Nat Genet. Author manuscript; available in PMC 2013 April 01.

Published in final edited form as:
Nat Genet. 2012 October ; 44(10): 1137–1141. doi:10.1038/ng.2395.

Dense fine-mapping study identifies new susceptibility loci for primary biliary cirrhosis

Jimmy Z Liu[1,12], **Mohamed A Almarri**[1,12], **Daniel J Gaffney**[1], **George F Mells**[2,3], **Luke Jostins**[1], **Heather J Cordell**[4], **Samantha J Ducker**[5], **Darren B Day**[2], **Michael A Heneghan**[6], **James M. Neuberger**[7], **Peter T Donaldson**[5], **Andrew J Bathgate**[8], **Andrew Burroughs**[9], **Mervyn H Davies**[10], **David E Jones**[5], **Graeme J Alexander**[3], **Jeffrey C Barrett**[1], **The UK PBC Consortium**[11], **The Wellcome Trust Case Control Consortium 3**[11], **Richard N Sandford**[2,13], and **Carl A Anderson**[1,13]

[1]Wellcome Trust Sanger Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge, UK.

[2]Academic Department of Medical Genetics, Cambridge University, Cambridge, UK.

[3]Department of Hepatology, Cambridge University Hospitals National Health Service (NHS) Foundation Trust, Cambridge, UK.

[4]Institute of Human Genetics, Newcastle University, Newcastle upon Tyne, UK.

[5]Institute of Cellular Medicine, Medical School, Newcastle University, Newcastle upon Tyne UK.

[6]Institute of Liver Studies, King's College Hospital NHS Foundation Trust, Denmark Hill, London, UK.

[7]The Liver Unit, Queen Elizabeth Hospital, Birmingham, UK.

[8]Scottish Liver Transplant Unit, Royal Infirmary of Edinburgh, Edinburgh, UK.

[9]Hepatology Department, Royal Free Campus, University College London Medical School, London, UK.

[10]The Liver Unit, St James's University Hospital, Leeds, UK

Abstract

We genotyped 2,861 cases from the UK PBC consortium and 8,514 UK population controls across 196,524 variants within 186 known autoimmune risk loci. We identified three loci newly associated with primary biliary cirrhosis (PBC) (with $P<5\times10^{-8}$), increasing the number of known susceptibility loci to 25. The most associated variant at 19p12 is a low-frequency non-synonymous SNP in *TYK2*, further implicating JAK/STAT and cytokine signalling in disease pathogenesis. A further five loci contained non-synonymous variants in high linkage disequilibrium (LD) ($r^2>0.8$)

Correspondence should be addressed to C.A.A. (carl.anderson@sanger.ac.uk). .
[11]A full list of members and affiliations is provided in the Supplementary Note.
[12]These authors contributed equally to this work.
[13]These authors jointly directed this work.

URLs University of Chicago eQTL Browser, http://eqtl.uchicago.edu/cgi-bin/gbrowse/eqtl/

Author contributions Study concept and design: D.J.G., G.F.M., L.J., H.J.C., M.A.H., J.M.N., P.T.D., D.E.J., G.J.A., A.J.B., A.B., M.H.D., J.C.B., The WTCCC3 management committee (see Supplementary Note), R.N.S., C.A.A.
Case ascertainment and phenotyping: S.J.D., D.B.D. and The UK PBC Consortium (see Supplementary Note)
Control sample ascertainment: The UK Blood Service Controls group (see Supplementary Note), The 1958 Birth Cohort Controls group (see Supplementary Note).
Genotyping: The WTCCC3 DNA, Genotyping and Informatics group (see Supplementary Note).
Statistical analysis: J.Z.L., M.A.A., D.J.G., L.J., The WTCCC3 data analysis group (see Supplementary Note), C.A.A.
Manuscript preparation: J.Z.L., M.A.A., D.J.G., C.A.A. All authors reviewed the final manuscript.

