thought to overlap functional DNA elements involved in gene regulation such as transcription factor binding sites, open chromatin or gene enhancer regions [117]. Functional annotation complements statistical fine-mapping methods by providing independent information about the likely biological function of each variant. In recent years functional annotation profiles have been developed across many hundreds of tissue and cell types collated into databases such as the Encyclopedia of DNA Elements (ENCODE) database [161].

I aimed to further define the function of those non-coding variants included within the credible sets from the above fine-mapping analysis, by assessing which credible causal variants overlapped functional regions of the genome. Several existing fine-mapping approaches have integrated functional annotation as a means of prioritising causal variants. *Paintor* is one example of such an approach which integrates association strength with

functional genomic annotation data to improve the accuracy in selecting credible causal variants for functional validation [155]. Genomic Annotation Shifter *(GoShifter)* is a statistical approach that tests for enrichment of functional annotations overlapping a disease-associated variant, as a means to prioritising variants for further functional follow-up [162]. *GoShifter* identifies all variants in high LD (defined as an $r^2>0.8$) with the lead GWAS variant, and the median size (in bp) of the tested annotation feature, X. *GoShifter* defines the 'locus' as the region between the two furthest SNPs linked with the lead variant, plus twice the median size of X. *GoShifter* identifies the proportion of loci in which at least one SNP overlaps X, and compares this to a null distribution of iterations, generated by repeated random shifting of the site of X within the locus. The p-value is computed as the proportion of iterations for which the number of overlapping loci is equal to or greater than that for the tested SNPs. *GoShifter* then uses stratified enrichment analysis to asses the significance of an overlap with X, independent of overlap with any other colocalising annotation, Y. This involves separating the locus into two fragments- that which overlaps Y, and that which does not, Y_0. X is then shifted separately within Y and Y_0 to generate the null distribution, and the significance of the observed overlap assessed by the proportion of loci in which any SNPs overlaps annotation X in Y or Y_0. The delta overlap describes the difference between the observed proportion of loci overlapping X and the mean proportion of loci overlapping X under the null derived by shifting and provides a measure of the effect size of the observed enrichment. In the absence of enrichment, the observed overlap will be close to the mean overlap of the null, and delta-overlap will be close to 0, whereas stronger enrichment corresponds with larger delta overlap. Finally, to identify loci in which the overlap between a SNP and an annotation is particularly informative and thus should be higher priority for further functional evaluation, the 'overlap score' is calculated. The overlap score describes the probability that each locus overlaps an annotation by chance, and is only computed for loci that overlap the annotation in the observed data. Loci with better (lower) overlap scores suggest significant enrichment and are therefore proposed to be higher priority for functional evaluation of causal variants.

To identify those non-coding credible variants that overlapped gene regulatory features, I used a modified version of *GoShifter*, with modifications similar to those implemented in a published study by Ulirsch *et al* [163]. These modifications included substitution of the high-LD variants with the credible set variants from my fine-mapping analysis. Therefore, the 'locus' provided to *GoShifter* was defined as the region between the two furthest credible SNPs linked with the variant with the highest PP of causality from fine-mapping, plus twice X (the median size in bp of the tested annotation). In the first stage of the analysis *GoShifter* therefore takes all PSC credible causal variants across all non-coding PSC loci and tests for regulatory features enriched across all PSC credible causal variants.

In the second stage, it tests for overlap of each locus with those enriched features from the first stage of the analysis, to prioritise credible causal variants based upon those with the lowest overlap score.

I used the ENCODE v4 database [161] for all annotations, including promoters, enhancers, histone acetylation marks and DNase-I hypersensitive sites for 28 whole tissue and immune cell sub-types, relevant to PSC. I defined relevant PSC tissues as any immune cell type and any tissue from an organ system affected by PSC. Instead of using a set of high LD variants, for each independent signal within each locus, I input the 95% credible set of variants defined from my fine-mapping study and performed 20,000 local shift iterations per annotation. I calculated delta-overlap scores to measure the enrichment of overlap between annotations and credible variants. I adjusted the enrichment p-values for multiple testing using the Benjamin Hochberg FDR correction at 5% [164]. I chose this less stringent form of multiple testing correction as some annotations are not independent (for example enhancer marks are made up from a combination of annotations) and therefore a more lenient method than the Bonferonni correction is required. I calculated overlap scores for those loci that overlapped annotations. Lower scores suggest significant enrichment and higher priority as causal variants. Although *GoShifter* does not define a overlap score threshold to interpret significance, I prioritised variants from credible sets based upon the variant or variants with the lowest feature overlap score per PSC risk locus.

For ease of reference, SNPs are referred to according to their RSID. The 15 PSC risk loci are numbered 1 to 15, and referred to according to their PSC risk region number, chromosome and bp position of the lead GWAS SNP for the region, according to Ensembl build 37 (see Table 2.1). Where a region is referred to according to a nearby candidate gene, it is important to note that candidate genes are assigned according to locality and biological plausibility and do not necessarily describe a proven causal association between variant and gene, unless specifically stated otherwise.

2.4 Results

I conducted fine-mapping of the fifteen PSC risk loci in the GWAS dataset published by Ji *et al* [42]. Genotyping, quality control (QC) and imputation had been previously conducted on the genotype data of 24,751 individuals of European ancestry, including 4,796 PSC cases 19,955 controls, as previously described by Ji *et al.* Fine-mapping of the PSC risk loci identified nineteen independent signals across fifteen risk loci (Figure 2.2). Evidence supported just one causal signal within eleven of the fifteen regions with >50% certainty. For seven of those regions the PP supporting one independent causal signal was >70%. In the remaining four regions the evidence supported the presence of two independent signals with >50% certainty. For each signal detected, variants were sorted

by the PP of causality and added to the credible set of associated variants until the sum of their PPs exceeded 95%. These credible sets ranged in size from one to sixty-two variants (Table 2.1). For all loci, review of the 1000 Genomes and UK10K data-sets revealed that there were no SNPs in high LD (r^2>0.8) with the most probable causal variants, missing from the analysis. Two of the fifteen risk loci resolved to a single causal variant with >95% certainty and three loci to larger credible sets where one credible variant was assigned >50% PP of causality. Of these five variants, one was enriched for a significant protein-coding change, one caused direct disruption to a splice site and three overlapped tissue-specific epigenetic marks in PSC-relevant tissue- and cell-types (Table 2.2).

Sensitivity analysis using a different LD matrix derived from UK10K reference cohort demonstrated that when considering one independent causal variant, choice of LD matrix did not affect the most probable SNP or PP of causality according to *FINEMAP*. As expected, when considering more than one independent causal variant in each region, the results of fine-mapping were more sensitive to the choice of LD matrix. For five of the fifteen loci, the most probable fine-mapped causal variant using LD from GWAS and UK10K remained the same. For seven of the fifteen loci the credible sets derived from both analyses were identical, with a slight redistribution of the PP of causality to a neighbouring, highly-correlated variant. Fine-mapping attempts in the remaining three loci resulted in credible sets containing more than forty credible causal variants with complex patterns of LD in all three regions, which was not improved with a different choice of LD matrix.

GoShifter identified significant enrichment of credible causal variants from all 19 independent signals with promoter and enhancers annotations in all five tested immune-cell types (B-cell, CD14+ monocytes, macrophage, peripheral blood mononuclear cells and T-regulatory cells) and eleven gastro-intestinal tissues (colonic mucosa, duodenal mucosa, gastroesophageal sphincter, large intestine, liver, duodenal muscle, rectal smooth muscle, rectal mucosa, Sigmoid colon, small intestine and transverse colon). It is important to note that *GoShifter* is applied only to variants within non-coding regions and therefore PSC regions 4 and 10 (loci within coding regions) and PSC region 15 (a splice site region), were not included in this analysis.

