Chapter 3. Discovery of primary sclerosing cholangitis risk loci and the genetic relationship with inflammatory bowel disease

3.1 Introduction

Primary sclerosing cholangitis (PSC) is a severe liver disease of unknown etiology that results in the fibrotic destruction of the bile ducts (Aadland *et al.,* 1987; Broome *et al.,* 1996; Farrant *et al.,* 1991). The pathogenesis of PSC is poorly understood, and due to the lack of effective medical therapy, PSC remains a leading indicator for liver transplantation in Northern Europe and the US (Karlsen *et al.,* 2010b), despite the relatively low prevalence (~10/100,000). Affected individuals are diagnosed at a median age of 30-40 years and suffer from an increased frequency of inflammatory bowel disease (IBD) (60-80%) (Karlsen and Kaser, 2011; Karlsen *et al.,* 2010b) and autoimmune diseases (25%) (Saarinen *et al.,* 2000). Conversely, approximately only 5% of patients with IBD develop PSC (Karlsen and Kaser, 2011; Karlsen *et al.,* 2010b). A 9-39-fold sibling recurrence risk indicates a strong genetic component to PSC risk (Bergquist *et al.,* 2008). In addition to multiple strong associations within the human leukocyte antigen (HLA) complex, recent association studies have identified genome-wide significant loci at 1p36 (*MMEL1*/*TNFRSF14*), 2q13 (*BCL2L11*), 2q37 (*GPR35*), 3p21 (*MST1*), 10p15 (*IL2RA*) and 18q21 (*TCF4*) (Ellinghaus *et al.,* 2012; Folseraas *et al.,* 2012; Karlsen *et al.,* 2010a; Melum *et al.,* 2011; Srivastava *et al.,* 2012).

In order to identify additional risk loci associated with PSC risk, 3,789 PSC cases from Europe and North America, along with 25,079 population matched controls, were genotyped on the Immunochip. The IBD status were also available for 3,283 of the PSC cases, and, along with results from a recent GWAS of IBD (Jostins *et al.,* 2012), allowed for powerful cross-phenotype genetic comparisons.

3.1.1 Chapter overview

In this chapter, I discuss the identification of twelve genome-wide significant PSC risk loci outside the HLA region, nine of which are implicated in PSC risk for the first time. Within the HLA region, HLA-allele imputation revealed five independent associations. Due to the high comorbidity with IBD (72% of cases have Crohn's disease (CD), ulcerative colitis (UC) or indeterminate IBD), investigating the shared and unique genetic basis between the two disorders has implications in understanding shared biology and disease classification. I investigated this sharing at PSC risk loci, and considered in aggregate IBD risk and variants genome-wide, showing the presence of both overlapping and distinct genetic architectures for PSC and IBD.

3.1.2 Contributions

The study design was conceived by the International PSC Genetics Study Group (IPSCSG). Cases and controls were ascertained through the IPSCSG and the International IBD Genetics Consortium (IIBDGC). Genotyping was performed at various centres described in section 3.2.1 and the Supplementary Note of Liu et al. (2013). GRAIL analysis was performed by Trine Folseraas. Quality control on unpublished GWAS data was performed by Sun-Gou Ji. All other analyses were performed by myself.

3.2 Methods

3.2.1 Samples, DNA extraction and genotyping

Recruitment of PSC cases was performed in 14 countries in Europe and North America (Table 3.1). Diagnosis of PSC was based on standard clinical,

biochemical, cholangiographic and histological criteria with exclusion of secondary causes of sclerosing cholangitis (Chapman *et al.,* 1980). Controls were recruited from blood donors, population-based studies as part of this study, or via the International Immunochip Consortium. See Supplementary Note of Liu *et al.* (2013) for details.

	Controls	PSC cases	Total
Scandinavia	4,324	917	5,241
North Central Europe	9,438	1,136	10,574
Southern Europe	580	115	695
UK	8,663	1,033	9,696
North America	2,074	588	2,662
Total	25,079	3,789	28,868

Table 3.1. Post-QC patient and control panels. PSC cases and controls in the study sorted by broad geographic panels (based on participating centre information, not genotypes). Scandinavia: Finland, Norway, Sweden; North Central Europe: Belgium, Germany, The Netherlands, Poland; Southern Europe: France, Greece, Italy, Spain; UK: United Kingdom; North America: Canada, USA.

DNA was extracted from whole blood, transformed lymphocytes or liver tissue using commercially available kits or an in-house out-salting method. DNA samples were genotyped using the Immunochip according to Illumina protocols. The NCBI build 36 (hg18) reference was used and normalised probe intensities were extracted for all samples passing standard laboratory quality control thresholds. All genotypes were called specifically for this study using optiCall (Shah *et al.,* 2012), but separately across each genotyping batch. Genotypes with a posterior probability lower than 0.7 were defined as unknown. All PSC cases were genotyped at the Institute of Clinical Molecular Biology in Kiel, Germany, or at the department of Genetics, University of Groningen and University Medical Centre Groningen, The Netherlands.

3.2.2 Quality control

SNPs with a call rate < 80% were removed prior to sample QC (n = 235). Per individual genotype call rate and heterozygosity rate were calculated using PLINK (Purcell *et al.,* 2007) and outlying samples were identified using Aberrant (Figure 3.1) (Bellenguez *et al.,* 2012), which identifies outliers from otherwise Gaussian distributions. A set of 20,837 LD-pruned (r^2 < 0.1) SNPs with minor

allele frequency > 10% present in both the Immunochip and the Illumina Omni2.5-8 array used in the 1000 Genomes Project (Genomes Project *et al.,* 2012) were used to estimate identity by descent and ancestry. For each pair of individuals with estimated identity by descent ≥ 0.9, the sample with the lower call rate was removed (unless case/control status was discordant between the pair, in which case both samples were removed, n = 92). Related individuals (0.1875 < identity by descent < 0.9) remained in the analysis to maximize power because the mixed model association analysis can correctly account for the relatedness between individuals. Principal components analysis was performed using SMARTPCA (Patterson *et al.,* 2006). Principal components were defined using population samples from the 1000 Genomes Project genotyped using the Illumina Omni2.5-8 genotyping array and then projected into PSC cases and controls, with non-European outliers identified using Aberrant and removed (Figure 3.2). Following sample QC, 3,789 PSC cases and 25,079 remained. SNPs with a minor allele frequency less than 0.1%, Hardy-Weinberg equilibrium $P < 10^{-5}$ in controls, call rate lower than 98%, or significant differential missing data rate between cases and controls ($P < 10^{-5}$) were excluded. After completion of marker QC, 131,220 SNPs were available for analysis – further reduced to 130,422 after cluster plot inspection of nominally associated SNPs. The genomic inflation factor (Devlin *et al.,* 1997) was calculated using 2,544 "null" SNPs. These SNPs were included on the Immunochip as part of replication panels for bipolar disease and other non-immune-related studies.

3.2.3 Imputation

Using 85,747 post-QC SNPs located in the Immunochip high density regions, additional genotypes were imputed using IMPUTE2 with the 1000 Genomes Phase 1 (March, 2012) reference panel of 1,092 individuals (Genomes Project *et al.,* 2012) and 744,740 SNPs. Imputation was performed separately across ten batches, with the case:control and country of origin ratios constant across batches. SNPs with a posterior probability less than 0.9 and those with differential missingness ($P < 10^{-5}$) between the 10 batches were removed, as

were SNPs failing the exclusion thresholds used for genotyped SNP QC. After imputation, a total of 208,852 SNPs were available for analysis.

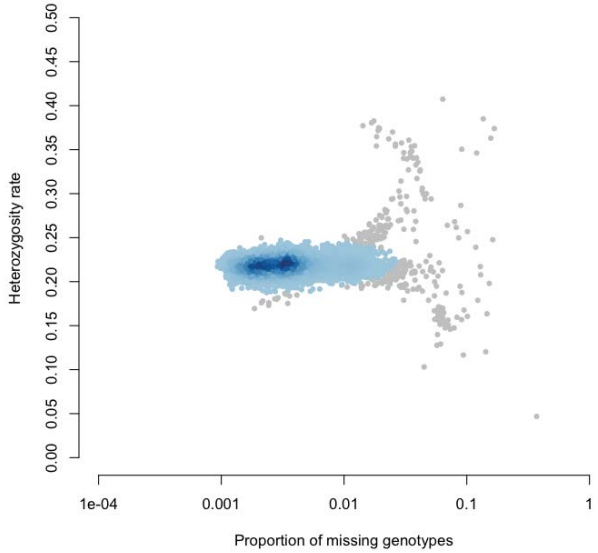

Figure 3.1. Heterozygosity rate and proportion of missing genotypes for PSC cases and controls. The grey points represent outlying individuals. Heterozygosity proportions and missingness were calculated using PLINK (Purcell et al., 2007). Outliers were detected using Aberrant (Bellenguez et al., 2012).