Europe PMC Funders Author Manuscripts

with the most associated variant at the locus. We found multiple independent common, low-frequency and rare variant association signals at five loci. Of the 26 independent non-HLA signals tagged on Immunochip, 15 have SNPs in B-lymphoblastoid open-chromatin regions in high LD (r^2>0.8) with the most associated variant. This study demonstrates how dense fine-mapping arrays coupled with functional genomic data can be utilized to identify candidate causal variants for functional follow-up.

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[1]. 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 (AI) liver disorder[1,2]. Family-based studies indicate a substantial genetic component to PBC susceptibility, with a sibling relative risk of ~10.5 in the UK[3]. Genome-wide association studies (GWAS) have identified 22 PBC risk loci, and highlighted the role of NFkB signaling, T-cell differentiation, Toll-like receptor and tumor necrosis factor signaling in disease pathogenesis[4-6]. 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[7]. Despite these advances, the specific causal variant at many of these loci remains unknown.

To better define risk variants and identify additional susceptibility loci, we performed a fine-mapping and association study using a cohort of 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 on the Immunochip, an Illumina Infinium array containing 196,524 variants (718 small insertions/deletions and 195,806 SNPs) across 186 known AI risk loci. SNPs were derived from population-based sequencing projects such as the 1000 Genomes project and autoimmune disease resequencing efforts[8,9]. Compared to GWAS arrays, Immunochip's increased marker density within known AI loci increases power to detect PBC associations within these selected key candidate genes, and provides a powerful means of fine-mapping known PBC loci as causal variants are more likely to be directly genotyped.

Following quality control (Online Methods), 143,020 polymorphic SNPs were available across 2,861 cases and 8,514 controls. (Supplementary Tables 1-2, Supplementary figures 1-6). A further 94,559 SNPs in the Immunochip fine-mapping regions were imputed using genotypes from the 1000 Genomes Project June 2011 release (Online Methods). The inflation factor inferred from 2,258 SNPs not associated with autoimmune disease showed only a modest inflation (λ=1.096, Online Methods), similar to that reported in our previous GWAS for PBC [6].

Of the 22 known PBC risk loci, 16 reached genome-wide significance ($P<5\times10^{-8}$) (Figure 1) and four showed nominal evidence of association ($5\times10^{-8}<P<5\times10^{-4}$) (Supplementary Table 3). Two PBC loci, 14q32 and 19q13, were not included on Immunochip as the array was designed before the publication of the most recent PBC GWAS[6]. At 12 of the genome-wide significant loci, the most associated SNP was different to that previously reported (Supplementary Table 3). There was little difference in the effect-size estimates between the GWAS tagging SNP and the most strongly associated Immunochip SNP (Supplementary Figure 7), although this may be due to a large proportion of overlapping samples between the two studies (Online Methods).

Stepwise conditional regression[10] revealed multiple independent signals at five loci, with 16p13 harboring three, and 3q25 four such associations (Table 1). At the 16p13 locus, the third independent signal, rs80073729, is a rare SNP (MAF<0.5%) recently associated with

celiac disease[9]. In the same study, Trynka et al.[9] 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. These results suggest that resequencing hundreds or thousands of cases across known GWAS loci will be a powerful means of identifying additional independent risk alleles. It is likely that these two rare SNP associations would have been missed using standard GWAS arrays due to poor tagging, unless they were directly genotyped. As these are rare SNPs, further replication in large independent cohorts will be required to confirm their associations. Haplotype association analysis at loci with multiple independent signals identified similar effect-size estimates suggesting that the causal variant is among, or is highly correlated with, genotyped SNPs (Supplementary Table 4). These additional independent association signals thus yield a more complete understanding of the genetic architecture of PBC and enable more informative genotype-based recall studies to be conducted.