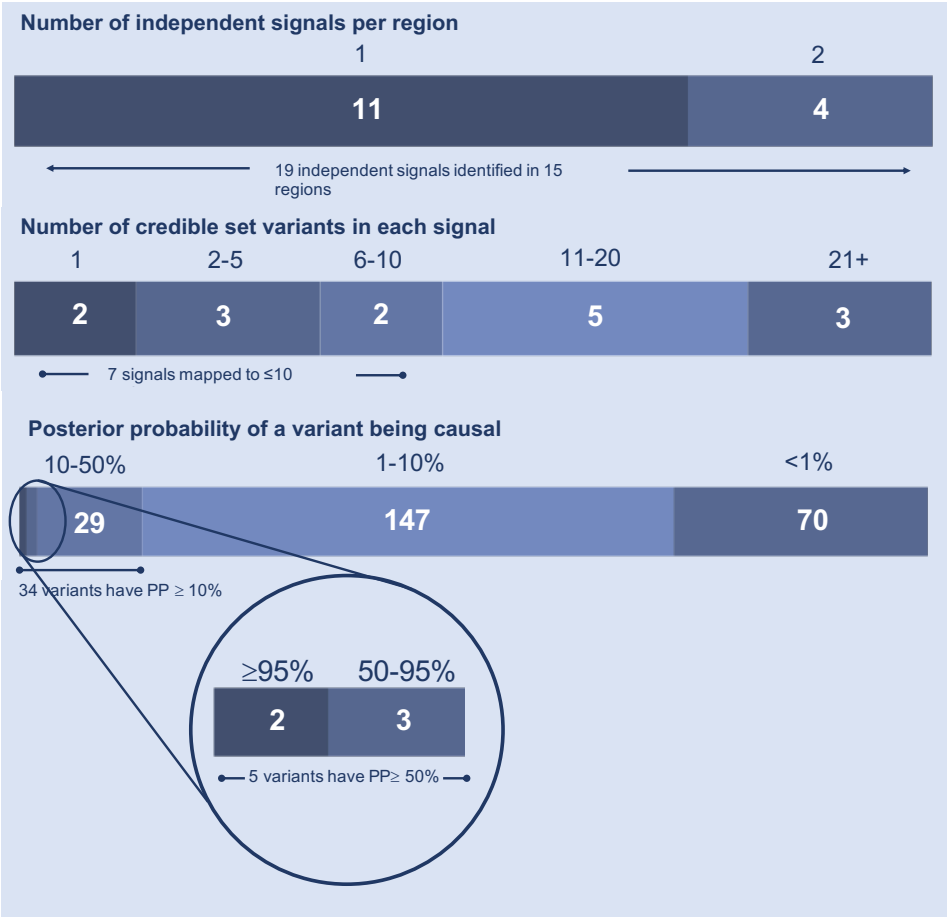

Figure 2.2: Summary of fine-mapping the PSC risk loci.

Table 2.1: Fine-mapping of PSC risk loci

			Region			Fine-mapping			
Region	**Signal**	**Chr**	**Candidate Gene**	**Lead GWAS SNP**	**Position (b37)**	**SNP PP_{max}**	**Position (b37)**	**PP Causal**	**Credible set size**
1	1	1	*MMEL1*	rs3748816	2526746	rs61763697	2810791	0.07	62
2	1	2	*BCL2L11*	rs72837826	111933001	rs72837826	111933001	0.18	12
3	1	2	*CD28*	rs7426056	204612058	rs5837875	204647878	0.19	6
	2					rs231799	204707417	0.17	
4	1	3	*MST1*	rs3197999	49721532	rs11716895	49762779	0.11	13
	2					rs13083791	49721798	0.07	
5	1	3	*FOXP1*	rs80060485	71153890	rs80060485	71153890	0.99	1
	2					rs36023390	71523093	0.14	
6	1	4	*IL2-IL21*	rs13140464	123499745	rs13119723	123218313	0.09	50
7	1	6	*BACH2*	rs56258221	91030441	rs7750271	91036225	0.20	12
8	1	10	*IL2RA*	rs4147359	6108439	rs4147359	6108439	0.46	5
9	1	11	*CCDC88B*	rs663743	64107735	rs35247680	63884747	0.61	2
	2					rs663743	64107735	0.41	
10	1	12	*SH2B3*	rs3184504	111884608	rs3184504	111884608	0.99	1
11	1	16	*CLEC16A*	rs725613	11169683	rs725613	11169683	0.16	12
12	1	18	*CD226*	rs1788097	67543688	rs1610555	67543147	0.08	44
13	1	19	*PRKD2*	rs313839	47221557	rs313839	47221557	0.23	14
14	1	21	*ETS2*	rs2836883	40466744	rs4817988	40468838	0.58	10
15	1	21	*UBASH3A*	rs1893592	43855067	rs1893592	43855067	0.62	5

PP; posterior probability of causality

Table 2.2: PSC risk loci overlapping gene regulatory features

Region	Signal	Chr	Cand. gene	FINEMAP SNP	FM PP	GoShifter SNP	FM PP	Overlaps promoter in these tissue	Overlaps enhancer in these tissues
1	**1**	1	MMEL1	rs61763697	0.07	rs60733400	0.02	SI,RM,L,CM,MD, CD14,SC,PBMC,TC,DM,LI,MR,SC	SC,SI,L,GOS,BC,MR,TC,CM,CD14, PBMC,LI,DM, MD
2	**1**	2	BCL2L11	rs72837826	0.18	rs72836345	0.18	SI,RM,L,CM,CD14,SC,PBMC,TC,DM,LI,MR,M	SC,GOS,BC,RM,TC,CM,CD14,PBMC
3	**1**	2	CD28	rs5837875	0.19	rs5837875	0.19	RM,CM,PBMC,L,DM,MR, RM,CD14,M,TR	SI,SC,CD14, PBMC,RM,GOS,CM,TR
	2			rs231779	0.16	rs231779	0.16	RM,CM,PBMC,L,DM,MR, RM,CD14,M,TR	SI,SC,CD14, PBMC,RM,GOS,CM,TR
5	**1**	3	FOXP1	rs80060485	0.99	rs80060485	0.99	CD14,MR,RM,BC,TR,M	SC,GOS,MR,RM,LI,BC,TR
	2			rs36023390	0.14	rs36023390	0.14	CD14,MR,RM,BC,TR	SC,GOS,MR,RM,LI,BC,TR,CM
6	**1**	4	IL2-IL21	rs13119723	0.09	rs67963613	0.01	SI,RM,L,CM,DM,SC,PBMC,CD14,TC, DM,LI,MR,RM,M,TR	SC,SI,L,GOS,BC,RM,MR,TC,CM,CD14,PBMC,L,MD,DM,TR
7	**1**	6	BACH2	rs7750271	0.20	rs7750271	0.20	SI,RM,L,CM,MD,SC,PBMC,CD14,TC, DM,LI,MR,RM,M,BC	SC,SI,L,GOS,BC,RM,MR,TC,CM,PBMC,CD14
8	**1**	10	IL2RA	rs4147359	0.46	rs4147359	0.46	CD14,M,L,DM,CD14,CM,SC,RM,TR,LI,CM,BC,SI,MR,MD	SC,SI,L,GOS,BC, RM,TC,CM, CD14,PBMC,MD, L,DM, CM,TRs,LI,BC,SI,MR
9	**1**	11	CCDC88B	rs35247680	0.61	rs35247680	0.61	SI,RM,L,CM, MD, CD14,SC,PBMC, TC,DM,LI,MR,M	L,SC,GOS,BC,RM, TC,CM,CD14,PBMC,LI ,SI,DM
	2			rs663743	0.41	rs663743	0.41	SI,RM,L,CM, MD, CD14,SC,PBMC, TC,DM,LI,MR,M,L	L,SC,GOS,BC,RM, TC,CM,CD14,PBMC,LI ,SI,DM,TC
11	**1**	16	CLEC16A	rs725613	0.16	rs113344842	0.02	SC,SI,L,GOS,BC,MR,TC,CM,CD14,RM,MD,DM,CM, PBMC,RM,TRs,LI	CM,CD14,PBMC,MR,RM,M,L,DM,CD14,CM,PBMC,SC,BC, RM,SI,TR,LI,CM,MR,MD
12	**1**	18	CD226	rs1610555	0.08	rs4891781	0.03	PBMC,SC,TR,MD,SI,PBMC	LI,RM
13	**1**	19	PRKD2	rs313839	0.23	rs313839	0.23	SI,RM,L,CM,MD,CD14,SC,PBMC,TC,DM, LI,MR,RM,M	SC,SI,L,GOS,BC,RM,MR,TC,CM,CD14,PBMC,LI,MD,DM
14	**1**	21	ETS2	rs4817988	0.58	rs2836883	0.05	CM,RM,CD14,PBMC,M,SI	SC,SI,L,GOS,RM,TC,CM, CD14,PBMC,MD,DM,BC