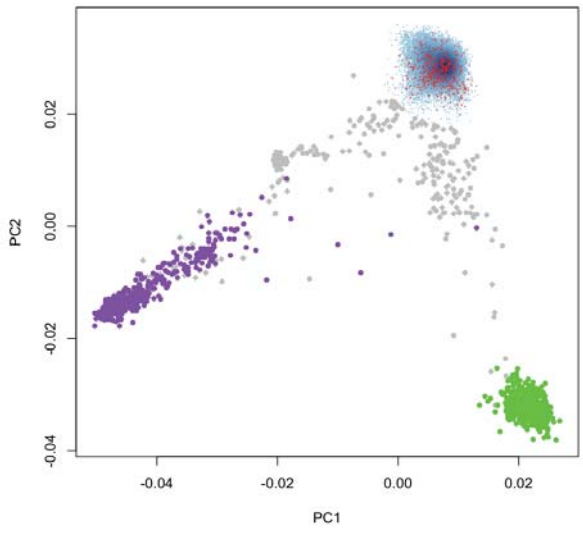

Figure 3.2. Principal components analysis of PSC cases and controls with 1000 Genomes Omni2.5-8 data. The red, purple and green points represent 1000 Genomes CEU (Utah residents with Northern and Western European ancestry), YRI (African) and CHB+JPT (Han Chinese and Japanese) populations respectively. The blue points represent the included PSC cases

and controls, overlapping the CEU population, with the grey points showing those who were identified as ancestry outliers (and therefore excluded). The principal components were generated using 20,837 common (MAF>0.10) SNPs overlapping between the Immunochip (this study) and the Omni2.5-8 array.

3.2.4 Association analysis

Case-control association tests were performed using a linear mixed model as implemented in MMM (Pirinen *et al.,* 2012). A covariance matrix, *R*, of a random effects component was included in the model to explicitly account for confounding due to population stratification and cryptic relatedness between individuals. This method has been shown to better control for population stratification than correction for principal components or meta-analyses of matched subgroups of cases and controls (Korte *et al.,* 2012; Sawcer *et al.,* 2011). *R* is a symmetric n×n matrix with each entry representing the relative sharing of alleles between two individuals compared to the average in the sample, and is typically estimated using genome-wide SNP data. To avoid biases in the estimation of *R* due to the design of the Immunochip, SNPs were first pruned for LD ($r^2 < 0.1$). Of the remaining SNPs, I then removed those that lie in the HLA region or have a minor allele frequency < 10%. Finally, I excluded SNPs that showed modest association ($P < 0.005$) with PSC in a linear regression model fitting the first 10 principal components as covariates. A total of 17,260 SNPs were used to estimate *R*. The following parameters were used in MMM: logOR = 2 (more accurate when genotypes are coded 0,1,2 and no predictors other than genotyes), mean_center = 1 (genotypes are mean-centred), impute_missing = 1 (missing genotypes are set to mean of non-missing genotypes), min_d = 0.1 (lower bound for accepted eigenvalues of *R*).

Due to computational limitations, I estimated the *R* matrix and performed all association analyses separately for UK (n = 9,696) and non-UK (n = 19,172) samples, and then combined the results using a fixed-effects (inverse-variance weighting) meta-analysis. This reduced the λ_{GC} (estimated using the 2,544 "null" SNPS and using the first 10 PCs as covariates) from 1.13 to 1.02 (Figure 3.3), suggesting population stratification was well-controlled for. Stepwise conditional regression was used to identify possible independent associations at

genome-wide significant loci. SNP×SNP interactions between all pairs of genome-wide significant SNPs were tested using the PLINK --epistasis command. Signal intensity plots of all non-HLA loci with association $P < 5\times10^{-6}$ were visually inspected using Evoker (Morris *et al.,* 2010). SNPs that clustered poorly were removed (N = 800).

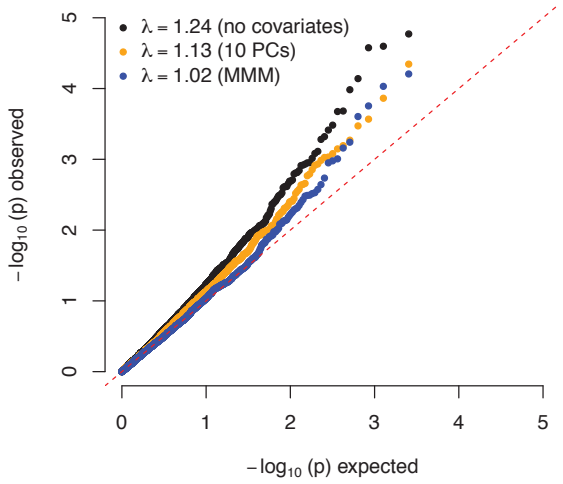

Figure 3.3. Quantile-quantile plots and genomic inflation factors of observed vs. expected P-values. Association tests were compared for logistic regression with no covariates, logistic regression with the first 10 principal components as covariates, and a linear mixed model implemented in MMM (Pirinen et al., 2012). Tests were performed on 2,544 "null" SNPs with no evidence for association with immune-related phenotypes. The dashed red line is y = x.

3.2.5 Functional annotation of risk loci

Gene regulatory elements from the Encyclopedia of DNA elements (ENCODE) and coding SNPs were annotated using HaploReg (Ward and Kellis, 2012). For each risk locus, SNPs in high linkage disequilibrium ($r^2 > 0.8$) with the most significantly associated SNP were assessed as to whether they lie within regions with promoter and enhancer marks, DNAse-I hypersensitivity, protein binding or regulatory motifs in one or more of 147 cell types. Expression quantitative trait loci (eQTLs) were collated from the University of Chicago eQTL Browser (http://eqtl.uchicago.edu/cgi-bin/gbrowse/eqtl/).

3.2.6 GRAIL and DAPPLE analyses

To assess the functional relationship among established genomic PSC risk regions, we performed a GRAIL pathway analysis. GRAIL is a statistical tool that uses text mining of published abstracts in the PubMed database to identify and quantify functional similarity among genes within disease-associated regions (Raychaudhuri *et al.,* 2009). The output GRAIL score is a significance score, Ptext, which is adjusted for multiple hypothesis testing. Sixteen PSC risk loci (7 known and 9 novel) were used as input for this analysis.

Similarly, DAPPLE assesses functional similarity through constructing networks of protein-protein interactions (Rossin *et al.,* 2011). Gene connectivity is assessed based on the number of direct and indirect (via other proteins) connections and a permuted P-value is calculated. The 16 PSC risk loci were used as input into DAPPLE. Genes with $P < 0.05$ were listed as causal candidates.

3.2.7 HLA imputation and association analysis

Imputation of HLA class I and II genes was performed using HLA*IMPv2 (Dilthey *et al.,* 2011; Leslie *et al.,* 2008). The imputation reference panel includes ~2,500 individuals of European ancestry with both genotype and classical HLA-allele type data. Cluster plots for all SNPs contributing to the imputation of HLA types were manually inspected and poorly clustered SNPs were removed. Case-control association was performed on HLA allele posterior probabilities using the mixed model framework described previously. Stepwise conditional regression was used to determine independent HLA association signals.

3.2.8 Heritability explained

The proportion of variance explained by the genome-wide significant SNPs and HLA alleles was calculated using a disease liability threshold model (Falconer and Mackay, 1996; So *et al.,* 2011) assuming a disease prevalence of 10/100,000 and multiplicative disease risk.

