Chapter 2. Discovery, refinement and functional genomics integration of primary biliary cirrhosis risk loci using the Immunochip

2.1 Introduction

Primary biliary cirrhosis (PBC) is characterized by the immune-mediated destruction of intra-hepatic bile ducts, resulting in chronic cholangitis, liver fibrosis and ultimately cirrhosis (Kaplan and Gershwin, 2005). With a UK prevalence of 35:100,000, rising to 94:100,000 women over 40 years of age, it is the most common autoimmune liver disorder (James *et al.,* 1999; Kaplan and Gershwin, 2005). Family-based studies indicate a substantial genetic component to PBC susceptibility, with a sibling recurrence risk of ~10.5 in the UK (Jones *et al.,* 1999). Genome-wide association studies (GtwitWAS) have identified 22 PBC risk loci, and highlighted the role of NFkB signaling, T-cell differentiation, Toll-like receptor and tumor necrosis factor signalling in disease pathogenesis (Hirschfield *et al.,* 2009; Liu *et al.,* 2010b; Mells *et al.,* 2011). Sixteen of these loci are also associated with other immune-mediated diseases such as multiple sclerosis, celiac disease and type 1 diabetes (T1D), shedding light on the involvement of common genes and pathways across these diseases (Zhernakova *et al.,* 2009). Despite these advances, the specific causal variant at many of these loci remains unknown.

To better define known risk variants and identify additional susceptibility loci, I performed an association study in 2,861 cases from the UK PBC

Consortium and 8,514 UK population controls from the 1958 British Birth Cohort and National Blood Service. All samples were genotyped using the Immunochip, an Illumina Infinium array containing 196,524 variants (718 small insertions/deletions and 195,806 SNPs). Two thirds of these variants reside in 186 loci with known associations with one or more autoimmune disorders, while most of the remaining variants were included as part of GWAS replication efforts for various autoimmune disorders (Cortes and Brown, 2011; Trynka *et al.,* 2011a). Compared with GWAS arrays, the Immunochip has increased marker density within known autoimmunity-associated loci, increasing the power to detect PBC associations within these selected candidate loci and providing a powerful means of fine mapping known PBC loci, as causal variants are more likely to be directly genotyped.

2.1.1 Chapter overview

In this chapter, I describe the results from an association study for PBC risk loci. In total, 19 loci reach genome-wide significance ($P < 5\times10^{-8}$), three of which are novel. One of these novel loci includes a low-frequency non-synonymous SNP in *TYK2*, further implicating JAK/STAT and cytokine signalling in PBC pathogenesis. Multiple independent common, low frequency and rare variant associations were found at five loci. Further investigation of one of these regions (3q25) showed that the most significantly associated signal in the locus was driven by a shared haplotype with two other SNPs, and that this top signal was no longer genome-wide significant when testing for association using a joint model of all signals in the region. Imputation and association testing of HLA haplotypes also confirmed three known independent genome-wide significant associations. Finally, I observed that 15 of the 26 independent non-HLA association signals overlapped with regions of open chromatin in B-lymphoblastoid cell lines as identified in the ENCODE project, though this was not significantly different compared to other cell lines when taking LD and the SNP composition on the Immunochip into account ($P = 0.06$).

2.1.2 Contributions

The study design was conceived by the Wellcome Trust Case Control Consortium 3 (WTCCC3) and the UK PBC Consortium. Case ascertainment and phenotyping were performed by the UK PBC Consortium. Controls were ascertained from the UK National Blood Service and the 1958 Birth Cohort Controls group. See Supplementary Note in Liu *et al*. (2012) for the full list of contributors. Sample and SNP quality control was performed by Mohamed Almarri. All other analyses, unless stated, were performed by myself.

2.2 Methods

2.2.1 Samples, DNA extraction and genotyping

All subjects were of self-declared British or Irish ancestry. Cases were collected by the UK PBC Consortium, which consists of 142 NHS trusts including all UK liver transplant centers. All individuals were over 18 years of age with probable or certain PBC. Three criteria were applied to diagnose the condition: a) a positive test for the presence of anitmitochondrial antibodies (titer 1:40 or higher), b) liver biopsy histology consistent with PBC, and c) liver biochemistry consistent with PBC (i.e. a higher level of bilirubin, aspartate transaminase, alanine transaminase, alkaline phosphatase or gamma-glutamyl transferase compared to the upper reference level). Diagnosis was documented as probable when two criteria were satisfied and certain if all three criteria were satisfied. A total of 2,981 cases were supplied by the UK PBC Consortium. 8,970 control samples were ascertained from the 1958 British Birth Cohort and the National Blood Service. This study contains 1,838 cases and 2,356 controls that were also included in a recent PBC GWAS (Mells *et al.,* 2011).

DNA was extracted from blood or saliva. Blood samples from PBC patients were extracted by the East Anglian Medical Genetics Service, while saliva samples were collected using an Oragene kit and DNA extracted at Source BioScience Healthcare. DNA samples were plated, normalized and shipped to the Wellcome Trust Sanger Institute for sample quality control.

Samples were genotyped on an Illumina iSelect HD custom genotyping array (Immunochip). All 2,981 cases and 4,537 controls were genotyped at the Wellcome Trust Sanger Institute. A further 4,433 control samples were genotyped at the Center for Public Health Genomics at the University of Virginia. Genotyping of control samples was coordinated by the Immunochip consortium for use in several Immunochip projects. The NCBI build 36 (hg18) map was used (Illumina manifest file Immuno_BeadChip_11419691_B.bpm). Normalized probe intensities were extracted for all samples passing standard laboratory QC thresholds and genotypes were called using optiCall (Shah *et al.,* 2012). Genotypes with an individual posterior probability lower than 0.7 were defined as unknown. optiCall was chosen because we found it to be more accurate in calling common and low-frequency variants on Immunochip compared to other established algorithms such as Illuminus (Teo *et al.,* 2007) and GenoSNP (Giannoulatou *et al.,* 2008; Shah *et al.,* 2012)

2.2.2 Quality control

Sample quality control (QC) was performed for each sample set separately. All monomorphic SNPs were removed prior to QC. Samples with a call rate lower than 98% and heterozygosity more than three standard deviations from the mean were excluded. A set of LD-pruned SNPs with minor allele frequency (MAF) > 20% were used to estimate identity by descent (IBD) and ancestry. For each pair of individuals with an estimated IBD > 18.75%, the sample with the lower call rate was removed. Principal component analysis was used to exclude samples of non-European ancestry (Price *et al.,* 2006) (Figure 2.1).

Following sample QC 2,861 cases and 8,514 controls remained (Table 2.1). SNPs with a minor allele frequency less than 0.1%, Hardy-Weinberg equilibrium $P < 10^{-6}$ in controls, call rate lower than 98%, or significantly different ($P < 10^{-5}$) call rate in cases vs. controls (or between the two control sets) were excluded. After marker QC 143,020 polymorphic SNPs were available for analysis (Table 2.2).