BC; B-cell, CD14; CD14+ monocyte, CM; Colonic mucosa, DM; Duodenal mucosa, GOS; gastroesophageal sphincter, LI; Large intestine, L; liver, M; macrophage, MD; Duodenal muscle.
MR; Rectal smooth muscle, PBMC; Peripheral blood mononuclear cell, RM; Rectal mucosa, SC; Sigmoid colon, SI; small intestine, TC; Transverse colon, TR; T-regulatory cell

2.4.1 Loci mapped to a single causal variant

Two of the fifteen PSC risk loci mapped to a single causal variants with $\geq 95\%$ PP of causality. The first single variant credible set was in PSC region 5 (Chr3:71153890), where the GWAS lead SNP, rs80060485 at position 71153890, was predicted to be causal with a PP of 99%. *FINEMAP* strongly supported the presence of a second independent signal within this region with 83% certainty. Signal 2 could not be well fine-mapped with 14% PP of causality for the most probable causal variant, rs36023390 at position 71523093 (Figure 2.3a). The presence of two independent causal variants was supported by the finding that the causal configuration with the highest PP contained both rs80060485 and rs36023390 and that these two SNPs were not correlated ($r^2=0$). *GoShifter* identified that the credible causal variant for signal 1, rs80060485, overlapped promoter and enhancer marks in three immune cell types and ten gastrointestinal tissue types (see Table 2.2).

The fine-mapped causal variant, rs80060485, occurs within an intron of *FOXP1* (forkhead box P1), a transcription factor with an important role in B- and T-cell differentiation. CD4+ T-follicular helper (T-FH) cells, are a specialised T-cell subset found in germinal centres, which interact with B-cells, inducing antibody formation and response. Foxp1 is a negative regulator of T-FH cell differentiation, directly and negatively regulating IL-21 production [165]. Foxp1-deficient CD4+ T cells preferentially differentiate into CD4+ T-FH cells, resulting in substantially enhanced germinal centre and antibody responses. T-FH cells can also be found in the periphery where they are characterised by the expression of chemokine receptor type 5 (CXCR5) and the inhibitory receptor, programmed death 1 (PD-1). Circulating T-FH cells lacking the chemokine (C-C motif) receptor 7 (CCR7), closely resemble lymphoid tissue-derived T-FH cells, that are pathogenic in autoimmunity [166]. Interestingly, the frequency of potentially pathogenic CCR7lowCXCR5+PD-1+CD4+ T-FH cells is increased in patients with PSC, compared to healthy donors [167], suggesting Foxp1 and T-FH cells may have an important role in PSC pathogenesis. Whilst it is not yet clear whether or how the expression of *FOXP1* is affected by the intronic rs80060485 variant, this analysis demonstrates that this variant overlaps several important markers for active enhancers, suggesting a mechanism via which this variant may exert a quantitative effect upon the expression of *FOXP1* or several other genes within the region.

The second locus fine-mapped to a single causal variant with $\geq 95\%$ PP of causality was PSC region 10 (Chr12:111884608). Fine-mapping confirmed that rs3184504, the lead GWAS SNP, was the most probable causal variant with 99% certainty (Figure 2.3b). The rs3184504 SNP, is a multi-allelic missense variant which is positioned within exon 3 of the *SH2B3* (Scr homology 2 adaptor protein 3) gene. The rs3184504*A and rs3184504*C alleles code for a basic polar arginine and the rs3184504*G allele codes for a polar glycine at position 262 in the pleckstrin homology domain of the SH2B3 protein. The minor allele

for this locus, present at a frequency of 15%, is the PSC risk increasing rs3184504*T allele, which codes for a non-polar tryptophan at this position. Analysis of the functional effect of this missense mutation using Ensembl's variant effect predictor (VEP) assigned the rs3184504*C>T SNP a PHRED-like scaled CADD score of 11.08, where a score of ≥10 indicates polymorphisms predicted to be within the 10% most deleterious substitutions in the human genome [168].