3.2.9 Prediction of PSC using IBD risk loci

Odds ratios (ORs) for Crohn's disease and ulcerative colitis in 163 IBD-associated SNPs were obtained from Jostins *et al.* (2012). I used the R package Mangrove (http://cran.r-project.org/web/packages/Mangrove) to generate risk scores and estimate each individual's probability of developing PSC among the 3,789 PSC cases and 25,079 controls assuming additive risk (log-additive OR). The performance of the predictor using either Crohn's disease or ulcerative colitis ORs was assessed by constructing a receiver operating characteristic (ROC) curve, which shows the proportion of true and false positives at each probability threshold. The area under the curve (AUC) was calculated to compare the predictive power of the ulcerative colitis and Crohn's disease ORs.

The DeLong method was used to test if the AUC using ulcerative colitis ORs was significantly different to the AUC using Crohn's disease ORs (DeLong *et al.,* 1988). The method is a non-parametric approach for test the alternative hypothesis that two (or more) AUCs estimated from different sets of predictors in the same samples are significantly different. As the AUC is equivalent to the Mann Whitney U statistic for comparing the distribution values from two samples, variances of correlated U statistics can be estimated using the approach of Sen (1960). The method is equivalent to a jackknife resampling approach for estimating the variance of the AUC (DeLong *et al.*, 1988).

3.2.10 Genetic correlation between PSC and IBD

Genome-wide SNP data were available for 5,322 Crohn's disease cases, 6,307 ulcerative colitis cases and 12,164 population matched controls genotyped as part of previous GWAS meta-analyses (Anderson *et al.,* 2011a; Franke *et al.,* 2010; Jostins *et al.,* 2012). Genome-wide data from an ongoing GWAS for PSC (2,871 cases and 12,019 controls) were also obtained (Sun-Gou Ji, personal communication). All datasets were obtained post-QC. The IBD dataset was imputed using the HapMap phase 2+3 reference panel, while the PSC dataset was imputed using a combined 1000 Genomes Phase I plus UK10K reference panel. Additional QC included removing 35 cases 3803 controls from the IBD dataset

that were duplicated or related with individuals in the PSC dataset using PLINK (pi_hat > 0.1). SNPs with a missingness rate of greater than 2% in the combined data were also removed. In total, 721,733 autosomal SNPs that overlap the two datasets remained. The top 20 Principal components estimated from the 1000 Genomes Phase I individuals were projected onto all IBD and PSC cases and controls.

The proportion of genetic variation (as tagged by common genome-wide SNPs) that is shared between PSC and IBD was estimated using the bivariate linear mixed-effects model implemented in GCTA (Lee *et al.,* 2012). The method uses genome-wide SNPs to estimate genetic similarities between pairs of individuals, and uses bivariate restricted maximum likelihood to estimate covariance components (r_G) of the linear mixed model. In all, each of four PSC subphenotypes (all PSC cases, PSC cases with UC, PSC cases with CD and PSC cases with no IBD) were tested against CD and UC. To test whether r_G is significantly different from 0 (i.e. there is no genetic overlap between the two phenotypes), r_G was fixed at 0 and a likelihood ratio test comparing this constrained model and the unconstrained model was applied.

3.3 Results and discussion

Following quality control and imputation, 208,852 SNPs from 3,789 cases and 25,079 population controls were available for analysis, of which 80,183 SNPs located in the Immunochip high density regions were imputed using the 1000 Genomes reference panel. Case-control association testing was performed using a linear mixed model as implemented in MMM to minimise the effect of population stratification (λ_{GC} = 1.02, estimated using 2,544 "null" SNPs).

3.3.1 Locus discovery

Twelve non-HLA genome-wide significant ($P < 5\times10^{-8}$) PSC susceptibility loci were identified, nine of which were implicated in PSC for the first time (Table 3.2, Figure 3.4, Figure 3.5). The most associated SNP within each locus was a common variant (all risk allele frequencies >0.18) of moderate effect (ORs

between 1.15 and 1.4) (Table 3.2). Genotype imputation and stepwise conditional regressions within each locus did not identify additional independent genome-wide significant signals, nor did genotype-genotype or sex-genotype interaction analyses.

For seven of the nine novel loci, the most significantly associated SNP in the locus was the same SNP or was in strong linkage disequilibrium (LD; $r^2 > 0.8$) with the original association reports for another disease (Table 3.3). The two exceptions were 11q23, where only independent disease associations ($r^2 < 0.01$) have so far been reported for colorectal cancer (Peters *et al.,* 2012), and 6q15, where the most significantly associated PSC variant, rs56258221 (OR = 1.23, P = 8.36×10^{-12}), is in low-to-moderate LD with the previously reported *BACH2* variants in Crohn's disease ($r^2 = 0.23$) and type 1 diabetes ($r^2 = 0.12$).

Chr	SNP[a]	RA[b]	RAF cases[c]	RAF controls[c]	*P*-value	OR (95%CI)	LD region[d] (Kb)	RefSeq genes in LD region	Notable nearby gene(s)[e]	Functional annotation[f]
1p36	rs3748816	A	0.698	0.656	7.41×10^{-12}	1.21 (1.14-1.27)	2,398-2,775	9	*MMEL1, TNFRSF14*	eQTL,MS, OC, PB, HM
2q33	**rs7426056**	A	0.277	0.229	1.89×10^{-20}	1.3 (1.23-1.37)	204,155-204,397	1	*CD28*	HM, OC
3p21	rs3197999	A	0.352	0.285	2.45×10^{-26}	1.33 (1.26-1.4)	48,388-51,358	90	*MST1*	eQTL,MS, OC, PB HM
4q27	**rs13140464**	C	0.871	0.836	8.87×10^{-13}	1.3 (1.21-1.4)	123,204-123,784	4	*IL2, IL21*	OC, PB
6q15	**rs56258221**	G	0.213	0.183	8.36×10^{-12}	1.23 (1.16-1.31)	90,967-91,150	1	*BACH2*	OC, PB
10p15	rs4147359	A	0.401	0.349	8.19×10^{-17}	1.24 (1.18-1.3)	6,070-6,206	2	*IL2RA*	PB
11q23	**rs7937682**	G	0.298	0.265	3.17×10^{-09}	1.17 (1.11-1.24)	110,824-111,492	19	*SIK2*	OC, PB, HM
12q13	**rs11168249**	G	0.506	0.466	5.49×10^{-09}	1.15 (1.1-1.21)	46,442-46,534	3	*HDAC7*	OC, PB, HM
12q24	**rs3184504**	A	0.527	0.488	5.91×10^{-11}	1.18 (1.12-1.24)	110,186-111,512	16	*SH2B3, ATXN2*	MS, OC, HM
18q22	**rs1788097**	A	0.518	0.483	3.06×10^{-08}	1.15 (1.1-1.21)	65,633-65,721	2	*CD226*	MS, OC, PB, HM
19q13	**rs60652743**	A	0.864	0.836	6.51×10^{-10}	1.25 (1.16-1.34)	51,850-51,998	6	*PRKD2, STRN4*	OC, PB, HM
21q22	**rs2836883**	G	0.777	0.728	3.19×10^{-17}	1.28 (1.21-1.36)	39,374-39,404	-	*PSMG1*	OC, PB, HM

Table 3.2. Association results of twelve non-HLA genome-wide significant risk loci for PSC. [a]SNPs from novel PSC-associated loci are shown in bold. [b]Risk increasing allele. [c]Risk allele frequency. [d]LD regions around lead SNPs were calculated by extending in both directions a distance of 0.1 centimorgans as defined by the HapMap recombination map. [e]Select candidate gene(s) within same LD region as the associated SNPs. [f]Denotes if there are SNPs with r2>0.8 with the hit SNP that have functional annotations: eQTL: expression quantitative trait locus, HM: overlaps a region of histone modification MS: missense mutation; OC: open chromatin; PB: protein binding.