Variants at three loci not previously reported as associated with PBC reached genome-wide significance threshold (Table 1). The most significant association on 19p12, rs34536443 (OR=1.91, $P=1.24\times10^{-12}$), is a low-frequency ($1\% \leq MAF<5\%$) non-synonymous SNP in the tyrosine kinase 2 gene (*TYK2*), previously associated with multiple sclerosis[11]. The locus has also been associated with T1D[12], psoriasis[13] and Crohn's disease[14], 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[9,15], rheumatoid arthritis[16] and T1D[17], 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, $P=2.87\times10^{-9}$), that was most strongly associated; The two SNPs are in high LD ($r^2 = 0.81$) and further studies are required to identify the causal variant underlying the PBC association signal at this locus. The most associated SNP in the 17q21 region, rs17564829 (OR=1.25, $P=1=2.15\times10^{-9}$), is located in *MAPT*, a gene that has been associated with cognitive symptoms in Parkinson's disease. While cognitive symptoms are a major part of the symptom complex associated with PBC, it remains to be seen if a) the true causal variant at the locus has its functional effect through *MAPT*, and b) if this functional effect then results in cognitive changes in individuals with PBC.

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[18,19]. *TYK2* is a member of the Janus kinase family, which transduce cytokine signals by phosphorylating STAT transcription factors. Couturier et al.[20] showed that heterozygotes for rs34536443 have significantly reduced *TYK2* activity, which promotes the secretion of Th2 cytokines[20]. 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[21], suggesting a possible role in helping prevent bacterial infection.

Candidate genes studies have implicated several HLA-DR alleles in PBC susceptibility, particularly the DRB1*08 allele[22-25]. To further characterize HLA risk variants, the classical HLA alleles (A, B, C, DQA1, DQB1 and DRB1) were imputed from genotyped SNPs in the MHC[26,27] (Online Methods). Fourteen HLA-alleles reached genome-wide significance and conditional analysis clustered these associations into four independent signals (Supplementary Table 5, Supplementary Figure 8). The most significant association was the HLA-DQA1*0401 allele (OR=3.06, $P =5.9\times10^{-45}$), which forms a haplotype with two other HLA class II alleles (DQB1*0402 and DRB1*0801) and is an established PBC risk locus[22-25]. 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 [22,23]. 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 liability explained by the 26 independent SNPs and four HLA-types are 4.9% and 1.4% respectively, which together account for 16.2% of the total PBC heritability of liability of 0.39 (Online Methods).

To identify candidate causal variants we searched for non-synonymous variants in high LD (r^2>0.8) with the most associated variants at each PBC risk locus. We found 39 such variants (of which 13 were directly genotyped) within seven risk loci (Table 1 and Supplementary table 6), including two variants at two of the loci newly associated with PBC in this study, *TYK2* and *SH2B3*. Functional follow-up studies are needed before these non-synonymous variants can be confirmed as the causal variants at these loci. As variation in gene expression is also likely to influence PBC risk, we evaluated the extent to which the most associated SNP at each locus tags expression quantitative trait loci (eQTLs) or regions of open chromatin. Regions of open chromatin are associated with gene regulatory elements including promoters, enhancers, silencers, insulators and locus control regions. Known eQTLs were collated from the University of Chicago eQTL (see URLs) Browser and Gaffney *et al.* (2012)[28]. Open chromatin regions in a range of cell lines were identified as part of the Encyclopedia of DNA Elements (ENCODE) project[29,30] using DNase I hypersensitive sites sequencing (DNase-seq) and formaldehyde-assisted isolation of regulatory elements sequencing (FAIRE-seq).

Of the 26 independent non-HLA genome-wide significant SNPs identified in this study, 15 have an r^2>0.8 with SNPs in DNase-seq or FAIRE-seq peaks in a B-lymphoblastoid cell line, and seven are also significant eQTLs in the same cell line (Table 1 and Supplementary Table 7-8). To test if the enrichment of open chromatin within the B-lymphoblastoid cell line was significantly greater than that for all other cell lines we began by grouping SNPs into independent loci. We sequentially identified the most associated SNP not already assigned to a locus and assigned this SNP, and others in weak LD (r^2>0.1) with it, to a new locus. We then calculated an enrichment score, E (Online Methods), using only candidate causal variants (r^2>0.8 to the most associated SNP in each locus) across all currently assigned loci. Considering only loci where the most associated SNP achieved genome-wide significance (N=21, excluding the HLA locus, SNPs outside Immunochip fine-mapping regions and SNPs with MAF<5%), Gm12878 had the highest enrichment score compared with the other cell lines, though the difference in enrichment just failed to reach significance (P = 0.068 Online Methods) (Figure 2). Failure to correctly account for LD between associated SNPs can bias the calculated degree of enrichment (Supplementary Figure 9). Our enrichment analysis protocol can be applied to other functional annotations and other disease phenotypes and will be well powered for traits with many genome-wide significant associations.