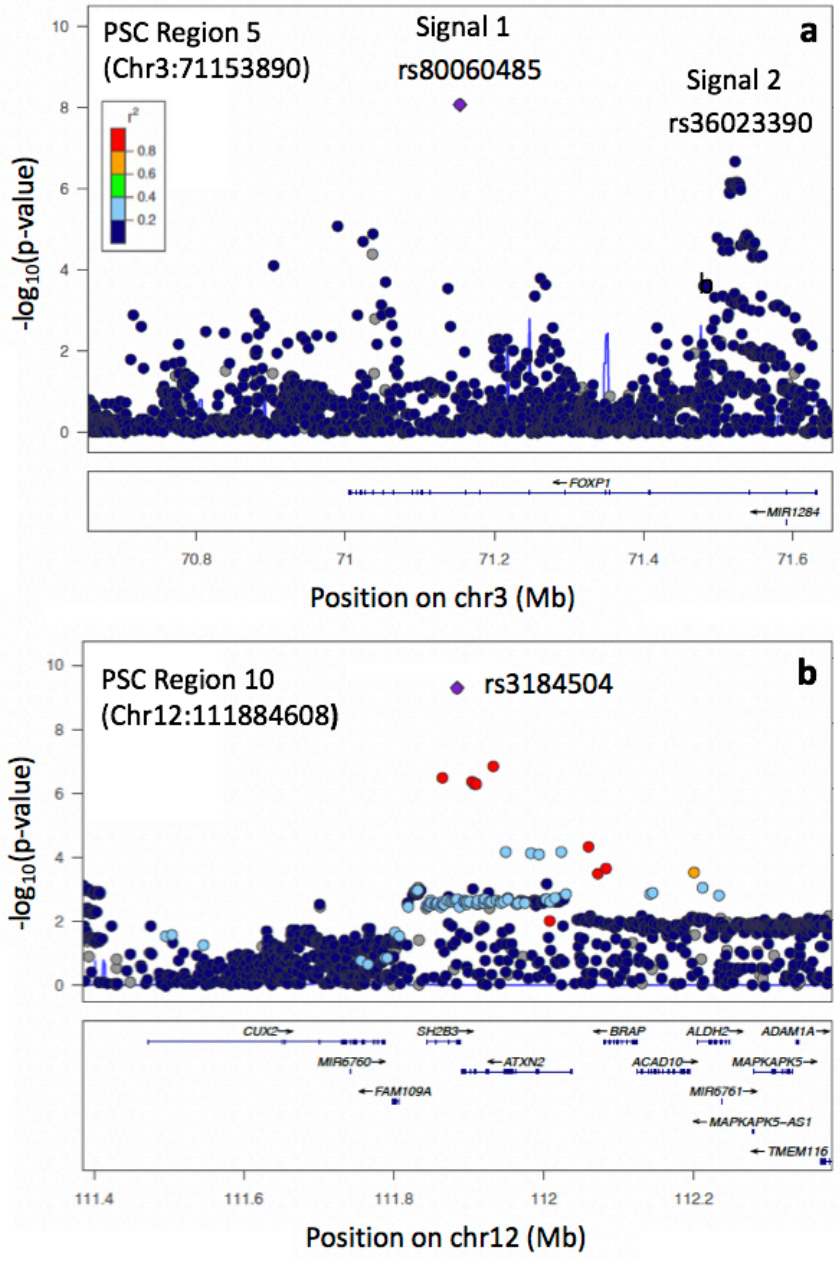

Figure 2.3: Regional association plots for PSC risk loci mapped to single variants.

SH2B3 is an interesting gene in the pathogenesis of PSC, as it is a negative regulator of T-cell activation, TNF production, and Janus kinase (JAK) 2 and 3 signalling. It

is known to encode the T-cell adapter protein LNK, which regulates T-cell receptor-, growth factor- and cytokine receptor-mediated signalling [169]. The *SH2B3* locus is a shared risk locus with several other IMDs and rs3184504 remains the lead SNP in GWAS of coeliac disease (CeD), rheumatoid arthritis (RhA), type 1 diabetes mellitus (T1DM) and autoimmune hepatitis (AIH) [169–171]. Fine-mapping of this risk locus in RhA predicted that rs3184504 was the most probable causal variant for this locus with 76% PP of causality [114]. Expression quantitative trait loci (eQTL) studies have shown that rs3184504 is associated with increased expression of genes involved in IFNγ production [172]. Furthermore, functional investigation of this locus has shown that peripheral blood mononuclear cells isolated from individuals homozygous for the rs3184504*A allele, which increases risk of RhA and T1DM, display increased production of pro-inflammatory cytokines in response to bacterial stimuli compared to individuals homozygous for the non-risk G allele [173]. The same study also suggested that the SH2B3 protein has an inhibiting function on the MDP-NOD2-RIP2 pathway, which responds to bacterial ligands, with disease-associated alleles causing diminished inhibitory activity of SH2B3. Unfortunately, they did not include analysis of individuals homozygous for the minor rs3184504*T allele, which not only increases the risk of PSC, but also of AIH [174], suggesting the resultant Arg262Trp amino acid substitution may contribute to an aberrant immune- and inflammatory-response targeted at the hepato-biliary system.

2.4.2 Variants with a greater than 50% posterior probability of causality

Three signals mapped to credible sets containing more than one variant, where one variant within each credible set had >50% PP of causality. The first was within PSC region 9 (Chr11:64107735), where rs35247680 at position 63884747 was predicted to be causal with 61% PP. This SNP is a non-coding variant within an intron of *MACROD1*. *GoShifter* demonstrated that this variant overlapped promoter marks enriched in four immune cell-types and nine gastrointestinal tissues and overlapped enhancer marks in four immune cell-types and eleven gastrointestinal tissues (Table 2.2), thereby suggesting several mechanisms via which this credible causal variant may regulate expression of nearby genes. There was evidence to support a second independent signal within this region with 68% certainty (Figure 2.4a). The most probable causal variant for signal 2 was the previously reported lead GWAS SNP for this locus, rs663743 at position 64107735. Independence of these two signals was supported by the fact that these variants were not highly correlated with one another (r^2=0.02). The rs663743 SNP is non-coding and within the 5' untranslated region overlapping a promoter region for *CCDC88B* (coiled-coil domain containing 88B). However with 41% PP of causality attributed to rs663743, signal

2 could not be considered fine-mapped.

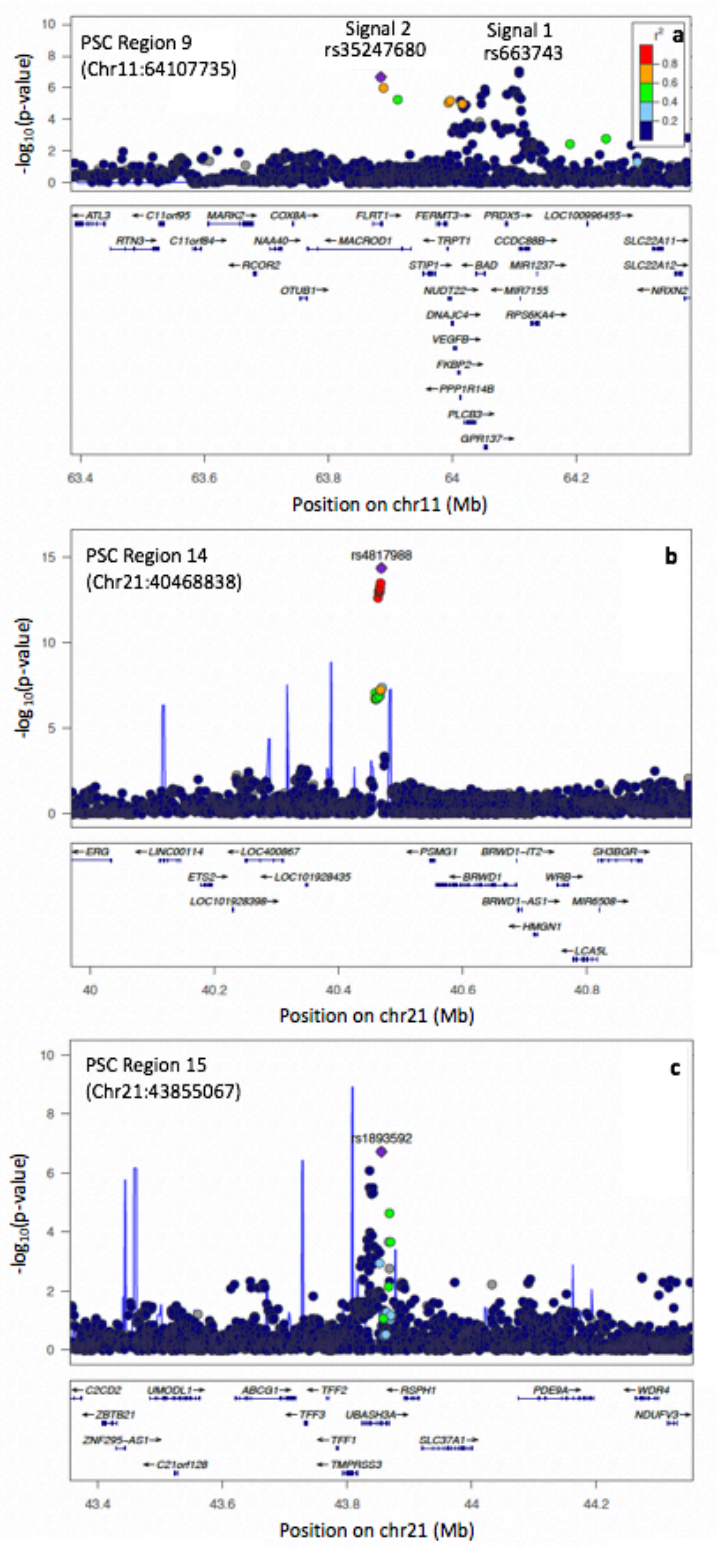

Figure 2.4: Regional association plots for PSC risk loci mapped to casual variants with >50% posterior probability of causality.