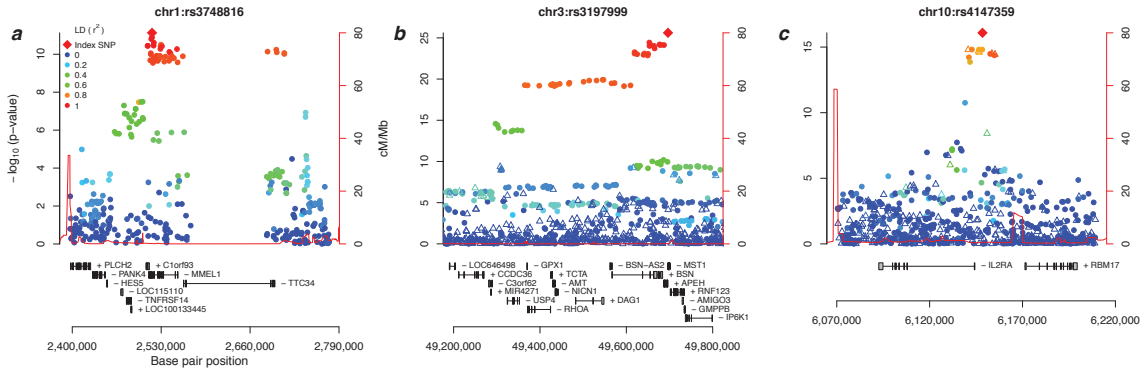

Figure 3.4. Regional association plots for genome-wide significant associations at previously established PSC risk loci. Filled-in circles are directly genotyped and hollow-triangles are imputed SNPs. The colour of the marker (see legend in panel a) illustrates the linkage disequilibrium between the most associated SNP and others in the locus.

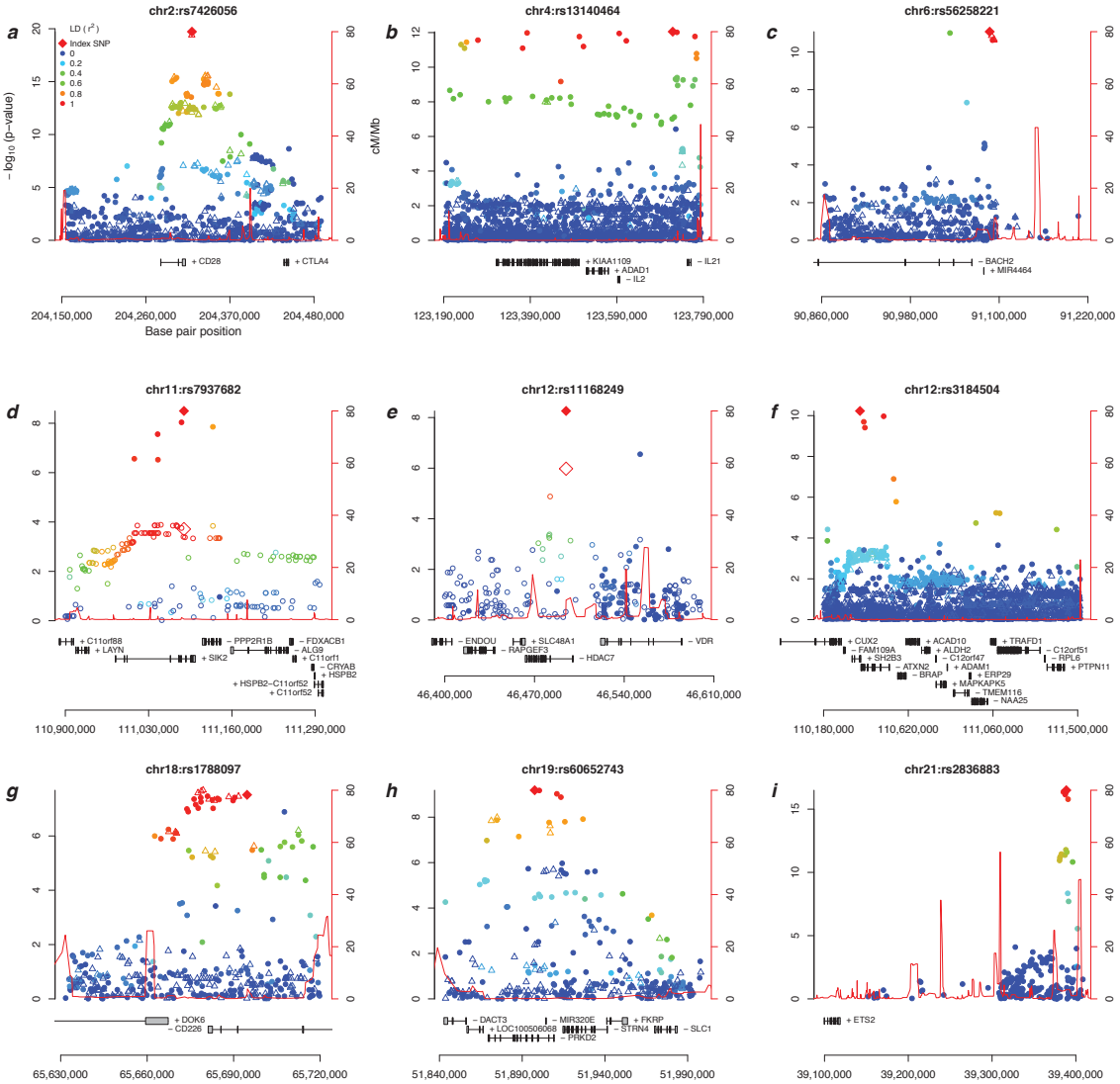

Figure 3.5. Regional association plots of nine newly associated PSC risk loci. In panels d and e, the most associated SNPs are located outside Immunochip fine-mapping regions. Association signals from the discovery panel of the largest PSC GWAS to date are shown as hollow circles and the most associated SNP as a hollow diamond (genotyped and imputed to HapMap release 22 SNPs, cases overlap with the current study).

Locus	SNP	Same signal	Different signal
1p36	rs3748816	CeD,RA,UC	
2q33	rs7426056	CeD	AA,CeD,CHD,GD,Ht,MI,RA,T1D
3p21	rs3197999	**CD**,**UC**	
4q27	rs13140464	CD,CeD,RA,UC	AA,T1D
6q15	rs56258221		CD,CeD,MS,T1D,Vi
10p15	rs4147359	AA,MS,RA	T1D,Vi
11q23	rs7937682		Colorectal cancer
12q13	rs11168249	**CD**,**UC**	
12q24	rs3184504	**BP**,**CeD**,Ch,CKD,**EC**,He,Hg,**Ht**,PBC,RVC,**T1D**	
18q22	rs1788097	T1D	
19q13	rs60652743	T1D	CLL
21q22	rs2836883	AS,**CD**,**UC**	

Table 3.3. Association of genome-wide significant PSC risk loci with other diseases. A SNP association for another disease is defined to be the same signal if this SNP is in high LD (r2 > 0.8) with one or more genome-wide significant PSC associated SNPs in the locus. Diseases highlighted in bold denote associations where the lead SNP is the same in both diseases. Previously associated SNPs were obtained from the Catalog of Published Genome-wide Association Studies (http://www.genome.gov/gwastudies). SNPs reported in other Immunochip experiments were available for CeD, CD, UC and PBC (Jostins et al., 2012; Liu et al., 2012; Trynka et al., 2011b). AA: Alopecia areata, AS: Ankylosing spondylitis, BP: Blood pressure, CD: Crohn's disease, CeD: Celiac disease, Ch: Cholesterol, CHD: Coronary heart disease, CKD: Chronic kidney disease, CLL: Chronic lymphocytic leukaemia, EC: Eosinophil counts, GD: Grave's disease, He: Haematocrit, Hg: Haemoglobin, HT: Hypothyroidism, MI: Myocardial infarction, MS: Multiple sclerosis, PBC: Primary biliary cirrhosis, RA: Rheumatoid arthritis, RVC: Retinal vascular calibre, T1D: Type 1 diabetes, UC: Ulcerative colitis, Vi: Vitiligo

3.3.2 Associations at previously reported non-HLA PSC risk loci

In the main association analysis, three out of six previously reported genome-wide significant ($P < 5\times10^{-8}$) non-HLA risk loci (rs3748816 at 1p36, rs3197999 at 3p21 and rs4147359 at 10p15) (Folseraas *et al.,* 2012; Melum *et al.,* 2011; Srivastava *et al.,* 2012) were genome-wide significant (Table 3.2, Figure 3.4). In a fourth locus, the genome-wide significant SNPs from the previous study (rs3749171 and rs4676410 at 2q37) (Ellinghaus *et al.,* 2012) failed genotyping in one of the genotyping batches and was excluded. However, the peak SNP in this dataset (rs2011743) was in moderate linkage disequilibrium (r2 = 0.29) with the lead SNP from the previous study (rs3749171) and showed nominal association, ($P = 5.0\times10^{-5}$, OR = 1.17, 95% CI 1.08-1.26). The previously reported PSC associations at 2q13 and 18q21 were not covered on the Immunochip (Ellinghaus *et al.,* 2012; Melum *et al.,* 2011).