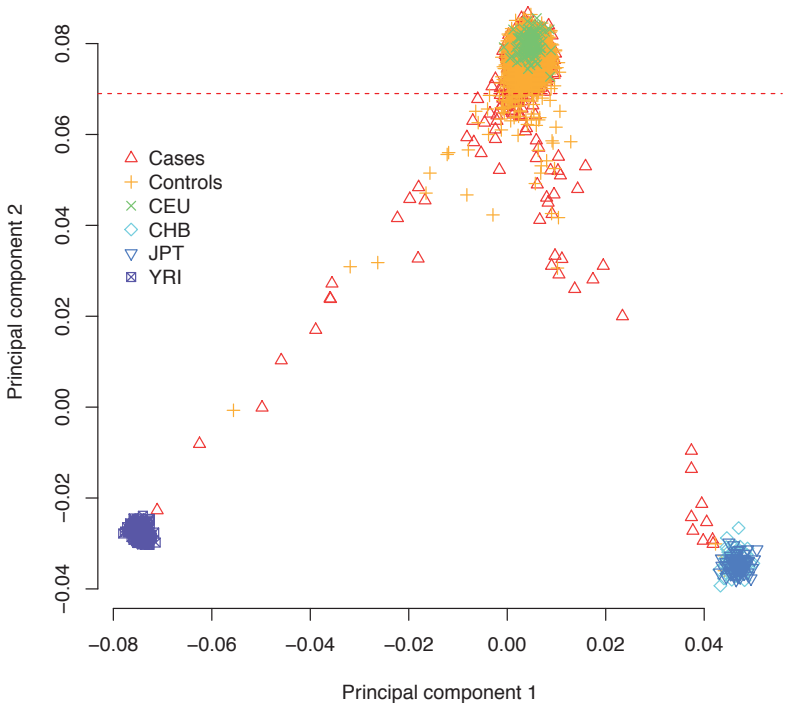


Figure 2.1. Principal component analysis of PBC cases and controls. The first two principal components were calculated for 18,995 SNPs that had MAF > 20% on the Immunochip and overlap those for the CEU, CHB, JPT and YRI HapMap samples. The red horizontal line indicates the exclusion threshold on the second principal component.

Sample	Heterozygosity/ missingness	Relatedness	Ancestry	Total[a]
Cases	29	47	65	140
Controls 1	70	187	32	224
Controls 2	37	169	53	232
Total	136	403	150	596

Table 2.1. Sample quality control. [a]Some samples failed more than one QC metric

Sample	HWE[a]	Call rate	MAF[b]	NRM[c]	Total Remaining[f]
Cases	-	8,301	39,504	9,362[d]	143,020
Controls 1	1,721	6,871	39,954	4,605[e]	
Controls 2	1,771	7,372	40,048		

Table 2.2. SNP quality control. [a]Hardy-Weinberg equilibrium. [b]Minor allele frequency. [c]Non-random missingnes (dbetween cases and controls, [e]between both sets of controls). [f]Some SNPs failed more than one QC metric.

The Immunochip contains 2,258 SNPs that were included as a replication panel for non-immune-mediated disorders. These SNPs were used as null markers to estimate the overall inflation of the distribution of association test statistics (Devlin and Roeder, 1999).

2.2.3 Imputation

Additional genotypes were imputed using 90,977 SNPs from the 186 Immunochip high density regions with the 1000 Genomes Phase I (interim) June 2011 release reference panel and IMPUTE2 (Howie *et al.,* 2009). Imputation was performed separately in three batches of 3,792, 3,792 and 3,791 individuals, with the case/control ratio constant across batches. SNPs with a posterior probability less than 0.9, IMPUTE INFO score < 0.5 and those with differential missingness ($P < 10^{-5}$) between the three batches were removed, as were those SNPs that failed the same exclusion thresholds used for the original Immunochip QC. After imputation, a total of 237,619 SNPs were available for analysis.

2.2.4 Association analysis

Case-control association tests were implemented using a standard one-degree of freedom Cochran-Armitage test for trend in PLINK v1.07 (Purcell *et al.,* 2007). Secondary associations were identified using step-wise logistic regression analysis conditioning on the allelic dosage of the primary signal in each significant locus. The process was repeated, conditioning on all independent genome-wide significant SNPs, until all genome-wide significant signals were accounted for (Cordell and Clayton, 2002). Cluster plots for all SNPs $P < 5\times10^{-6}$ were manually checked using Evoker (Morris *et al.,* 2010), and poorly called SNPs were removed from further study.

2.2.5 HLA Imputation

Imputation of six classic HLA alleles (class I: HLA-A, HLA-B and HLA-C, class II: HLA-DQA1, HLA-DQB1 and HLA-DRB1) was performed using the prediction algorithm proposed by Leslie et al. and implemented in the program HLA*IMP (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. Case-control association was performed on HLA allele posterior probabilities generated from HLA*IMP using logistic regression to account for genotype uncertainty following imputation. Stepwise conditional logistic regression was used to identify independent association signals among the 21 HLA-alleles that reached $P < 0.0001$.

2.2.6 Variance in disease risk explained

The variance in disease risk explained by the 26 independent genome-wide significant SNPs and four HLA-alleles was estimated using a disease liability threshold model (Falconer and Mackay, 1996; So *et al.,* 2011) assuming a disease prevalence of 40/100,000 and log-additive risk. A review of population-based epidemiological studies of PBC found prevalence rates varied from 1.9 to 40.2 per 100,000 depending on the surveyed population, time of survey and phenotype definitions (Boonstra *et al*., 2012). The choice of using 40/100,000 in variance explained calculations was based on three recent large population-based surveys in European populations, where prevalence was estimated to be 38.3-40.2/100,000 (Podda *et al*., 2013).

2.2.7 eQTL analysis

Expression quantitative trait loci (eQTLs) within genome-wide significant loci were collated from the University of Chicago eQTL Browser (http://eqtl.uchicago.edu/cgi-bin/gbrowse/eqtl/) and a study by Gaffney *et al.,* (2012). The eQTL Browser contains significant eQTLs that were identified in recent studies across multiple cell lines and populations, while Gaffney *et al.,* reanalysed gene expression data from 210 lymphoblastoid cell lines using a total of 13.6M SNPs from the 1000 Genomes Project. For more details, see Gaffney, *et al.* (2012) and references listed in http://eqtl.uchicago.edu/cgi-bin/gbrowse/eqtl/.