The second region mapped to a credible set containing a variant with >50% PP of causality was in PSC region 14 (Chr21:40466744). The lead GWAS variant for this region,

rs4817988, is highly correlated (r^2>0.8) with more than ten other variants of genome-wide significance that lie in close proximity, all with similar measures of association (Figure 2.4b). Fine-mapping of this region identified a 95% credible set containing ten of these high-LD variants with the most probable causal variant for this region, rs4817988 at position 40468838 with 58% PP of causality. The next most probable causal variants were rs2836884 at position 40467643 (8% PP) and rs2836883 at position 40466744 (5% PP). This locus is located 5' upstream of *PSMG1*, which is the commonly quoted candidate gene for this region, based upon its genomic locality and evidence from the study of paediatric IBD colon where levels of PSMG1 were increase compared to healthy colon [175]. Importantly, although rs4817988 is located in a non-coding region, it overlaps a CTCF transcription factor binding site and has been correlation with the expression of *ETS2 (v-ets avian erythroblastosis virus E26 oncogene homolog 2)* in the GTEx eQTL analysis of whole blood [176]. *GoShifter* prioritised rs2836883, followed closely by rs4817988 with the lowest overlap scores and found these variants to overlap promoter marks in three immune cell types and three gastrointestinal tissue types (Table 2.2). This same locus has also been associated with IBD and has been the subject of an IBD fine-mapping study [56]. Huang, Fang, Jostins *et al* resolved the *ETS2* locus to a credible set of 10 variants, with a 39% PP attributed to the most probable variant, rs9977672. In PSC, I mapped this region to a 10 variant credible set, 8 of which overlapped with the IBD credible set for this region, however prioritising a different variant, rs4817988 at position 40468838, as causal with 58% certainty. The two non-overlapping variants within the IBD credible set, one of which is the most probable IBD fine-mapped variant, are both present within the PSC data-set. It is likely that for this region, the same variant is causal is both PSC and IBD, although further analysis is required to validate this hypothesis.

The third locus fine-mapped to a credible set containing a variant with >50% PP of causality was PSC region 15 (Chr21:43855067). Ji *et al* reported an association between the Chr21:43855067 locus and PSC risk, driven by lead SNP rs1893592, and proposed *UBASH3A* as the most likely gene affected by this risk locus on the basis that this SNP was an eQTL of *UBASH3A* in one B-cell only [129] and two whole blood analyses [118, 177] (Figure 2.4c). Fine-mapping of this region confirmed that rs1893592 at position 43855067, which is located three bases downstream of the 10^{th} exon of *UBASH3A* within the splice consensus sequence, was the most probable causal variant in this region with 62% PP of causality. The PSC risk reducing rs1893592*C allele, disrupts the conserved 5' splice donor sequence at this position, and is predicted to cause partial retention of the downstream intron and possible non-stop mediated decay [178]. I fine-mapped this locus to a credible set containing just four additional variants, each located within intronic regions of *UBASH3A*, but with a low individual probability of causality; rs11203203 (14%), rs3788013 (9%), rs9974339 (6%) and rs876498 (6%), and all in low LD (r^2<0.6)

with the most probable SNP, rs1893592. Interestingly, this locus has also been reported as associated with T1DM, where a fine-mapping study has identified the second most probably PSC SNP, rs11203203 at Chr21:43836186, as the most probable causal variant for this locus in T1DM with 39% PP of causality [114]. In this T1DM fine-mapping study, the 95% credible set contained four variants, of which only rs11203203 is contained with both the PSC and T1DM credible sets. Notably, a review of the summary statistics from both data-sets showed that SNPs from both credible sets were considered within both the PSC and T1DM fine-mapping analyses. Whilst it is possible that different SNPs within this same locus may precipitate different IMDs, it is more likely, where the credible sets overlap, that it is the same causal variant responsible for both IMDs. Further work to colocalise the signals in PSC and T1DM at this locus would be helpful to establish a shared causal variant.

2.4.3 Variants with a greater than 20% posterior probability of causality

Fine-mapping of two loci resulted in credible sets containing at least one causal variant with >20% PP of causality. Although these loci could not be considered fine-mapped, a large credible set with >20% PP attributed to one SNP could, in combination with functional annotation, provide useful information about the potential causal variants within a locus. The first locus containing at least one causal variant with >20% PP of causality was PSC region 8 (Chr10:6108139 region). The lead GWAS SNP for this locus, rs4147359 (Chr10:6108139), located upstream of *IL2RA*, was predicted to be the most probable causal variant for this region with 46% PP of causality. The 95% credible set included four other variants, two intergenic and two intronic variants (Figure 2.5a). These variants were found to be enriched for overlapping regulatory regions in PSC-relevant tissues. *GoShifter* identified variants within this locus as potentials for functional follow up with one of the lowest overlaps scores across all credible variants from all non-coding loci observed for rs4147359, also the most probable causal variant from fine-mapping of this locus. This variant is located within an intergenic region and overlaps a marker of active transcription, H3K36me3. This suggests that the mechanism via which rs4147359 may increase PSC risk is through modulation of an active transcription histone acetylation mark, although the downstream gene and direction of effect cannot be identified from either of these analyses. *FINEMAP* could not distinguish whether there were one or two independent signals within the region, with equal evidence for both, although the most probable causal configuration contained just one single variant, rs4147359. Interestingly, this locus has been previously fine-mapped in a study of individual and combined summary statistics for T1DM and RhA [114]. In the combined T1DM and RhA data, this locus

was fine-mapped to a credible set containing 3 SNPs; rs706778 (89% PP), rs7072793 (4% PP) and rs7096384 (3% PP). Reassuringly, two of these T1DM/RhA credible set variants (rs706778 and rs7072793) were also included within the PSC credible set, with 12% and 9% PP of causality respectively. Both PSC and T1DM/RhA GWAS datasets included all SNPs within the credible sets for both of these IMDs. Given that the T1DM/RhA fine-mapping study included data from 11,475 RhA cases and 9,334 T1DM cases, compared to the 4,796 PSC cases analysed in this study, the T1DM/RhA fine-mapping study was better powered to fine-map individual rick loci. Therefore, where fine-mapping of risk loci within one data-set is inconclusive, the sharing of genetic architecture between IMDs means that other fine-mapping studies of the same locus can be informative.