3.3.3 Candidate gene prioritisation

To prioritize candidate genes within the non-HLA genome-wide significant loci, I searched for nonsynonymous coding and known eQTLs among the SNPs in high LD ($r^2 > 0.8$) with the most associated SNPs. Risk loci were als functionally annotated using data from the ENCODE project (Ward and Kellis, 2012). Networks were constructed based on known protein-protein interactions (DAPPLE) (Rossin *et al.,* 2011) and the text mining published literature (GRAIL) (Raychaudhuri *et al.,* 2009) to identify potentially important disease-relevant genes. For four of the 12 genome-wide significant loci, the same gene (*MMEL1*, *MST1*, *SH2B3*, and *CD226*) was annotated by more than one method (Table 3.4), suggesting these as candidates for further investigation at these loci.

Two newly associated loci are located outside of the Immunochip fine mapping regions (Figure 3.5). At 11q23, the most strongly associated SNP, rs7937682 (OR = 1.17, $P = 3.18\times10^{-9}$), is located in an intron of salt-inducible kinase 2 (*SIK2*), which both influences the expression of interleukin-10 in macrophages and Nur77, an important transcription factor in leukocytes (Hanna *et al.,* 2011). The association at 12q13 is with an intronic SNP (rs11168249, OR = 1.15, $P = 5.49\times10^{-9}$) within the histone deacetylase 7 (*HDAC7*) gene, which has also been associated with IBD (Jostins *et al.,* 2012). *HDAC7* has been implicated in negative selection of T cells in the thymus (Kasler *et al.,* 2011), a key factor in the development of immune tolerance. A role for *HDAC7* in PSC etiology is supported by the novel association at 19q13, where the most associated SNP, rs60652743 (OR = 1.25, $P = 6.51\times10^{-10}$) is located within an intron of serine-threonine protein kinase D2 (*PRKD2*). When T cell receptors of thymocytes are engaged, *PRKD2* phosphorylates *HDAC7*, leading to nuclear exclusion of *HDAC7* and loss of its gene regulatory functions, ultimately resulting in apoptosis and negative selection of immature T cells (Dequiedt *et al.,* 2003; Dequiedt *et al.,* 2005). Interestingly, this negative selection takes place due to a loss of HDAC7-mediated repression of Nur77 (regulated by SIK2) (Clark *et al.,* 2012), linking three novel PSC loci to this pathway.

Locus	SNP	ENCODE	eQTL[b]	Missense[b]	GRAIL[c]	DAPPLE[c]	No. of genes
1p36	rs3748816	P,E,D,PB,RM	*MMEL1*	*MMEL1*			1
2q33	rs7426056	E,D,PB,RM			*CD28*		1
3p21	rs3197999	P,E,D,PB,RM	*USP4*	*BSN,MST1*	*GPX1,MST1*		5
4q27	rs13140464	D,PB,RM			*IL2*		1
6q15	rs56258221	D,PB,RM			*BACH2*		1
10p15	rs4147359	PB,RM			*IL2RA*		1
11q23	rs7937682	P,E,D,PB,RM			*CRYAB,HSPB2*	*SIK2*	3
12q13	rs11168249	E,D,PB,RM			*VDR*		1
12q24	rs3184504	P,E,D,RM		*SH2B3*	*SH2B3,TRAFD1*	*C12orf51*	3
18q22	rs1788097	E,D,PB,RM		*CD226*	*CD226*		1
19q13	rs60652743	P,E,D,PB,RM					0
21q22	rs2836883	E,D,PB,RM			*ETS2*		1

Table 3.4. Candidate functional annotations and genes among genome-wide significant PSC risk loci. [a]SNPs in high LD ($r^2 > 0.8$)with the lead SNP that overlap one or more of the following ENCODE annotations in at least one of 147 cell types identified using HaploReg (Ward and Kellis, 2012). P: promotor histone markers; E: enhancer histone markers; D: DNAse-I hypersensitivity; PB: protein binding; RM: regulatory motifs. [b]SNPs in high LD with the most significantly associated SNP in the locus that are either known eQTLs or missense mutations. [c]Genes implicated by GRAIL, DAPPLE or functional similarity networks that show nominally significant ($P < 0.05$) number of connections.

3.3.4 HLA association

The associations at the HLA complex at 6p21 were refined by imputing HLA haplotypes at *HLA-A*, *HLA-C*, *HLA-B*, *HLA-DRB1*, *HLA-DQB1*, *HLA-DQA* and *HLA-DPB1* (Dilthey *et al.,* 2011). Imputation was highly accurate at 2 digit level for *HLA-B* and *HLA-DRB1*, with >96% and 98% concordance respectively when compared with previous in-house sequencing-based HLA typing data (Karlsen *et al.,* 2007; Melum *et al.,* 2011). The lead SNP in the HLA complex (rs4143332; $P = 6.39\times10^{-249}$) was in perfect linkage disequilibrium with the lead SNP in the previous genome-wide association study (rs3134792, $r^2 = 1.0$) (Melum *et al.,* 2011), and in almost perfect linkage disequilibrium with HLA-B*08:01 ($r^2 = 0.996$ with imputed HLA- B*08:01 in this dataset). HLA-B*08:01 is encoded on the ancestral HLA-B*08:01-DRB1*03:01 haplotype (AH8.1) which is associated with multiple autoimmune diseases (Candore *et al.,* 2002).

Stepwise conditional analysis was performed including both SNP and HLA haplotypes. The SNP rs4143332 (tagging HLA-B*08:01) and a complex HLA class II association signal determined by HLA-DQA1*01:03 and SNPs rs532098, rs1794282 and rs9263964 explain all of the genome-wide significant HLA

association signals in the data (Figure 3.6). Stepwise conditional regression with only HLA alleles showed significant associations with the established PSC haplotypes HLA-B*08:01, HLA-DQA*01:03, HLA-DQA*05:01, DRB1*15:01 and DQA*01:01, confirming previously reported associations with HLA haplotypes in PSC (Table 3.5) (Chapman *et al.,* 1983; Donaldson *et al.,* 1991; Donaldson and Norris, 2002; Wiencke *et al.,* 2007).

The HLA-DRB1*15:01 association overlaps with that of ulcerative colitis (risk increasing) and Crohn's disease (risk decreasing) (Okada *et al.,* 2011; Stokkers *et al.,* 1999). Since imputed genotypes at the class II region were only available for four (HLA-DRB1, HLA-DQB1, HLA-DQA1 and HLA-DPB1) out of 20 loci (Horton *et al.,* 2004), further studies involving direct sequencing of all HLA class II loci along with assessments of their protein structure and peptide binding are required to causally resolve the link between this HLA subregion and PSC development (Hov *et al.,* 2011; Hovhannisyan *et al.,* 2008).

		Per-allele model		Full model	
HLA allele	MAF	OR	P-value	OR	P-value
B*08:01	0.12	2.82	3.70×10^{-246}	2.53	3.79×10^{-80}
DQA*01:03	0.07	2.23	1.20×10^{-100}	3.66	7.43×10^{-167}
DQA*05:01	0.16	2.39	6.00×10^{-175}	1.87	5.41×10^{-36}
DRB1:15:01	0.14	1.04	0.28	1.57	7.41×10^{-35}
DQA*01:01	0.09	0.83	1.20×10^{-6}	1.31	6.60×10^{-15}

Table 3.5. Odds ratio and P-value of independent HLA allele associations with PSC. Five independent HLA allele associations were identified via stepwise conditional analysis. The per-allele model denotes ORs and P-values of each HLA-allele from a univariate model (no covariates), while the full model includes all five HLA alleles as covariates in a multivariate model. Association testing was performed using the linear mixed model implemented in MMM (Pirinen *et al.*, 2012).