In summary, we have used dense genotyping across autoimmune disease associated loci to better define the genetic architecture of known PBC risk loci. We have identified additional independent genome-wide significant associations at five loci, and have identified potentially causal protein-coding and regulatory variants within many disease associated loci. We also identified three new PBC risk loci, bringing the total number of associated loci to 25, and confirmed HLA-allele associations by imputing HLA-types. Furthermore, we have combined our SNP data with large-scale functional genomics annotations to identify the cell types in which the PBC associated variants are likely to be acting.

Online Methods

Ethical approval

This study was approved by the Research and Development Departments of all National Health Service (NHS) Trusts participating in this study and by the Oxford Research Ethics Committee C (Oxford REC C reference 07/H0606/96).

Samples

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. This study contains 1,838 cases and 2,356 controls included in our recent PBC GWAS[6].

DNA extraction

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.

Genotyping

Samples were genotyped on an Illumina iSelect HD custom genotyping array (Immunochip). All 2,981 cases and 4537 controls were genotyped at the Wellcome Trust Sanger Institute. A further 4433 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[31]. 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[32] and GenoSNP[33].

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 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[34] (Supplementary Figures 1-3). Following sample QC 2,861 cases and 8,514 controls remained (Supplementary Table 1). SNPs with a minor allele frequency

less than 0.1%, Hardy-Weinberg equilibrium $P < 10^{-6}$, 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 (Supplementary Table 2).

Statistical Methods

Genomic Inflation factor—The Immunochip contains 2,258 SNPs that lie in regions associated with bipolar disease. These were used as null markers to estimate the overall inflation of the distribution of association test statistics[35].

Imputation—Using the 90,977 SNPs from the cleaned Immunochip set that were in fine-mapped regions, additional genotypes were imputed using the 1000 Genomes Phase I (interim) June 2011 release reference panel and IMPUTE2[36]. Imputation was performed separately in three batches of 3792, 3792 and 3791 individuals, with the case:control ratio constant across batches. SNPs with a posterior probability less than 0.9 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.

Association analysis—Case-control association tests were implemented using a standard one-degree of freedom Cochran-Armitage test for trend in PLINK v1.07[37]. 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[10]. Haplotype association was performed in PLINK using logistic regression. Cluster plots for all SNPs $P<5\times10^{-6}$ were manually checked using Evoker[38], and poorly called SNPs were removed from further study (Supplementary Figure 11).

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[26,27]. Case-control association was performed on HLA allele posterior probabilities generated from HLA*IMP using logistic regression to account for genotype uncertainty following imputation. Pairwise conditional logistic regression was used to identify independent association signals among the 21 HLA-alleles that reached $P < 0.0001$.

Heritability explained—The heritability explained by the 26 independent genome-wide significant SNPs and four HLA-alleles was estimated using a liability threshold model[39,40] assuming a disease prevalence of 40/100,000, log-additive risk and a sibling relative risk ratio of 10.5[3].

eQTL analysis—eQTLs within genome-wide significant loci were collated from the University of Chicago eQTL Browser (see URLs) and a study by Gaffney et al., (2012)[28]. 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)[28] and references listed in the Chicago eQTL Browser (see URLs).