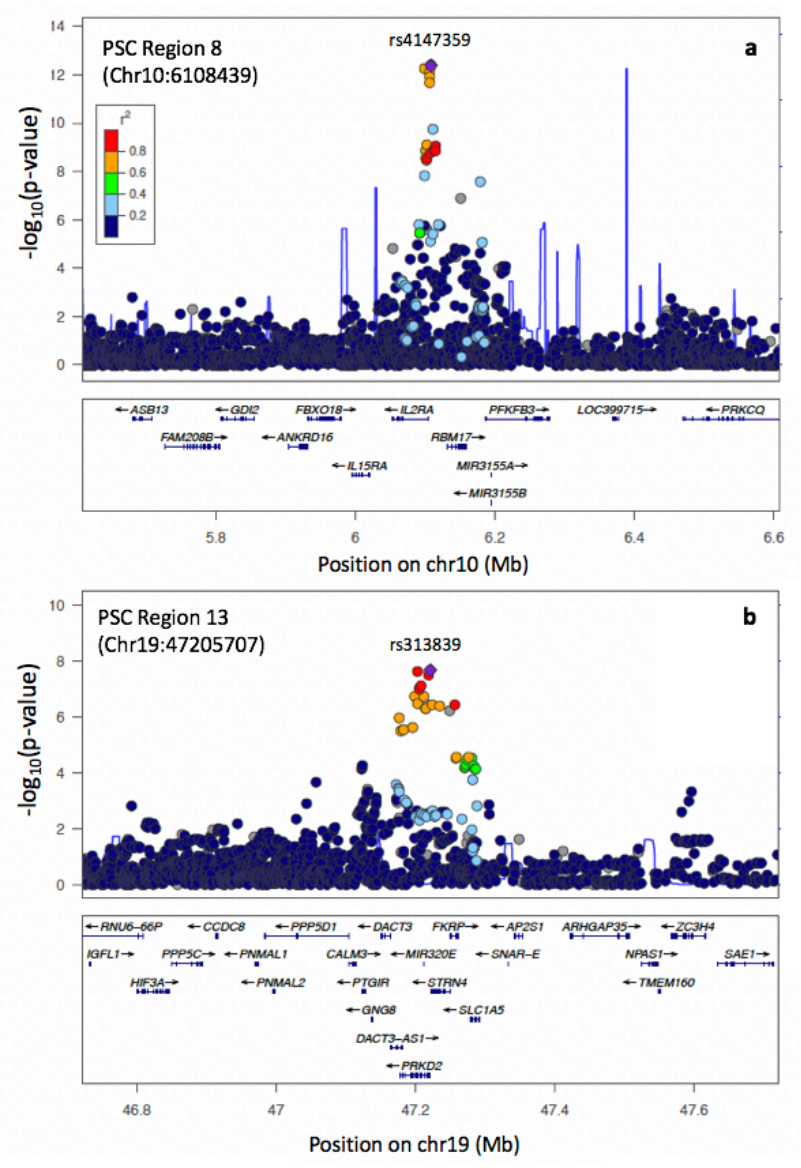

Figure 2.5: Regional association plots for PSC risk loci mapped to casual variants with >20% posterior probability of causality.

The second locus containing at least one causal variant with >20% PP of causality was PSC region 13 (Chr19:47205707). This locus was fine-mapped to a credible set containing fourteen variants, with the most probable causal variant for the region being rs313839 at position 47221557 with 23% PP of causality (Figure 2.5b). The rs313839 variant lies within an intron region that has been associated with binding of many transcription factors though chromatin immunoprecipitation (ChIP)-sequencing studies and overlaps a promoter region for a gene called *PRKD2*. Notably, rs313839 was also prioritised by *GoShifter* as it overlaps transcription activation marks, H3K4me3 and H3K27ac, in CD14+ monocytes. This finding is in keeping with the known role of *PRKD2* in monocyte migration and adhesion [179], [180]. Analysis of the major and minor allele sequences using *PROMO*, which identifies putative transcription factor binding sites [181], showed that rs313839 C>G (where rs313839*G is the PSC risk increasing allele) resulted in loss of binding motifs for transcription factors LEF-1 TCF1A, TCF-4E, DEF:GLO:SQUA, TCF-3 and ADR1, and gain of motifs for EIIE-A, VSF-1, V-MYB and MYB2. Although further investigation is required to identify the genes affected by these transcription factors, the results of this study suggest that the most probable causal variant, rs313839, modulates transcription factor binding, with evidence supporting *PRKD2* as a likely candidate gene for this locus.

2.4.4 Loci not well-resolved with fine-mapping

Several regions could not be fine-mapped to credible sets with the majority of the PP for causality attributed to one variant. However, four regions were resolved to relatively small credible sets. PSC region 3 (Chr2:204612058) was mapped to a six variant credible set, with the majority of the PP for causality (19%) attributed to rs5837875 at position 204647878, a variant located in an transcription factor binding site and associated with expression of *CD28* in one whole blood eQTL analysis [118]. PSC region 2 (Chr2:111933001), region 7 (Chr6:91030441) and region 11 (Chr16:11169683) were each resolved to credible sets of twelve variants, in which the respective most probable causal variants had PPs of 0.18, 0.20 and 0.16. Although PSC region 2 (Chr2:111933001) and region 11 (Chr16:11169683) have been reported as associated with IBD and PBC respectively, the Chr2:111933001 locus has not been considered in either of the published IBD fine-mapping studies [56], [60] and there have been no published fine-mapping studies in PBC, to date. PSC region 7 (Chr6:91030441) has been fine-mapped to a credible set of nine variants in RhA [114]. Whilst the most probable causal variants differed between PSC and RhA, all nine variants within the RhA credible set were contained within the PSC credible set.

Four PSC risk loci were not well resolved with fine-mapping, defined by large credible sets with the most probable causal variant assigned $\leq 10\%$ PP of causality. PSC region 4 (Chr3:49721532) *MST1* (Figure 2.6a) and region 6 (Chr4:123499745) *IL2-IL21* (Figure 2.6b) both contain many variants in tight LD with one another, extending over a wide

genomic region of >500Mb, all with very similar strengths of association. Both of these loci have been associated with IBD, however fine-mapping in IBD was unable to resolve the Chr3:49721532 *MST1* and Chr4:123499745 *Il2-Il21* loci, with credible sets containing 437 and 29 variants respectively, a likely consequence of the extended, complex patterns of LD observed within these regions. PSC region 1 (Chr1:2526746) (Figure 2.7a) and region 12 (Chr18:67543688) (Figure 2.7b) both contain many variants with similar strengths of association, all in tight LD with one another. Under such circumstances *FINEMAP*, which utilises subtle differences in strengths of association between tightly correlated variants, performs less well, and prioritisation of non-coding causal variants using functional annotation of genomic regions become more important to infer causality. The *GoShifter* overlap scores for these two loci were comparatively high, compared to other loci, suggesting *GoShifter* was unable to easily prioritise causal variants from these loci in comparison with other PSC risk loci. However, based upon the variant with the lowest overlap score relative to other credible variants within the same locus, for PSC region 1 (Chr1:2526746) *GoShifter* prioritised rs60733400 at position 2516781 with an overlap score of 0.17. This variant overlaps several regulatory features including H3K27ac, an active enhancer marker in CD14+ monocytes. For PSC region 12 (Chr18:67543688), *GoShifter* prioritised rs4891781 at position 67524646 with an overlap score of 0.14, as it overlapped an H3K9ac mark in peripheral blood mononuclear cells (PBMCs). This variant is in tight LD with the most probable causal variant from fine-mapping, and lies only 10Kb upstream. However in both cases the overlap scores were relatively high (>0.14) suggesting that in comparison to the other PSC risk loci, these loci should not be prioritised for further for functional follow-up. PSC region 12 (Chr18:67543688) has been associated with T1DM, however fine-mapping of locus in T1DM was no more successful, with a reported credible set containing 32 variants [114].