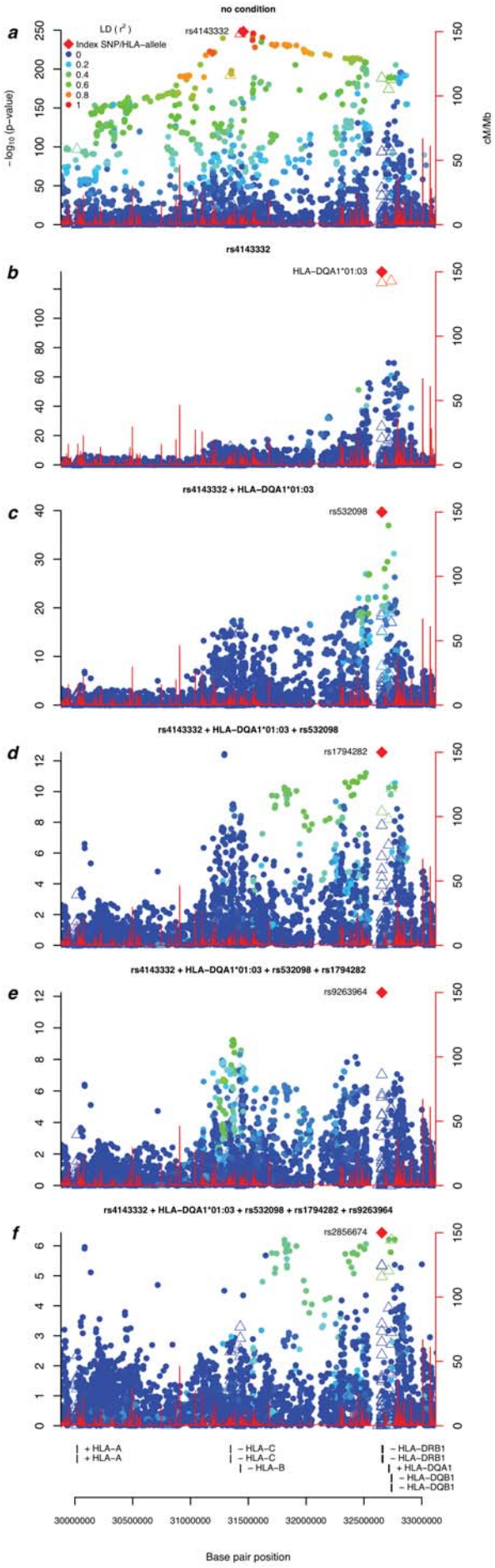

Figure 3.6. Regional association plots from stepwise conditional regression in the HLA complex in PSC. Regional association plots show both SNPs (filled circles and imputed HLA alleles (hollow triangles). Panel a displays associations with no conditioning, showing the peak association with rs4143332, which is in strong linkage disequilibrium with HLA-B*08:01. Panel b shows the association results when conditioned on rs4143332, panel c conditioned on rs4143332 and HLA-DQA1*01:03 and so on. Points are coloured according to linkage disequilibrium with the most strongly associated variant (see panel a for colour legend), which is sown as a filled in red diamond. Recombination rates in the region are shown by the red lines (in cM/MB).

3.3.5 Genetic overlap with IBD

IBD subphenotypes were available for 3,285 of the 3,789 cases (Table 3.6). Although 72% of the PSC patients in this study have a diagnosis of concomitant IBD, only half of the genome-wide significant loci were associated with IBD in the recent International IBD Genetics Consortium (IIBDGC) GWAS meta-analysis (Jostins *et al.,* 2012), despite the greater sample size of that study (~75,000 cases and controls) (Figure 3.7). Three of the six PSC risk alleles with no evidence of association in IBD (*BACH2*, *IL2RA* and *PRKD2*) contain other variants nearby that are associated with IBD. Across the 12 PSC loci, there was greater similarity between the OR estimates for PSC and ulcerative colitis than for PSC and Crohn's disease. Indeed, all but one of the CD/UC ORs for PSC-only risk alleles are > 1, suggesting that some of these may also be IBD risk loci, or that these ORs are partly driven by the small number of IBD cases in Jostins *et al.* who also have PSC.

Subphenotype	N
Crohn's disease	355
Ulcerative colitis	1898
Indeterminate IBD	108
No IBD	922
Unknown	506
	3789

Table 3.6. IBD Subphenotypes among PSC cases.

Significant genetic overlap between PSC and IBD was also observed at the 163 known IBD risk loci. While only six of these loci exceeded genome-wide significance in the PSC association analysis, 123 of the 163 IBD risk loci showed the same direction of effect ($P = 5.07\times10^{-11}$) (Figure 3.8). If the two phenotypes were unrelated, this fraction would be closer to 50%. This positive correlation in the direction of effects was stronger for loci associated with just UC (74% concordance, P = 0.0053) than those only associated with CD (60% concordance, P = 0.1). The greatest concordance was seen for loci that were associated with both CD and UC (80% concordance, $P = 1.1\times10^{-11}$).

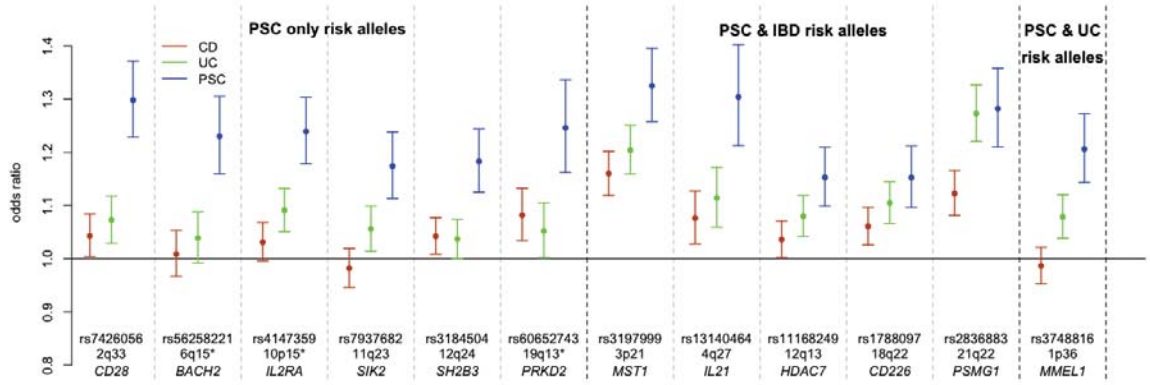

Figure 3.7. Odds ratio comparisons for PSC risk loci in IBD. IBD ORs and designation of loci as UC, CD or both (IBD) were obtained from Jostins et al. (2012). Error bars represent 95% confidence intervals. *The PSC associated alleles at 6q15 (*BACH2*), 10p15 (*IL2RA*) and 19q13 (*PRKD2*) are independent of the reported IBD associations (r2 < 0.3) but are located in the same broad genetic regions as the IBD-associated SNPs.

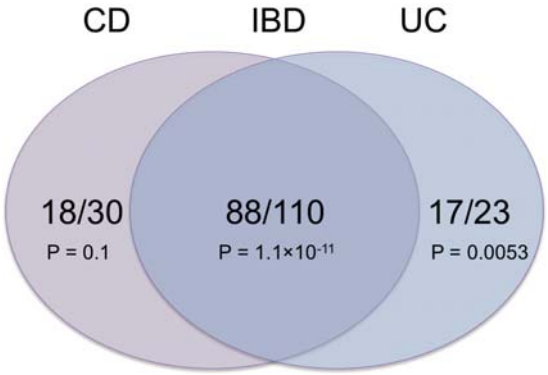

Figure 3.8. Venn diagram of directions of effect in PSC of SNPs associated with either CD, UC or both (IBD). The numbers within each segment denote the number of variants that have the same direction of effect in PSC as CD/UC/IBD over the total number of CD/UC/IBD variants. P-values were obtained from a binomial test (H1: proportion ≠ 0.5).

This trend for a greater genetic similarity between PSC and UC than CD also extends to the aggregate effect sizes at these loci. I used the Crohn's disease and ulcerative colitis OR estimates for the 163 IBD-associated loci to generate risk scores and predict case/control status in the PSC sample. There was a significantly greater area under the receiver operating characteristic curve (AUC) when prediction was performed using UC ORs compared to CD ORs (UC AUC = 0.62, CD AUC = 0.56, $P = 1.2 \times 10^{-57}$) (Figure 3.9).