Enrichment of open chromatin regions—The ENCODE project annotated regions of open chromatin using two techniques, the direct sequencing of DNaseI cleavage sites

(DNase-seq: sixteen different cell lines) and formaldehyde-assisted isolation of regulatory elements sequencing (FAIRE-seq: ten different cell lines)[29,30]. Both methods isolate nucleosome-depleted regions of DNA and map reads from next-generation sequencing to determine their location. The overlap of peaks between the two assays ranges from 30-40% depending on the cell type, and regions identified uniquely by DNase-seq or FAIRE-seq often represent relevant biological processes[29]. Positions of discrete DNase-seq and FAIRE-seq peaks were estimated from the base overlap signal (BOS) at each base-pair[29]. We quantified the evidence for the open chromatin peaks using a Poisson distribution where lambda equals the mean BOS across all Immunochip SNPs. Supplementary Figure 12 shows the relative position of open chromatin peaks and associated SNPs within significantly associated loci.

For both DNase-seq and FAIRE-seq data, we estimated the amount of enrichment for open chromatin peaks among significant loci across the ENCODE cell lines. SNPs were first grouped into independent loci; we sequentially identified the most associated SNP not already assigned to a locus and assigned this SNP, and others in weak LD with it ($r^2>0.1$), to a new locus. After the addition of each new locus, we 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 ($r^2>0.8$ to the most associated SNP) that lie within open chromatin peaks across the selected loci and the total number of SNPs within the loci, respectively. OC_{ichip} and N_{ichip} are the equivalent measures across all SNPs within Immunochip fine mapping regions We only included the fine-mapping 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. The OC_{ichip} values for each of the cell lines are given in Supplementary Table 9. 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. We 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 chi-square test for the difference in proportions.

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

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

The PBC sample collection was funded by the Isaac Newton Trust, the PBC Foundation, The Addenborooke's Charitable Trust and the Wellcome Trust (085925/Z/08/Z). The PBC Genetics Study is a portfolio study of the National Institute for Health Research Comprehensive Clinical Research Network (NIHR CRN, portfolio reference 5630). The project is also supported by the Wellcome Trust (WT090355/A/09/Z, WT090355/B/09/Z, 098051). Genotyping of samples at the University of Virginia utilized resources provided by the Type 1 Diabetes Genetics

Consortium, a collaborative clinical study sponsored by the National Institute of Diabetes and Digestive and Kidney Diseases, the National Institute of Allergy and Infectious Diseases, the National Human Genome Research Institute, the National Institute of Child Health and Human Development, and the Juvenile Diabetes Research Foundation International and is supported by U01-DK-062418. We would like to thank the UK Medical Research Council and Wellcome Trust for funding the collection of DNA for the British 1958 Birth Cohort (MRC grant G0000934, WT grant 068545/Z/02). We acknowledge use of DNA from The UK Blood Services collection of Common Controls (UKBS collection), funded by the Wellcome Trust grant 076113/C/04/Z, by the Wellcome Trust/Juvenile Diabetes Research Foundation grant 061858, and by the National Institute of Health Research of England. The collection was established as part of the Wellcome Trust Case-Control Consortium. G.F.M. is a Clinical Research Training Fellow of the Medical Research Council (G0800460). G.F.M. is also supported by a Raymond and Beverly Sackler Studentship.

We are grateful to the PBC Foundation for helping us to establish the PBC Genetics Study, for endorsing it, and for encouraging members of the Foundation to contribute samples. We thank all of the research nurses who assisted with participant recruitment in collaborating centers. We thank the staff in the NIHR CRN and Clinical Research Collaboration (CRC) Cymru for providing invaluable support. We are grateful to K. Chittock and his colleagues at Source Bioscience for performing DNA extraction. We thank O. Burren for designing the participant database and for providing information technology support. We are grateful to Alexander Dilthey for providing support regarding HLA*IMP. We thank J. Stone for coordinating the Immunochip design and production at Illumina. We also thank the members of each disease consortium who initiated and sustained the cross-disease Immunochip project and shared control genotypes. Finally, we thank the individuals who contributed the DNA samples used in this study.