2.2.8 Enrichment of open chromatin regions

The Encyclopedia of DNA Elements (ENCODE) project annotated regions of open chromatin using the direct sequencing of DNase-I hypersensitive sites (DNase-seq: sixteen different cell lines) (Myers *et al.,* 2011; Song *et al.,* 2011). The approach involves isolating nucleosome-depleted regions of DNA and mapping reads from next-generation sequencing to determine their location. I estimated the amount of enrichment for open chromatin peaks among significant PBC risk loci across the ENCODE cell lines. SNPs were first grouped into independent loci – beginning with the most strongly associated SNP (the "lead SNP"), I assigned SNPs in moderate LD with the lead SNP ($r^2 > 0.1$) to the associated locus, while those in high LD ($r^2 > 0.8$) were also considered candidate causal SNPs. The process then proceeds to the next most significantly associated SNP (that had not already been assigned to a locus), and assigned to the next locus this SNP along with those in moderate and high LD to this new locus, and so on. After the addition of each new locus, I calculated E,

$$E = \frac{OC_{loci;}/N_{loci}}{OC_{ichip}/N_{ichip}}$$

where, for a given cell line, OC_{loci} and N_{loci} are the number of candidate causal SNPs ($r^2 > 0.8$ with the lead SNP(s)) that lie within open chromatin peaks across the selected loci and the total number of SNPs within the loci ($r^2 > 0.1$ with the lead SNP(s)), respectively. OC_{ichip} and N_{ichip} are the equivalent measures across all SNPs within Immunochip high density regions. I only included the high density regions to increase the likelihood that the causal variant was assayed, and excluded SNPs in the HLA and those with MAF < 0.05 to avoid possible biases due to LD structure. To compare *E* between cell lines, the number of candidate causal SNPs in open chromatin ($OC_{loci:allcells}$) and the total number SNPs in open chromatin ($OC_{ichip:allcells}$) were first calculated for the union of open chromatin peaks across all cell lines other than that being evaluated. I then tested the alternative hypothesis that, for a given cell line, the proportion $OC_{loci}/OC_{ichip} > OC_{loci:allcells}/OC_{ichip:allcells}$ using a one-sided binomial test.

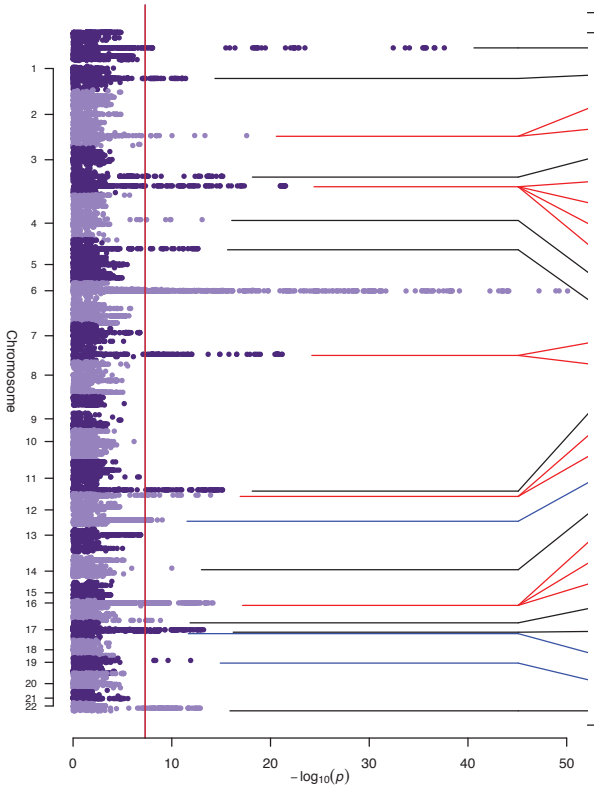
To ensure that the test was well calibrated under the null hypothesis I undertook 1000 permutations of PBC case control labels, repeating the association and enrichment analyses for each permutation. Comparing the observed level of enrichment at the top 21 loci to the equivalent from the permutations I obtained a similar, non-significant empirical P-value of 0.073 indicating that the proposed enrichment analysis is well calibrated under the null. A 95% confidence interval for *E* was estimated using the permutations.

2.3 Results and discussion

Following quality control, 143,020 polymorphic SNPs were available across 2,861 cases and 8,514 controls. (Table 2.1, Table 2.2, Figure 2.1). A further 94,559 SNPs in the Immunochip fine-mapping regions were imputed using genotypes from the 1000 Genomes June 2011 release. The inflation factor inferred from 2,258 SNPs not associated with autoimmune disease showed only a modest inflation (λ=1.096), similar to that reported in a previous GWAS study that included 4,194 overlapping samples (Mells *et al.,* 2011).

2.3.1 Replicating known PBC risk loci

Sixteen of the 22 known PBC risk loci reached genome-wide significance ($P < 5\times10^{-8}$) (Figure 2.2) and four showed nominal evidence of association ($5\times10^{-8} < P < 5\times10^{-4}$). Two PBC risk loci, 14q32 and 19q13, were not included on Immunochip as the array was designed before the publication of the most recent PBC GWAS (Mells *et al.,* 2011). At 12 of the genome-wide significant loci, the most associated SNP was different to that previously reported (Figure 2.3). There was little difference in the effect-size estimates between the GWAS tagging SNP and the most strongly associated Immunochip SNP, although this may partly be due to a large proportion of overlapping samples between the two studies. Nevertheless, the similarities in ORs despite the denser coverage of the Immunochip suggests that the ORs of tag-SNPs in GWAS adequately reflect the true ORs, and that synthetic associations are unlikely to explain the associations at these risk loci (Anderson *et al.,* 2011b).