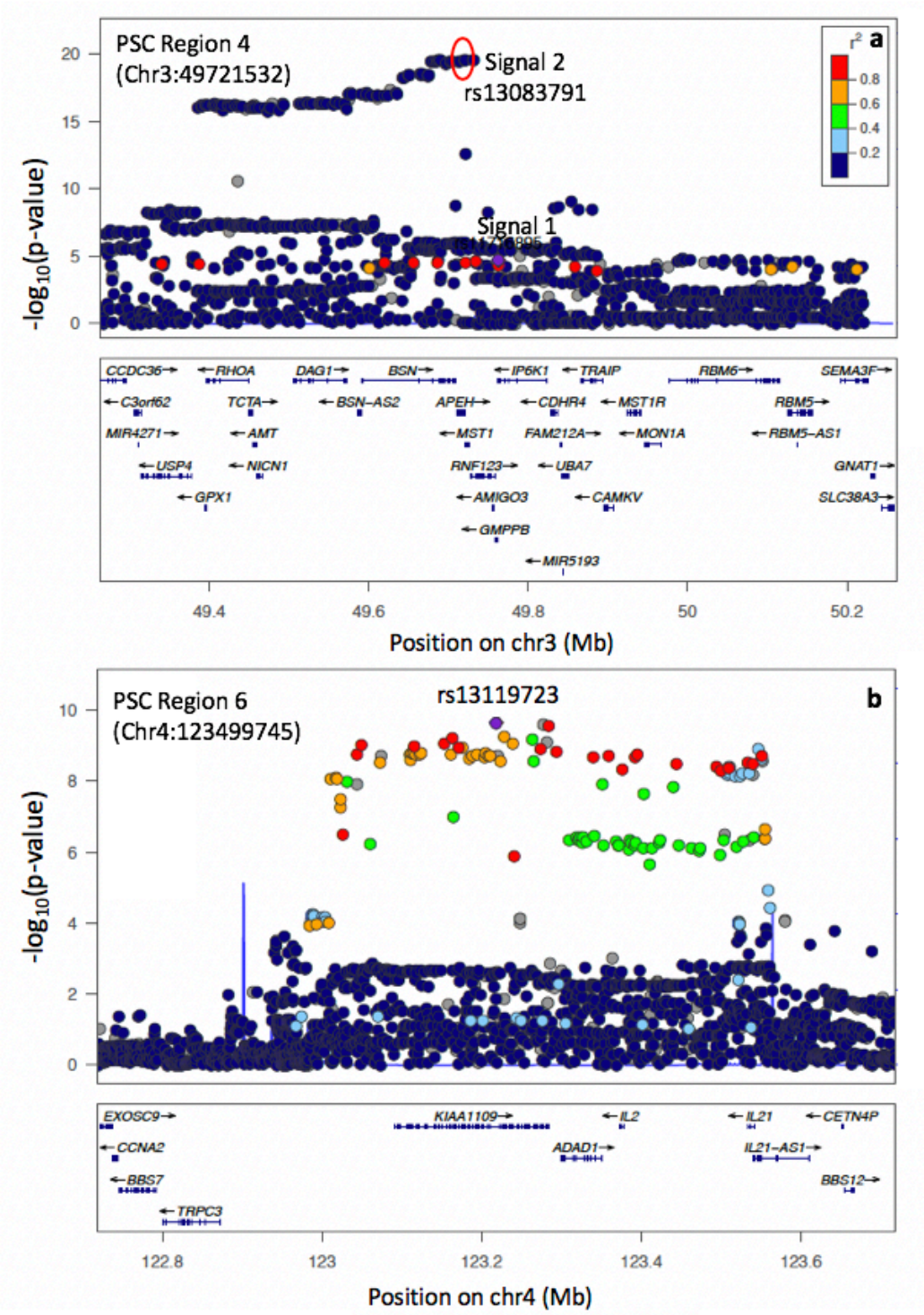

Figure 2.6: Regional association plots for PSC risk loci not well resolved with fine-mapping.

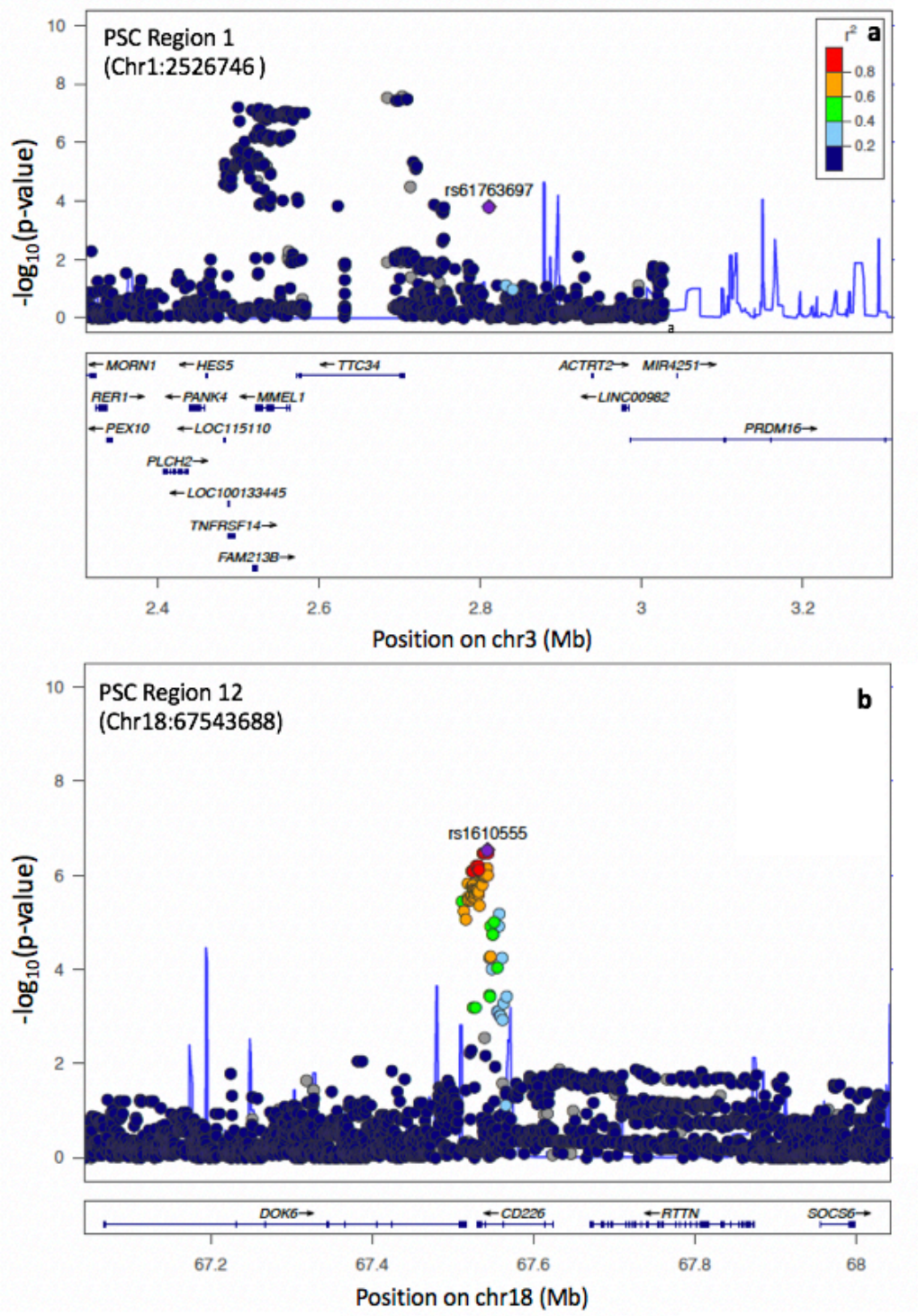

Figure 2.7: Regional association plots for PSC risk loci not well resolved with fine-mapping.