References

1. Kaplan MM, Gershwin ME. Primary biliary cirrhosis. N Engl J Med. 2005; 353:1261–73. [PubMed: 16177252]
2. James OF, et al. Primary biliary cirrhosis once rare, now common in the United Kingdom? Hepatology. 1999; 30:390–4. [PubMed: 10421645]
3. Jones DE, Watt FE, Metcalf JV, Bassendine MF, James OF. Familial primary biliary cirrhosis reassessed: a geographically-based population study. J Hepatol. 1999; 30:402–7. [PubMed: 10190721]
4. Hirschfield GM, et al. Primary biliary cirrhosis associated with HLA, IL12A, and IL12RB2 variants. N Engl J Med. 2009; 360:2544–55. [PubMed: 19458352]
5. Liu X, et al. Genome-wide meta-analyses identify three loci associated with primary biliary cirrhosis. Nat Genet. 2010; 42:658–60. [PubMed: 20639880]
6. Mells GF, et al. Genome-wide association study identifies 12 new susceptibility loci for primary biliary cirrhosis. Nat Genet. 2011; 43:329–32. [PubMed: 21399635]
7. Zhernakova A, van Diemen CC, Wijmenga C. Detecting shared pathogenesis from the shared genetics of immune-related diseases. Nat Rev Genet. 2009; 10:43–55. [PubMed: 19092835]
8. Cortes A, Brown M. Promise and pitfalls of the Immunochip. Arthritis Res Ther. 2011; 13:101. [PubMed: 21345260]
9. Trynka G, et al. Dense genotyping identifies and localizes multiple common and rare variant association signals in celiac disease. Nat Genet. 2011; 43:1193–201. [PubMed: 22057235]
10. Cordell HJ, Clayton DG. A unified stepwise regression procedure for evaluating the relative effects of polymorphisms within a gene using case/control or family data: application to HLA in type 1 diabetes. American journal of human genetics. 2002; 70:124–41. [PubMed: 11719900]
11. Ban M, et al. Replication analysis identifies TYK2 as a multiple sclerosis susceptibility factor. Eur J Hum Genet. 2009; 17:1309–1313. [PubMed: 19293837]
12. Wallace C, et al. The imprinted DLK1-MEG3 gene region on chromosome 14q32.2 alters susceptibility to type 1 diabetes. Nat Genet. 2010; 42:68–71. [PubMed: 19966805]
13. Strange A, et al. A genome-wide association study identifies new psoriasis susceptibility loci and an interaction between HLA-C and ERAP1. Nat Genet. 2010; 42:985–90. [PubMed: 20953190]
14. Franke A, et al. Genome-wide meta-analysis increases to 71 the number of confirmed Crohn's disease susceptibility loci. Nat Genet. 2010; 42:1118–25. [PubMed: 21102463]
15. Hunt KA, et al. Newly identified genetic risk variants for celiac disease related to the immune response. Nat Genet. 2008; 40:395–402. [PubMed: 18311140]