SNP[a]	RA[b]	Chr:Position[c]	RAF[d]	OR[e]	P–value[f]	Gene(s)	Annotation[g]
rs72678531	G	1:67,571,033	0.17	1.61	2.47×10^{-38}	*IL12RB2*	OC
rs2488393	A	1:195,988,863	0.21	1.28	4.29×10^{-12}	*DENND1B,CRB1*	
rs3024921	A	2:191,651,517	0.06	1.62	2.59×10^{-18}	*STAT1,STAT4*	
rs7574865	A	2:191,672,878	0.22	1.31	1.38×10^{-13}	*STAT1,STAT4*	
rs2293370	G	3:120,702,624	0.80	1.39	6.84×10^{-16}	*TMEM39A,POGLUT1,TIMMDC1,CD80*	NS
rs2366643	A	3:161,219,179	0.57	1.35	3.92×10^{-22}	*IL12A*	OC
rs62270414	G	3:161,122,353	0.15	1.41	5.74×10^{-17}	*IL12A*	OC
rs668988	G	3:161,198,245	0.43	1.26	4.73×10^{-09}	*IL12A*	OC
rs80014155	A	3:161,108,087	0.004	3.44	2.64×10^{-11}	*IL12A*	
rs7665090	G	4:103,770,651	0.52	1.26	8.48×10^{-14}	*MANBA,NFKB1*	NS,eQTL
rs6871748	A	5:35,921,739	0.72	1.30	2.26×10^{-13}	*IL7R,CAPSL,SPEF2,UGT3A1*	NS
rs35188261	A	7:128,470,775	0.17	1.52	6.52×10^{-22}	*IRF5,TNPO3*	OC
rs3807307	G	7:128,366,438	0.47	1.22	4.12×10^{-09}	*IRF5,TNPO3*	OC,eQTL
rs80065107	A	11:118,187,977	0.79	1.39	7.20×10^{16}	*DDX6*	OC
rs1800693	G	12:6,310,270	0.40	1.27	1.18×10^{-14}	*TNFRSF1A,LTBR,SCNN1A*	OC
rs11064157	A	12:6,362,910	0.25	1.23	1.69×10^{-09}	*TNFRSF1A,LTBR,SCNN1A*	
rs11065979	A	12:110,543,940	0.44	1.20	2.87×10^{-09}	*ATXN2,BRAP,SH2B3*	NS
rs911263	A	14:67,823,346	0.71	1.26	9.95×10^{-11}	*RAD51B*	
rs1646019	G	16:11,267,181	0.71	1.31	6.72×10^{-15}	*SOCS1,CLEC16A,PRM1,PRM2*	OC,eQTL
rs12708715	C	16:11,085,325	0.68	1.29	2.69×10^{-08}	*SOCS1,CLEC16A,PRM1,PRM2*	OC,eQTL
rs80073729	A	16:11,281,298	0.004	2.96	2.69×10^{-08}	*SOCS1,CLEC16A,PRM1,PRM2*	OC
rs11117433	G	16:84,577,017	0.77	1.26	1.41×10^{-09}	*IRF8*	OC
rs8067378	G	17:35,304,874	0.52	1.26	6.05×10^{-14}	*ORMDL3,ZPBP2,GSDMB,IKZF3*	NS,OC,eQTL
rs17564829	G	17:41,362,429	0.24	1.25	2.15×10^{-09}	*CRHR1,MAPT*	NS,OC,eQTL
rs34536443	G	19:10,324,118	0.95	1.91	1.23×10^{-12}	*TYK2*	NS
rs2267407	A	22:38,076,996	0.23	1.29	1.29×10^{-13}	*SYNGR1,PDGFB,RPL3*	OC,eQTL

Figure 2.2. Manhattan plot and table of genome-wide significant PBC risk loci. Novel loci are coloured in blue. Loci with multiple independent risk loci are coloured in red. [a]Most significantly associated SNP in locus. [b]Risk allele. [c]Base-pair position (NCBI36). [d]Risk allele frequency. [e]Odds ratio. [f]P-value for primary signals calculated from the Cochran-Armitage test for trend. Secondary signals calculated from stepwise logistic regression. [g]Whether SNPs in high linkage disequilibrium (r2 > 0.8) with the lead SNP overlap one of more of the following annotations: eQTL (expression quantitative trait loci), NS (non-synonymous SNP), OC (open chromatin).

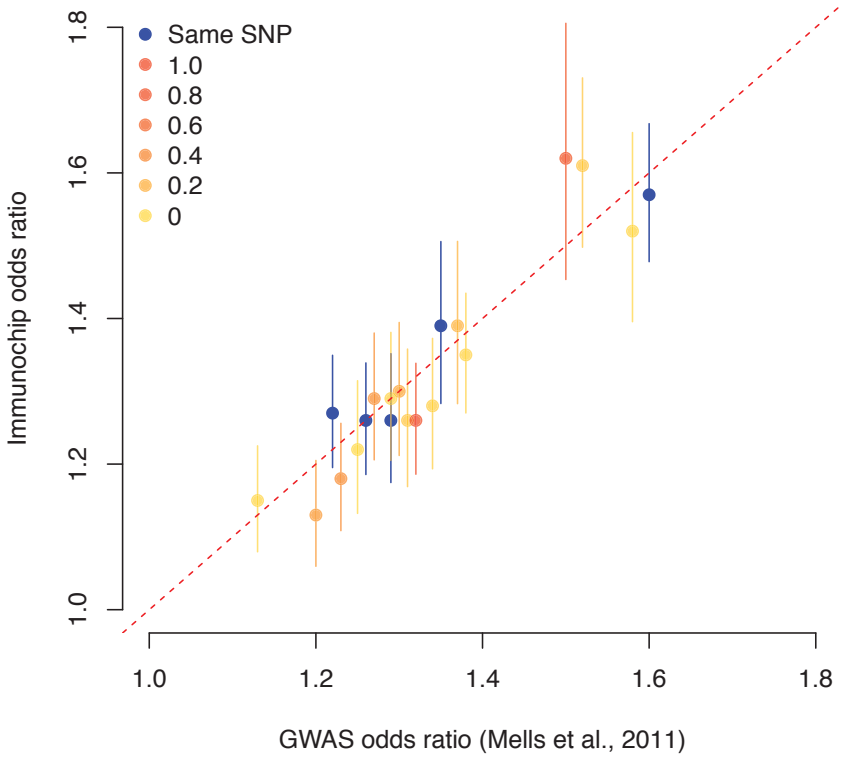


Figure 2.3. PBC risk loci odds ratios from this study vs. those from Mells *et al.* (2011). Colours denote the LD (r^2) between the SNP reported in Mells *et al.* and the most significant SNP in the same locus in this study. Error bars represent OR 95% confidence intervals. The red dashed line is $y = x$.

2.3.2 Multiple independent signals

Stepwise conditional regression (Cordell and Clayton, 2002) revealed multiple independent signals at five loci, with 16p13 harbouring three, and 3q25 four such associations (Figure 2.2, Figure 2.4). At the 16p13 locus, the third independent signal, rs80073729, is a rare SNP (MAF < 0.5%) recently associated with celiac disease (Trynka *et al.,* 2011). In the same study, Trynka et al. (2011) also identified multiple independent signals at 3q25, though rs80014155, a rare SNP that best tags the fourth independent PBC association at this locus, was not among them.