2.5 Discussion

In this chapter I perform the first fine-mapping analysis of risk loci associated with PSC. Using an established Bayesian fine-mapping method I was able to fine-map five of these risk

loci to a credible set containing a variant with >50% PP of causality, two of which were fine-mapped to a single causal variant with >95% PP of causality. Of these five variants, one is enriched for a significant protein-coding change, one is predicted to cause direct disruption to a splice site and three overlap tissue-specific epigenetic marks in PSC-relevant cell- and tissue-types. Stringent prior quality control and filtering of GWAS data means that for common variant associations, the results of this fine-mapping analysis are likely to be robust. This is supported by the finding that PSC credible sets are significantly enriched for variants that overlap enhancer or promoter regions in PSC-relevant tissues and cell-types. This analysis however, illustrates some of the challenges associated with fine-mapping. Only seven loci were mapped to credible sets containing ≤ 10 credible causal variants and for only 5 of these loci was is possible to identify a single causal variant as a promising candidate with >50% PP. For several loci, the presence of large credible sets with multiple plausible causal variants, each with low PPs of causality means that functional annotation is essential for prioritisation. The generation of precise annotation maps in disease-relevant tissues will therefore be crucial to our ability to further interpret these risk loci. Nevertheless, the identification of causal variants for even just a few loci remains a valuable outcome.

Precise fine-mapping should frequently point to the same variant in different diseases with shared risk loci. IBD remains the disease which shares the most genetic architecture with PSC. Of the 15 PSC risk loci fine-mapped within this study, it has been previously reported that five loci; Chr3:49721532 *(MST1)*, Chr4:123499745 *(Il2-Il21)*, Chr12:111884608 *(SH2B3)*, Chr18:67543688 *(CD226)* and Chr21:40466744 *(ETS2)*, demonstrate strong evidence for a shared causal variant with IBD [42]. Three of these loci, Chr3:49721532 *(MST1)*, Chr4:123499745 *(Il2-Il21)* and Chr21:40466744 *(ETS2)*, have been the subject of fine-mapping in IBD [56, 60]. Huang, Fang, Jostins *et al* resolved the *ETS2* locus to a credible set of ten variants, with a 39% PP attributed to the most probable variant, rs9977672. In PSC, I fine-mapped this region to a ten variant credible set, eight of which overlap with the IBD credible set for this region, however prioritising a different variant, rs4817988 at position 40468838, as causal with 58% certainty. The two non-overlapping variants within the IBD credible set, one of which is the most probable IBD fine-mapped variant, are both present within the PSC dataset. It is likely that for this region, the same variant is causal is both PSC and IBD, supporting a higher prior in any future fine-mapping studies and consideration of these two additional IBD credible variants in any future functional follow-up studies. Fine-mapping in IBD was however unable to resolve the Chr3:49721532 *MST1* and Chr4:123499745 *Il2-Il21* loci, with credible sets containing 437 and 29 variants respectively, a likely consequence of the extended, complex patterns of LD observed within these regions. This is an important negative finding, as for those regions not well resolved by fine-mapping in PSC, it has been suggested that the future

use of larger GWAS sample sizes will enable the statistical resolution of more risk loci, due to the ability to better distinguish subtle difference in LD between tightly correlated variants. Sample sizes in IBD GWAS dwarf those of PSC, with the numbers of subjects now approaching 60,000. However despite this, for these high LD regions, fine-mapping was no more successful in the larger samples sizes of IBD. One remedy to this might be to leverage LD from other ethnicities by undertaking GWAS in populations with different LD structure to improve fine-mapping resolution [113]. However, to date, GWAS in PSC have only included individuals of European ancestry, and the number of non-European individuals included in IBD GWAS is comparatively small.

Several of the PSC risk loci in this study have been reported as risk loci and fine-mapped in RhA and T1DM. Similar to IBD, this provides an important means of verifying the precision of these PSC fine-mapping results and where PSC fine-mapping has not resolved a locus to a small credible set, it provides the opportunity to review fine-mapping results from IMDs with larger sample sizes and thus greater power to differentiate between highly correlated SNPs. However, whilst we know there is significant sharing of risk loci between IMDs, to date, there have been no studies that determine shared risk loci between PSC and other IMDs, outside of IBD. In order to conclusively prove that a risk locus is shared between two traits, and thus the results from the fine-mapping of one trait are applicable to both traits, some form of statistical analysis is required. Colocalisation is a statistical means of assessing the probability that the signal observed in two traits e.g. a PSC risk locus and a T1DM risk locus, is driven by the same causal variant. Whilst colocalisation does not define which is the true causal variant for each colocalising trait, in combination with fine-mapping it provides a powerful means of determining shared genetic architecture and resolving causal variants. This is an analysis explored in the following chapter of this thesis.

An important step in using genetic risk loci to further our biological understanding of disease causation is to identify the genes impacted. When the fine-mapped variant falls within a coding region of the genome, this can be relatively straightforward. For example, fine-mapping of PSC region 10 (Chr12:111884608) identified rs3184504, a missense variant located in exon three of *SH2B3*, as the most probable causal variant with 99% certainty. *SH2B3* is an interesting gene in the pathogenesis of PSC, as it is a negative regulator of T-cell activation, TNF production, and Janus kinase (JAK) 2 and 3 signalling, with several studies exploring the allele-specific effects of this SNP on immune- and inflammatory-response and subsequent risk of IMD. However, the majority of genetic associations with IMD fall within non-coding genomic regions, and PSC is no exception. Of the 15 PC-risk loci fine-mapped in this study, only two loci were fine-mapped to variants within a coding region. For the remaining 13 loci, identifying the precise genes impacted by the non-coding variants, and the direction of effect on gene expression remains challenging. For example,

fine-mapping and annotation of the Chr19:47205707 *PRKD2* locus supported rs313839 at position 47221557 as the most probable causal variant. The PSC risk increasing allele at this position is predicted to cause direct disruption to a binding site for multiple transcription factors. However, identifying the gene affected by this mutation, and whether the PSC risk increasing allele results in increased or decreased expression of that gene is not possible from fine-mapping and annotation alone.

One means of identifying the genes affected by the many non-coding risk loci is via colocalisation analysis with variants that exert a quantitative effect upon gene expression (eQTL). Colocalisation can be performed between any two types of traits, for example two disease risk loci, or a disease risk locus and an eQTL locus. For example, the Chr19:47205707 *PRKD2* locus is also a T1DM locus, which has been shown to colocalise with an eQTL in monocytes [182]. It is likely that the gene affected by this same PSC locus is also *PRKD2*, however colocalisation is a statistical means of measuring the probability that this is true. Colocalisation with eQTLs allows us to identify the genes impacted by non-coding risk loci, in addition to identifying the direction of effect a particular disease risk allele has upon downstream gene expression. Following on from fine-mapping, an important next step to infer biological understanding from genetic risk loci in PSC is therefore to identify the genes impacted, by colocalisation with eQTLs mapped in relevant cell types, an analysis which is presented in the next chapter of this thesis.

Fine-mapping of genomic regions associated with disease risk is an important step in understanding the biological mechanisms via which risk variants exert their effect to cause disease. Through fine-mapping, we can filter the often many hundreds of potential causal variants within a locus to a single variant or set of variants responsible for the observed association. In many cases, functional annotation of these credible sets gives us insight into the mechanisms via which they alter gene expression and thus the biological pathways that may be important in disease causation. Colocalisation with eQTLs measured in disease-relevant cell-types and tissues is an important next step for identifying those genes, cell-types and biological pathways affected by disease risk loci, and will bring us one step closer to understanding the causal biology of PSC.