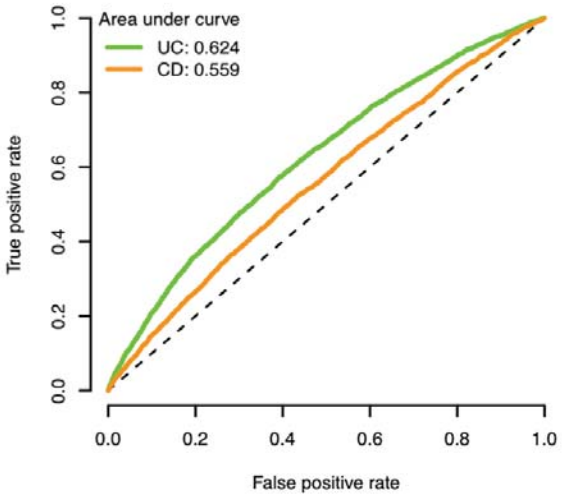

Figure 3.9. Predicting PSC using OR estimates from CD and UC risk loci. The green and orange lines represent the ROC curves for discriminating PSC cases from population controls using UC and CD ORs estimated in Jostins et al. (2012) respectively. The dashed diagonal line us y = x, and specifies the ROC curve of a random predictor.

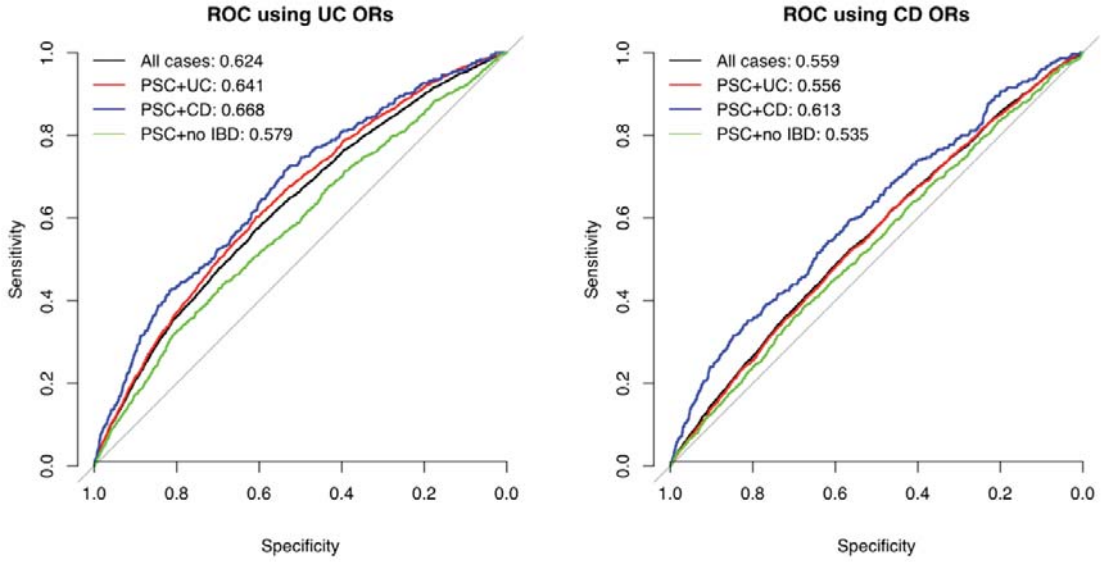

Figure 3.10. Predicting the IBD subphenotypes of PSC patients using OR estimates from CD and UC risk loci. PSC cases were divided into whether they have also been diagnosed with CD, UC, or no IBD, and the performance of the UC and CD ORs predictors assessed for each subphenotype.

That prediction accuracy is greater when performed using UC ORs suggests that PSC is genetically more similar to UC than CD, consistent with clinical observations of greater comorbidity between PSC and UC than CD. However, this

conclusion creates a slight circular argument. It may well be that the higher number of PSC patients with UC than CD is driving this improved prediction. To test this, I repeated the analysis on predicting subsets of PSC cases. PSC cases were divided into whether they have UC, CD, or no IBD (hence referred to as PSC+UC, PSC+CD or PSC+no IBD repsectively) (Table 3.6), and an AUC was estimated using UC and CD ORs on their ability to distinguish each PSC-IBD subset with controls. Notably, the results show that the better predictive performance using UC ORs extends to both PSC+UC and PSC+CD, with little difference in AUCs between the two subsets (PSC+UC AUC = 0.64, PSC+CD AUC = 0.67) (Figure 3.10). This suggests that the previous predictive performance on all PSC cases using UC ORs was not driven by the greater comorbidity between PSC and UC than with CD.

Thus far, I have used the PSC risk loci identified in this study and the IBD risk loci identified in Jostins *et al.* (2012) to illustrate genetic risk factors that are shared and those that are unique to the two diseases. I next considered the degree of sharing that exists genome-wide. Using a linear mixed model that simultaneously considers the effects of all genome-wide SNPs on a phenotype, it is possible to estimate the size of additive genetic variance component, or the total proportion of variance explained, of these SNPs (Yang *et al.,* 2010). In a bivariate extension of the method, it is also possible to estimate additive covariance components due to the SNPs, and provide an estimate of the genetic correlation (r_G) between two phenotypes (Lee *et al.,* 2012). I estimated the degree of genetic correlation between PSC and IBD using individual-level genotype data from an on-going PSC GWAS (2,871 cases and 12,091 controls) (Sun-Gou Ji, personal communication) and from previous IBD GWAS meta-analyses (5,322 CD caess, 6,307 UC cases and 12,164 controls) (Franke *et al.,* 2010; Anderson *et al.,* 2011a; Jostins *et al.,* 2012).

When considering all PSC cases, the genetic correlation was higher between PSC and UC (r_G = 0.47) than PSC and CD (r_G = 0.21), in line with the previous results showing that overlap at specific risk loci (Figure 3.11). Repeating the analysis on subsets of PSC according to their IBD diagnoses (Table 3.6) also showed similar levels of genetic correlation, with the exception of between

PSC+no IBD patients and CD (r_g = 0.038), which was not significantly different from 0 (P = 0.23). Removing the HLA region from the analysis increased estimates of r_G for both between PSC and CD (r_G = 0.26) and PSC and UC (r_G = 0.55), suggesting that variants in the HLA complex confer different effects on PSC and IBD. This is not surprising given differences in effect sizes between HLA variants in PSC and UC. For instance, rs4143332, which tags the HLA-B*08:01 haplotype, shows no evidence of association in UC (P = 0.12; UC data from Chapter 4), while it is by far the strongest associated variant in PSC in this study ($P = 6.39\times10^{-249}$) (Table 3.5 and Figure 3.6).

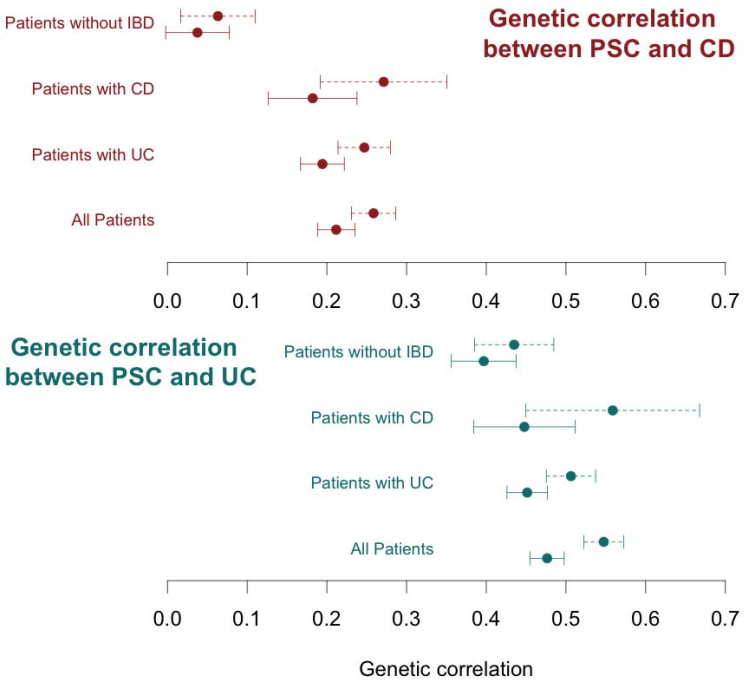

Figure 3.11. Genetic correlation (r_G) estimates using genome-wide SNP data between CD/UC and PSC subphenotypes. Error bars represent standard errors. The dashed error bars and points represent r_G estimates when the HLA region is excluded.