Further dissection of the four independent signals in 3q25 region revealed a complex genetic architecture. While stepwise conditional regression revealed

four independently associated SNPs (henceforth referred to as SNPs 1 through 4, as ordered according to P-value), jointly modelling the four SNPs in a multiple logistic regression model revealed large differences in the strength of association for SNPs 1 and 2 when compared with univariate regression (Figure 2.4, Table 2.3). For SNP 1, the strength of association fell from $P = 5.58\times10^{-22}$ to $P = 5.23\times10^{-7}$. Conversely, the strength of association for SNP 2 increased from $P = 6.75\times10^{-12}$ to $P = 2.93\times10^{-25}$. These changes are in part driven by the LD patterns within this region. SNP 1 is in moderate LD with SNPs 3 ($r^2 = 0.322$, D' = 0.75) and 4 (r2 = 0.004, D' = 0.91), such that the risk increasing alleles for all three SNPs reside more often on the same haplotype background. Thus the strong association signal for SNP 1 from a univariate association test is partly driven by its correlation with SNPs 3 and 4. Conversely, SNPs 2 and 3 are also in moderate LD ($r^2 = 0.10$, D' = 0.90), although in this case, the risk increasing allele of SNP 2 more often shares the same haplotype as the risk decreasing allele of SNP 3. Hence, the signal for SNP 2 is diluted by the risk decreasing effects of SNP 3 when performing a univariate test. Indeed, by accounting for the effects of independent SNPs in this region, it appears that SNP 2 is the most strongly associated signal ($P = 2.93 \times 10^{-25}$) while SNP 1 is no longer genome-wide significant ($P = 5.23 \times 10^{-7}$).

	SNP	RAF[a]	RA[b]	uncond OR[c]	uncond P[c]	cond OR[d]	cond P[d]
1	rs2366643	0.57	A	1.36	5.58×10^{-22}	1.22	5.23×10^{-7}
2	rs62270414	0.15	G	1.32	6.75×10^{-12}	1.59	2.93×10^{-25}
3	rs668998	0.44	G	1.26	1.98×10^{-14}	1.31	3.05×10^{-11}
4	rs80014155	0.004	A	3.07	6.66×10^{-10}	3.44	2.64×10^{-11}

Table 2.3. Unconditioned and conditioned association results for the four independent signals at 3q25. aRisk allele frequency. bRisk allele. cOdds ratios and P-values from univariate (unconditioned) association tests. dOdds ratios and P-values from multiple logistic regression model that includes all four SNPs as covariates.

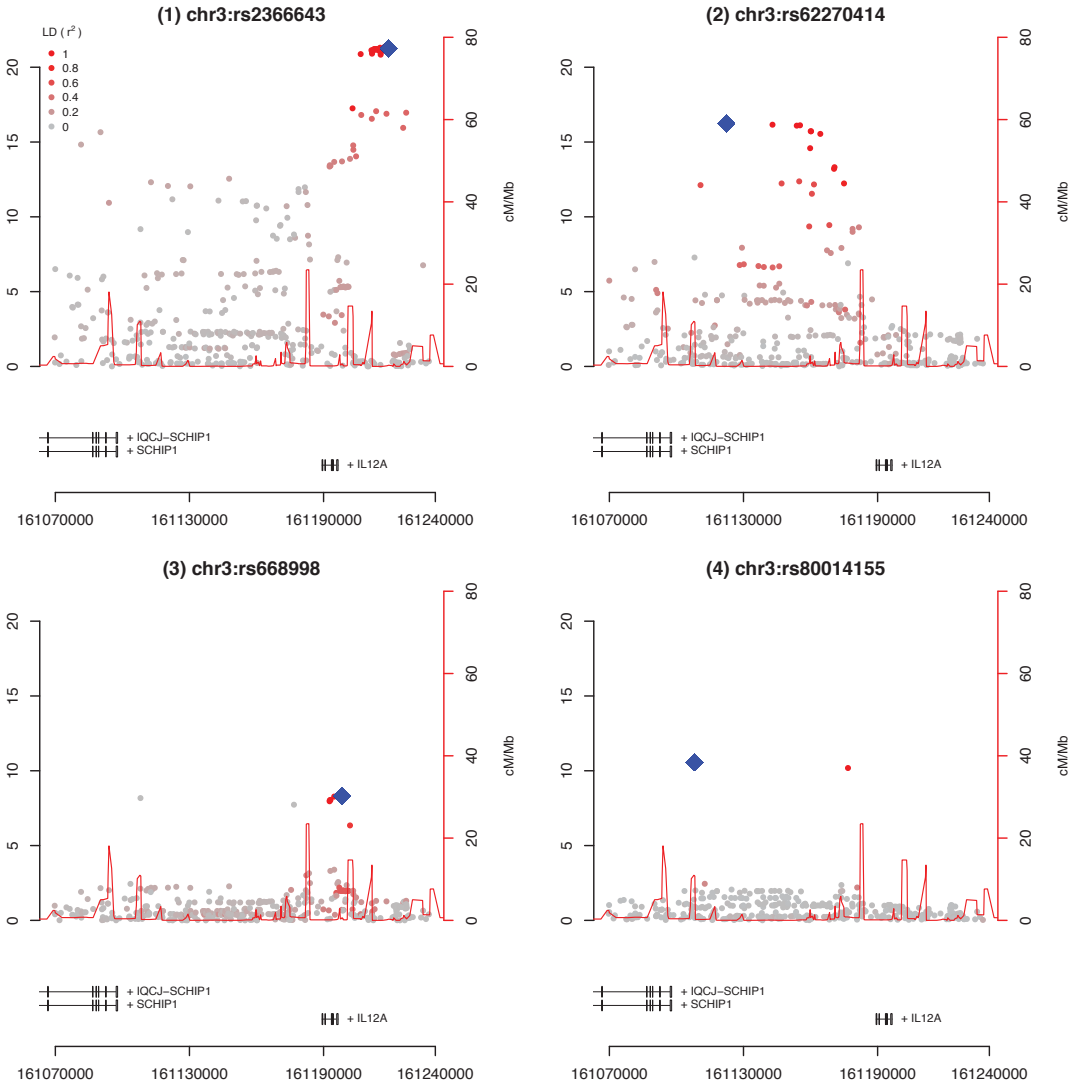


Figure 2.4. Multiple independent signals at 3q25 from stepwise conditional regression. Panel 1 shows the regional association plot of all SNPs with no conditioning. Panel 2 shows association results when conditioned on rs2366632. Panel 3 shows association results when conditioned on rs2366632 and rs62270414. Panel 4 shows association results when conditioned on rs2366632, rs62270414 and rs668998. The colour gradient (from red to grey) represents the strength of LD between the lead SNP (in blue) and others in the region.

The architecture of the 3q25 region demonstrates how the haplotype structure between associated SNPs can both dilute and inflate marginal association signals, such that stepwise regression may not completely reveal the true effect sizes of multiple associated SNPs within a region. In these situations, a two-stage procedure consisting of SNP-selection using conditional regression

and then performing joint multiple regression including the selected SNPs, may be more appropriate.

The identification of multiple independent signals show that resequencing efforts in large number of cases across known GWAS loci will be a powerful means of identifying additional independent signals (Hunt *et al.,* 2013). It is likely that the two rare SNP associations at 3q25 and 16p13 would have been overlooked using standard GWAS arrays due to poor tagging, unless they were directly genotyped. For example, in a case control study of 10,000 cases and 10,000 controls, there is only 0.07% power to detect association with the closest tagging SNP of rs80073739 on the Illumina Human1M chip, rs11649025 (minor allele frequency = 10%, r^2 = 0.04, D' = 1), at $P < 5\times10^{-8}$. These additional independent association signals thus yield a more complete understanding of the genetic architecture of PBC and enable more informative genotype-based recall and fine-mapping studies to be conducted.

2.3.3 Novel PBC risk loci

Three newly-associated PBC risk loci reached genome-wide significance (Figure 2.2). The strongest association on 19p12, rs34536443 (OR = 1.91, $P = 1.24\times10^{-12}$), is a low-frequency (MAF = 0.05) non-synonymous SNP in the tyrosine kinase 2 gene (*TYK2*), and is also associated with multiple sclerosis (Ban *et al.,* 2009). The locus has also been implicated in T1D (Wallace *et al.,* 2010), psoriasis (Strange *et al.,* 2010) and Crohn's disease (Franke *et al.,* 2010), although rs34536443 was not genotyped as part of these studies. For T1D and psoriasis, the strongest associations were to common SNPs that reside on the same haplotype (rs2304256: r^2 = 0.06, D' = 0.9 and rs280519: r^2 = 0.03, D' = 1). The most associated SNP in Crohn's disease and the second psoriasis signal (rs12720356) is independent of rs34536443 (r^2 = 0, D' = 0.003). The 12q24 locus has been associated with celiac disease (Hunt *et al.,* 2008; Trynka *et al.,* 2011a), rheumatoid arthritis (Stahl *et al.,* 2010) and T1D (Barrett *et al.,* 2009), though it was a non-synonymous SNP in *SH2B3*, rs3184504 (OR = 1.19, $P = 1.11\times10^{-8}$), rather than the most significant SNP in this study, rs11065979 (OR =

1.2, P = 2.87×10^{-9}), that was most strongly associated. The two SNPs are in high LD (r^2 = 0.81) and further studies are required to narrow the set of potential causal variants underlying the PBC association signal at this locus. The most associated SNP in the 17q21 region, rs17564829 (OR = 1.25, P = 2.15×10^{-9}), is located in *MAPT*, a gene that has been associated with cognitive symptoms in Parkinson's disease. While cognitive symptoms sometimes associated with PBC, it remains to be seen if the true causal variant at the locus has its functional effect through *MAPT*, and whether this functional effect then results in cognitive changes in PBC patients.

Both *TYK2* and *SH2B3* are involved in the production of cytokines, adding to the evidence that cytokine imbalances play a role in PBC and other autoimmune diseases (Rong *et al.,* 2009; Wang *et al.,* 2010a). *TYK2* is a member of the Janus kinase family, which transduce cytokine signals by phosphorylating STAT transcription factors. Couturier *et al.* (2011) showed that heterozygotes for rs34536443 have significantly reduced *TYK2* activity, which promotes the secretion of Th2 cytokines (Couturier *et al.,* 2011). For *SH2B3*, carriers of the A risk allele of rs3184504 show a moderate increase in production of cytokines and stronger activation of the NOD2 recognition pathway compared to carriers of the G allele (Zhernakova *et al.,* 2010), suggesting a possible role in helping prevent bacterial infection.

2.3.4 Associations with HLA haplotypes

Candidate genes studies have implicated several HLA-DR alleles in PBC susceptibility, particularly the DRB1*08 allele (Donaldson *et al.,* 2006; Invernizzi *et al.,* 2008; Mullarkey *et al.,* 2005; Wassmuth *et al.,* 2002). However, such studies were hindered by small sample sizes resulting in low power. As the Immunochip includes much denser SNP coverage of the MHC, it is expected that more HLA-types will be able to be imputed at greater accuracy than using traditional GWAS SNP chips. Here, the classical HLA alleles (HLA-A, B, C, DQA1, DQB1 and DRB1) were imputed from genotyped SNPs in the MHC (Dilthey *et al.,* 2011; Leslie *et al.,* 2008). Fourteen HLA-alleles reached genome-wide significance and conditional

analysis clustered these associations into four independent signals (Table 2.4). The most significant association was the HLA-DQA1*0401 allele (OR = 3.06, P = 5.9×10^{-45}), which forms a haplotype with two other HLA class II alleles (DQB1*0402 and DRB1*0801) and is an established PBC risk locus (Donaldson *et al.,* 2006; Invernizzi *et al.,* 2008; Mullarkey *et al.,* 2005; Wassmuth *et al.,* 2002). The second and third most significant clusters, DQB1*0602 (OR = 0.64, P = 2.32×10^{-15}) and DQB1*0301 (OR = 0.70, P = 6.48×10^{-14}) both have protective effects, confirming previous studies showing suggestive associations between these loci and PBC susceptibility (Donaldson *et al.,* 2006; Mullarkey *et al.,* 2005). The fourth most associated cluster, DRB1*0404 (OR = 1.57, P = 1.22×10^{-9}) has not been previously associated with PBC. The variance in disease liability explained by the 26 independent SNPs and four HLA-types are 4.9% and 1.4% respectively.

Haplotype	HLA type	Freq Cases	Freq Controls	OR	P-value
1	HLA*DQA1:0401	0.063	0.022	3.07	5.90×10^{-45}
	HLA*DQB1:0402	0.06	0.021	3.04	1.91×10^{-42}
	HLA*DRB1:0801	0.054	0.018	3.18	1.14×10^{-40}
	HLA*B:3905	0.01	0.003	5.48	4.81×10^{-12}
2	HLA*DQB1:0602	0.09	0.132	0.64	2.32×10^{-15}
	HLA*DRB1:1501	0.092	0.135	0.65	2.78×10^{-15}
	HLA*DQA1:0102	0.136	0.184	0.69	4.19×10^{-15}
	HLA*B:0702	0.109	0.144	0.73	4.93×10^{-10}
3	HLA*DQB1:0301	0.134	0.179	0.7	6.48×10^{-14}
	HLA*DRB1:1101	0.015	0.032	0.33	2.14×10^{-13}
	HLA*DQA1:0501	0.193	0.24	0.75	4.76×10^{-12}
	HLA*DRB1:1104	0.008	0.018	0.24	3.72×10^{-9}
4	HLA*DRB1:0404	0.072	0.052	1.57	1.22×10^{-9}
	HLA*DQB1:0302	0.133	0.104	1.34	6.96×10^{-9}

Table 2.4. Genome-wide significant HLA-type associations. Conditional analysis revealed four independent haplotypes.

2.3.5 Functional annotations and enrichment of open chromatin regions among risk loci

To identify candidate causal variants, I searched for non-synonymous variants in high LD ($r^2 > 0.8$) with the most associated variants at each PBC risk locus. I identified 39 such variants (of which 13 were directly genotyped) within seven risk loci (Figure 2.2), including two of the novel PBC associations identified in this study, *TYK2* and *SH2B3*. Functional follow-up studies are needed before

these non-synonymous variants can be confirmed as the causal disease variants at these loci.