The previous sets of analyses looked at three levels of the genetic overlap between PSC and IBD: within 12 PSC risk loci, within 163 IBD risk loci, and genome-wide. Taken together, the results demonstrate that there is indeed a high degree of genetic overlap between PSC and IBD, that this overlap is stronger between PSC and UC, and does not appear dependent on the IBD-status of the

PSC patient. Given the unclear aetiology of PSC, this raises questions about how pleiotropy can arise. Is PSC a direct result of IBD (and in particular, UC), in which a number of genetic and environmental modifiers affecting existing IBD patients give rise to PSC? Or is PSC a distinct disorder in its own right that shares phenotypic features and genetic risk factors with IBD, much in the same way that CD and UC are considered distinct?

In order to answer help answer these nosological questions, it is important to distinguish between the various situations in which pleiotropy can arise. If it is assumed that a single causal variant underlies a locus that is associated with two correlated phenotypes, the observed pleiotropy can either be mediated by shared biology (biological pleiotropy) or via only one of the phenotypes (mediated pleiotropy). In the former case, the causal variant may reflect molecular processes that result in distinct pathological features (e.g. in different cell types), leading to increased risk for both diseases. In the later case, apparent pleiotropy will be observed if the first phenotype directly causes the second, such that associations with the second phenotype are due entirely to this phenotypic correlation. These two models are illustrated in Figure 3.12, where PSC and UC are modelled as two distinct phenotypes that share a causal genetic variant, or where PSC is a direct consequence of UC. Mediated pleiotropy can be tested by looking for an association in the second phenotype in individuals where the first phenotype is not present. If the association signal persists, then the observed pleiotropy is more likely due to shared biology rather than being mediated by one of the phenotypes (Solovieff *et al.,* 2013). More generally, Mendelian randomisation can also be used to tease out the causal relationships, however, neither approach can distinguish between the models of pleiotropy when there exists one or more confounding factors that affect both the phenotypes and is also influenced by genotype (Lawlor *et al.,* 2008).

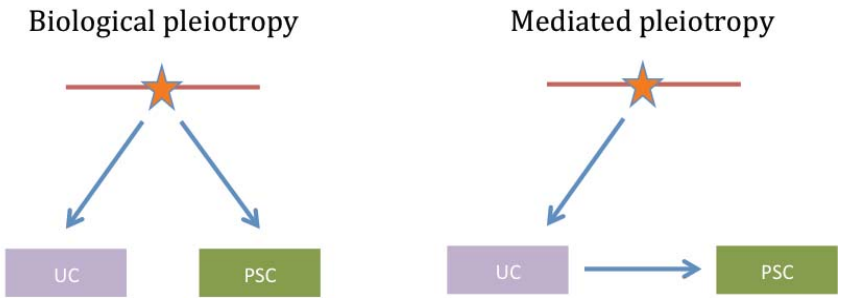

Figure 3.12. Two models of pleiotropy. The star represents a causal genetic variant tagged by a SNP associated to both UC and PSC. The arrows indicate the directions of causality between the SNP and phenotypes. Figure adapted from Solovieff et al. (2013).

Of the 12 genome-wide significant PSC risk variants identified here, six were also reported to be associated with UC (Figure 3.7). If PSC is partly mediated via IBD (and UC in particular), then it may be that these observed PSC associations at UC SNPs are due to mediated rather than biological pleiotropy. To test this, I again stratified PSC cases into subsets of whether they were also diagnosed with UC (n = 1,898) or had no IBD (n = 922) (Table 3.6). I then repeated the association analysis for each subset against controls. There was no evidence for any differences in odds ratios at any of the PSC and UC-associated variants, nor for that matter, any of the other six genome-wide significant PSC risk variants (Figure 3.13). It would have also been possible to stratify PSC cases into those with CD or indeterminate IBD, though there were much fewer samples of these and hence little power to detect any differences.

While these results suggest that common biology rather than phenotypic correlation explains the pleiotropy between PSC and IBD at these loci, caution must be applied when extrapolating these to all PSC and IBD associated loci. Firstly, this analysis was only performed on risk loci that were detected using all PSC cases. Hence variants that affect all PSC individuals are much more likely to be discovered than those associated with only a subphenotypes of PSC. Secondly, it remains to be seen how many of the non-IBD PSC cases will go on to develop IBD. The two diseases share some gastrointestinal symptoms, and while IBD precedes PSC in the majority of cases, the onset of both conditions may be separated be several years (Saich and Chapman, 2008). Finally, larger sample sizes will be required to obtain an accurate assessment of whether any

associations at additional risk loci, especially those with known IBD associations, are driven by the correlated phenotypes or shared biology.

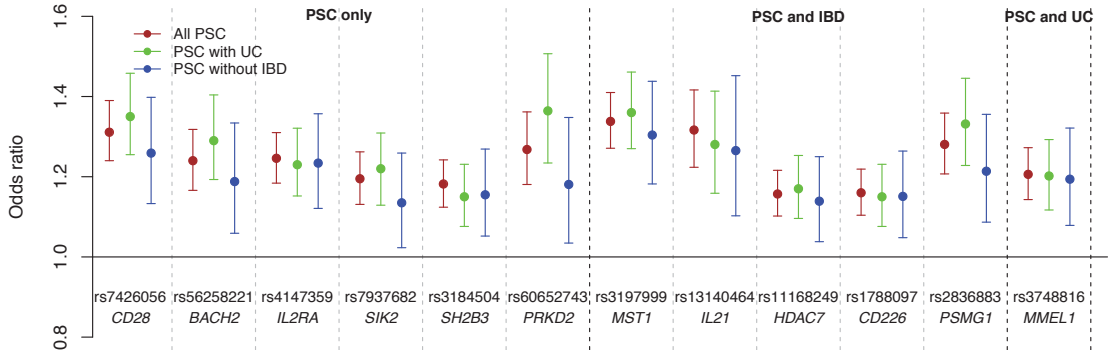

Figure 3.13. Odds ratios of PSC risk loci calculated using all PSC cases compared with odds ratios calculated using PSC+UC and PSC+no IBD subphenotypes. The error bars represent 95% confidence intervals. Designation of whether loci are associated with just PSC or both PSC and IBD follows that of Figure 3.7.

Overall, I showed that the genetic overlap between PSC and IBD is pervasive, and that this overlap is stronger between PSC and UC than between PSC and CD, mirroring the phenotypic comorbidity between the diseases. Within PSC risk loci, the genetic effects appear independent of whether UC was diagnosed along with PSC, suggesting that these loci reflect shared biology between the two diseases rather than a UC to PSC causal relationship. Incorporating association results with disease relevant functional genomic datasets may provide leads in uncovering the mechanisms behind this pleiotropy: does the causal variant result in distinct pathological features in different cells types, and do these differences reflect different disease states? I will explore approaches of integrating functional genomic datasets with disease risk loci to help answer these types of questions in Chapter 5.

3.4 Conclusion

Through genotyping of 3,789 PSC cases and 25,079 controls using the Immunochip, this study identified 12 non-HLA genome-wide significant loci, of

which nine are implicated in PSC for the first time. Network analysis using GRAIL and DAPPLE, along with searching for known eQTLs and coding variants revealed at least one candidate gene in at 11 of these loci, three of which are linked by genes that interact with each other to mediate T cell apoptosis (*SIK2*, *HDAC7* and *PRKD2*), offering new leads into the pathogensis of PSC.

The data also convincingly show pervasive overlap between genetic variants that affect PSC and IBD, and that this overlap is greater between PSC and UC than between PSC and CD, reflecting the observed comorbidity between the disorders. As many as half the variants are shared between PSC and UC when considering PSC and UC risk loci, as well as the genetic covariance between the two disorders tagged by SNPs genome-wide. Stratifying PSC cases into those with and without UC strongly suggests that this overlap is due to biological rather than mediated pleiotropy. This study demonstrates the utility of cheap high-density genotyping arrays in discovering novel loci and enabling powerful cross-phenotype comparisons.