As variation in gene expression is also likely to influence PBC risk, I evaluated the extent to which the most associated SNP at each locus tags expression quantitative trait loci (eQTLs) or regions of open chromatin. Known eQTLs were collated from the University of Chicago eQTL Browser (http://eqtl.uchicago.edu/cgi-bin/gbrowse/eqtl/) and Gaffney *et al*. (2012). Open chromatin regions in a range of cell lines were identified as part of the Encyclopedia of DNA Elements (ENCODE) project (Myers *et al.,* 2011; Song *et al.,* 2011) using DNase I hypersensitive sites sequencing (DNase-seq). Of the 26 independent non-HLA genome-wide significant SNPs identified in this study, 15 have an r^2>0.8 with SNPs that overlap DNase-seq peaks in a B-lymphoblastoid cell line (GM12878), and seven are also significant eQTLs in the same cell line (Figure 2.2).

To test if the enrichment of GM12878 open chromatin in regions was significantly greater than that for all other cell lines, associated SNPs were grouped into independent loci, and an enrichment score calculated for all loci that contained a genome-wide significant SNP (Section 2.2.8). Overall, GM12878 had the highest enrichment score compared with the other cell lines, though the difference in enrichment was non-significant (P = 0.068) (Figure 2.5).

The enrichment analysis protocol described here is predicated on the observation that the majority of complex disease risk loci do not lie within protein coding regions, and are likely to influence disease through their effects on gene expression, perhaps in a cell-specific manner. GWAS loci are indeed enriched for eQTLs (Nicolae *et al.,* 2010), though assigning causality to an associated variant based on eQTLs remains challenging due to LD and uncertainty over the precise regulatory mechanisms. Integrating functional genomic annotations may help bridge this gap between disease risk loci and eQTLs. Here, I used regions of open chromatin (as measured by DNase-I hypersensitivity) as it is a general indicator of potential regulatory activity (Bell *et al.,* 2011). These accessible regions make up 1-2% of the genome of a given

cell type, and are correlated with a range of other regulatory factors such as promoter and enhancer histone marks and transcription factor binding sites. Genetic variation in these regions have been shown to modify chromatin accessibility and transcription factor binding, which in turn lead to changes in gene expression (Degner *et al.,* 2012; Kasowski *et al.,* 2010). As such, variants within these regions that are in high LD with disease risk variants are good causal candidates, and enrichment in certain cell types may point to the relevant cells of interest in a particular disease.

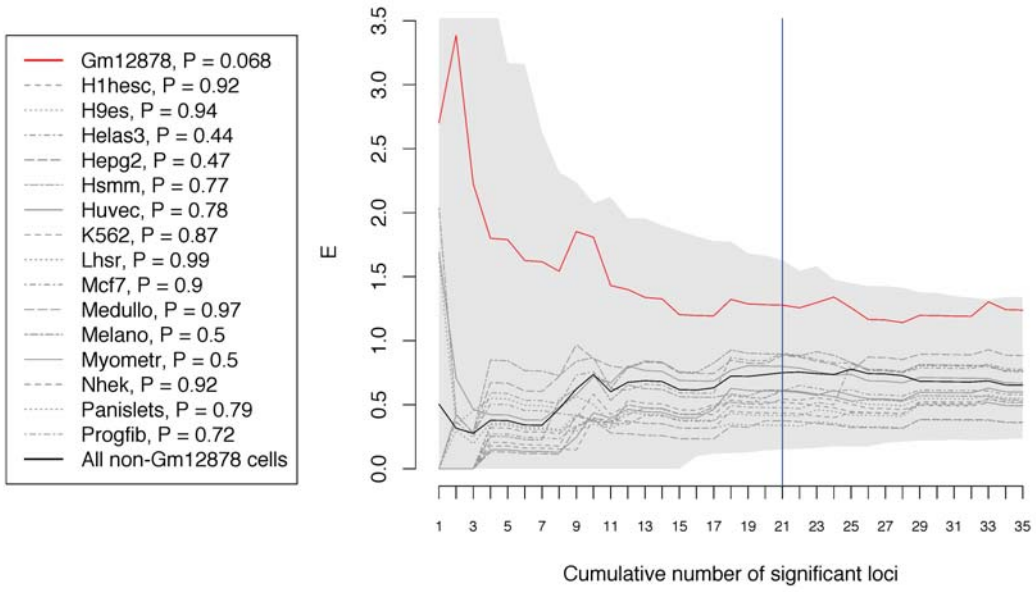


Figure 2.5. Enrichment of DNase-seq peaks among PBC risk loci in Gm12878 compared to other ENCODE cell lines. The relative enrichment (E) of SNPs within DNase-seq peaks was calculated across the 21 most associated loci. There is suggestive, though non-significant, evidence that genome-wide significant loci ($P < 5\times10^{-8}$ - vertical blue line) are more likely to lie within DNase-seq peaks in B-lymphoblastoid cell lines (solid red line) than they are to lie within the union of all other annotated cell lines (solid black line) (P = 0.068). Dotted grey lines denote E for other annotated cell lines. The shaded grey area represents the 95% confidence interval of E for Gm12878 from 1000 permutations. Cell types: Gm12878: B-lymphoblastoid, H1hesc: embryonic stem cells, H9es: embryonic stem cells, Helas3: cervical carcinoma, Hepg2: liver carcinoma, Hsmm: skeletal muscle myoblasts, Huvec: umbilical vein endothelial cells, K562: leukemia, Lhsr: prostate epithelial cells, Mcf7: mammary gland adenocarcinoma, Medullo: medulloblastoma, Melano: epidermal melanocytes, Myometr: Myometrial cells, Nhbe: bronchial epithelial cells, Nhek: epidermal keratinocytes, Panislets: pancreatic islets, Progfib: fibroblasts.

Fifteen of the 25 non-HLA PBC risk loci overlap regions of open chromatin in the GM12878 B-lymphoblastoid cell lines. While this number appears not to be significant when compared with the other ENCODE cell lines (P = 0.068), as a classical autoimmune disorder with a well-defined antibody presence (Jones, 2003), PBC risk is likely to be influenced by B cell activity. Moreover, it is important in these types of analyses not to bias results due to LD. Had I naively performed the enrichment analysis based on association P-value thresholds rather than pre-binning SNPs into independent loci, the evidence for enrichment with GM12878 open chromatin would have been much stronger (P = 0.0012) (Figure 2.6). This P-value enrichment approach, while seen in other studies (Maurano *et al.,* 2012), has the potential to bias results when SNPs that are moderately correlated with each other are counted multiple times if they overlap with functional annotations. In contrast, the approach presented here only considers a single potential causal variant per locus (i.e. SNPs that are in LD with the most strongly associated SNP is removed from further consideration), with all other variants in moderate LD are excluded from further analysis.

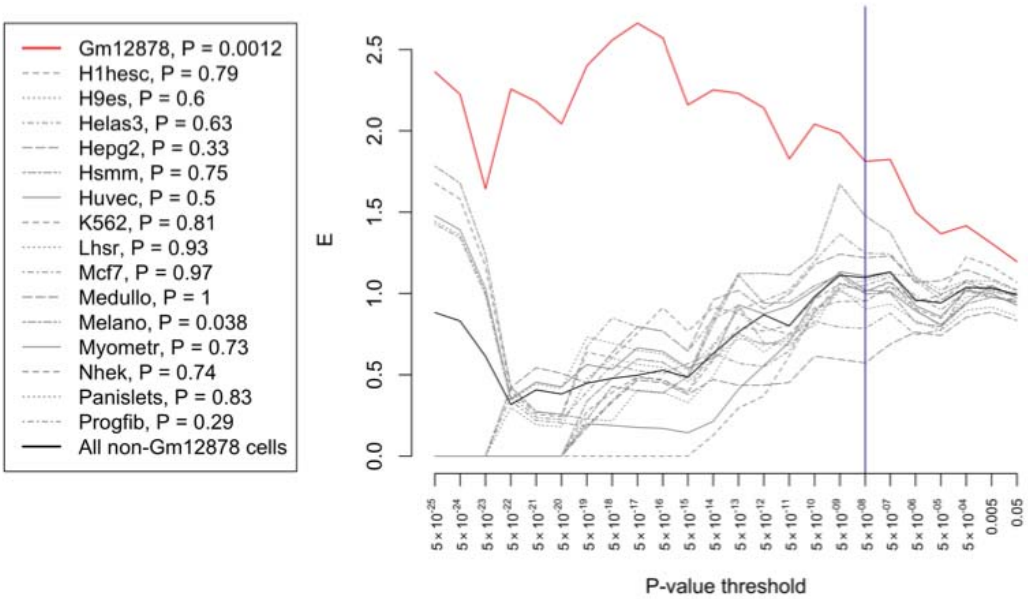


Figure 2.6. Enrichment of DNase-seq peaks among PBC risk loci calculated from P-value bins.

The thresholds used to define causal candidates at associated loci ($r^2 > 0.8$ with most associated SNP) as well as SNPs in LD to exclude ($r^2 > 0.1$) are somewhat subjective choices. The $r^2 > 0.8$ cut-off is based on a ubiquitous definition for which a SNP is "tagged" (Wang *et al.*, 2005), and is used throughout this chapter and the remainder of this thesis when defining causal candidate SNPs. This threshold has previously been shown to be effective at trading off power and the number of SNPs that need to be genotyped (de Bakker *et al.*, 2005), assessing the coverage of genotyping arrays (Barrett and Cardon, 2006) and resolving haplotypes (Carlson *et al.*, 2004). The threshold of $r^2 > 0.1$ to remove SNPs in LD with the most associated SNP in a locus (the lead SNP) was used to ensure that genome-wide significant SNPs whose signals are driven by their moderate LD with a much more strongly associated SNP are not considered causal candidates, while also allowing for additional truly independent variants to be counted. This approach yielded 21 genome-wide significant loci (Figure 2.5). Of the 26 independently associated SNPs identified in this study (Figure 2.2), 22 reside in the Immunochip high density regions considered in this enrichment analysis. The one ostensibly independently associated SNP that was excluded (rs668988) had $r^2 = 0.32$ with another lead SNP, rs2366643. Raising the r^2 threshold of 0.1 would have meant more SNPs denoted as "independently loci" even if their signals were entirely driven by a more strongly associated SNP nearby. For instance, at the rs72678531 ($P = 2.47\times10^{-36}$) locus, a second genome-wide significant SNP (rs17129749; $P = 3.50\times10^{-8}$) would have been declared independent had a minimum r^2 threshold of 0.3 been used. On the other hand, a lower threshold may exclude truly independently associated SNPs. Overall, any choice of LD thresholds involves trade-offs between excluding SNPs in LD and capturing truly independent association signals.

Finally, it should be noted that enrichment in a certain cell type does not automatically implicate that cell in disease. A certain amount of enrichment may be expected given that many gene promoters are active across multiple cell types (e.g. housekeeping genes). Between two given cell types, 30-40% of open chromatin regions may be shared (Song *et al.,* 2011). Moreover, lack of enrichment cannot rule out that cell's involvement in disease. This study was

also limited by the availability of cell types where the same functional genomic annotations were obtained in a consistent manner. It is likely that similar studies in autoimmune disorders will incorporate annotations from a range of immune cells (e.g. various types T cells, monocytes, dendritic cells, macrophages). For instance, recent approaches examining gene expression in murine immune cells found significant enrichment for B cell expressed genes among systemic lupus erythematosus risk loci, CD4 T cell genes among rheumatoid arthritis loci, and dentritic cell genes among Crohn's disease loci (Hu *et al.,* 2011; Jostins *et al.,* 2012). Moreover, the power to detect enrichment will only increase as the list of associated risk loci ever expands.

2.4 Conclusion

Through genotyping of 2,861 PBC cases and 8,514 controls on the Immunochip genotyping array, three novel PBC risk loci were identified, including a low-frequency non-synonymous SNP in *TYK2*, further implicating the JAK-STAT and cytokine signalling in disease pathogenesis. Together, these newly discovered risk loci in conjunction with 16 previously known loci offer further leads into the biological pathways that underlie PBC risk.

Within the 186 high density regions, the Immunochip includes ~90,000 directly genotyped SNPs compared with ~10,000 SNPs on the Illumina Human-660W Quad array used in Mells *et al.*, (2011). This denser coverage suggests that common causal variants are more likely to be genotyped directly or offer better tagging than SNPs from GWAS arrays. Reassuringly, odds ratios at known loci did not significantly differ to those from Mells *et al.*, (2011) despite the lead SNP changing at all but five of these loci. This suggests that the loci discovered in this study were primarily driven by sample size rather than SNP density, and that further GWAS of ever larger sample sizes will continue to discovery new risk loci.

The dense coverage also allows for greater refinement of the genetic architecture at risk loci. Multiple independent association signals were identified at five loci, including low-frequency and rare variants that are poorly tagged on GWAS arrays. At the 3q25 locus, four independent signals were identified,

though the effect sizes at two of the SNPs varied significantly when assessed under a joint model than when considering SNPs one at a time, highlighting that the haplotype structure of these regions with multiple signals should be considered when reporting association results.

Finally, I also explored the potential of integrating association results with large-scale functional genomic annotations to identify the cell types in which PBC associated variants are likely to be influencing disease. Future association studies in larger sample sizes in combination with disease-relevant functional genomic datasets will greatly improve the understanding of PBC and other complex disorders.