16. Stahl EA, et al. Genome-wide association study meta-analysis identifies seven new rheumatoid arthritis risk loci. Nat Genet. 2010; 42:508–514. [PubMed: 20453842]
17. Barrett JC, et al. Genome-wide association study and meta-analysis find that over 40 loci affect risk of type 1 diabetes. Nat Genet. 2009; 41:703–707. [PubMed: 19430480]
18. Wang D, et al. CD4+CD25+ but not CD4+Foxp3+ T cells as a regulatory subset in primary biliary cirrhosis. Cell Mol Immunol. 2010; 7:485–490. [PubMed: 20729906]
19. Rong G, et al. Imbalance between T helper type 17 and T regulatory cells in patients with primary biliary cirrhosis: the serum cytokine profile and peripheral cell population. Clin Exp Immunol. 2009; 156:217–225. [PubMed: 19302244]
20. Couturier N, et al. Tyrosine kinase 2 variant influences T lymphocyte polarization and multiple sclerosis susceptibility. Brain. 2011; 134:693–703. [PubMed: 21354972]
21. Zhernakova A, et al. Evolutionary and Functional Analysis of Celiac Risk Loci Reveals SH2B3 as a Protective Factor against Bacterial Infection. Am J Hum Genet. 2010; 86:970–977. [PubMed: 20560212]
22. Donaldson PT, et al. HLA class II alleles, genotypes, haplotypes, and amino acids in primary biliary cirrhosis: a large-scale study. Hepatology. 2006; 44:667–74. [PubMed: 16941709]
23. Mullarkey ME, et al. Human leukocyte antigen class II alleles in Caucasian women with primary biliary cirrhosis. Tissue Antigens. 2005; 65:199–205. [PubMed: 15713222]
24. Wassmuth R, et al. HLA class II markers and clinical heterogeneity in Swedish patients with primary biliary cirrhosis. Tissue Antigens. 2002; 59:381–7. [PubMed: 12144621]
25. Invernizzi P, et al. Human leukocyte antigen polymorphisms in Italian primary biliary cirrhosis: a multicenter study of 664 patients and 1992 healthy controls. Hepatology. 2008; 48:1906–12. [PubMed: 19003916]
26. Leslie S, Donnelly P, McVean G. A statistical method for predicting classical HLA alleles from SNP data. Am J Hum Genet. 2008; 82:48–56. [PubMed: 18179884]
27. Dilthey AT, Moutsianas L, Leslie S, McVean G. HLA*IMP--an integrated framework for imputing classical HLA alleles from SNP genotypes. Bioinformatics. 2011; 27:968–72. [PubMed: 21300701]
28. Gaffney DJ, et al. Dissecting the regulatory architecture of gene expression QTLs. Genome biology. 2012; 13:R7. [PubMed: 22293038]
29. Song L, et al. Open chromatin defined by DNaseI and FAIRE identifies regulatory elements that shape cell-type identity. Genome Res. 2011; 21:1757–67. [PubMed: 21750106]
30. Myers RM, et al. A user's guide to the encyclopedia of DNA elements (ENCODE). PLoS Biol. 2011; 9:e1001046. [PubMed: 21526222]
31. Shah TS, et al. optiCall: A robust genotype-calling algorithm for rare, low frequency and common variants. Bioinformatics. 2012; 28:1598–1603. [PubMed: 22500001]
32. Teo YY, et al. A genotype calling algorithm for the Illumina BeadArray platform. Bioinformatics. 2007; 23:2741–6. [PubMed: 17846035]
33. Giannoulatou E, Yau C, Colella S, Ragoussis J, Holmes CC. GenoSNP: a variational Bayes within-sample SNP genotyping algorithm that does not require a reference population. Bioinformatics. 2008; 24:2209–2214. [PubMed: 18653518]
34. Price AL, et al. Principal components analysis corrects for stratification in genome-wide association studies. Nat Genet. 2006; 38:904–9. [PubMed: 16862161]
35. Devlin B, Roeder K. Genomic control for association studies. Biometrics. 1999; 55:997–1004. [PubMed: 11315092]
36. Howie BN, Donnelly P, Marchini J. A flexible and accurate genotype imputation method for the next generation of genome-wide association studies. PLoS genetics. 2009; 5:e1000529. [PubMed: 19543373]
37. Purcell S, et al. Tool Set for Whole-Genome Association and Population-Based Linkage Analyses. Am J Hum Genet. 2007; 81:559–575. [PubMed: 17701901]
38. Morris JA, Randall JC, Maller JB, Barrett JC. Evoker: a visualization tool for genotype intensity data. Bioinformatics. 2010; 26:1786–7. [PubMed: 20507892]

39. So HC, Gui AH, Cherny SS, Sham PC. Evaluating the heritability explained by known susceptibility variants: a survey of ten complex diseases. Genet Epidemiol. 2011; 35:310–7. [PubMed: 21374718]

40. Falconer, D.; Mackay, T. Introduction to Quantitative Genetics. Longman: 1996. p. 464

Europe PMC Funders Author Manuscripts

Europe PMC Funders Author Manuscripts

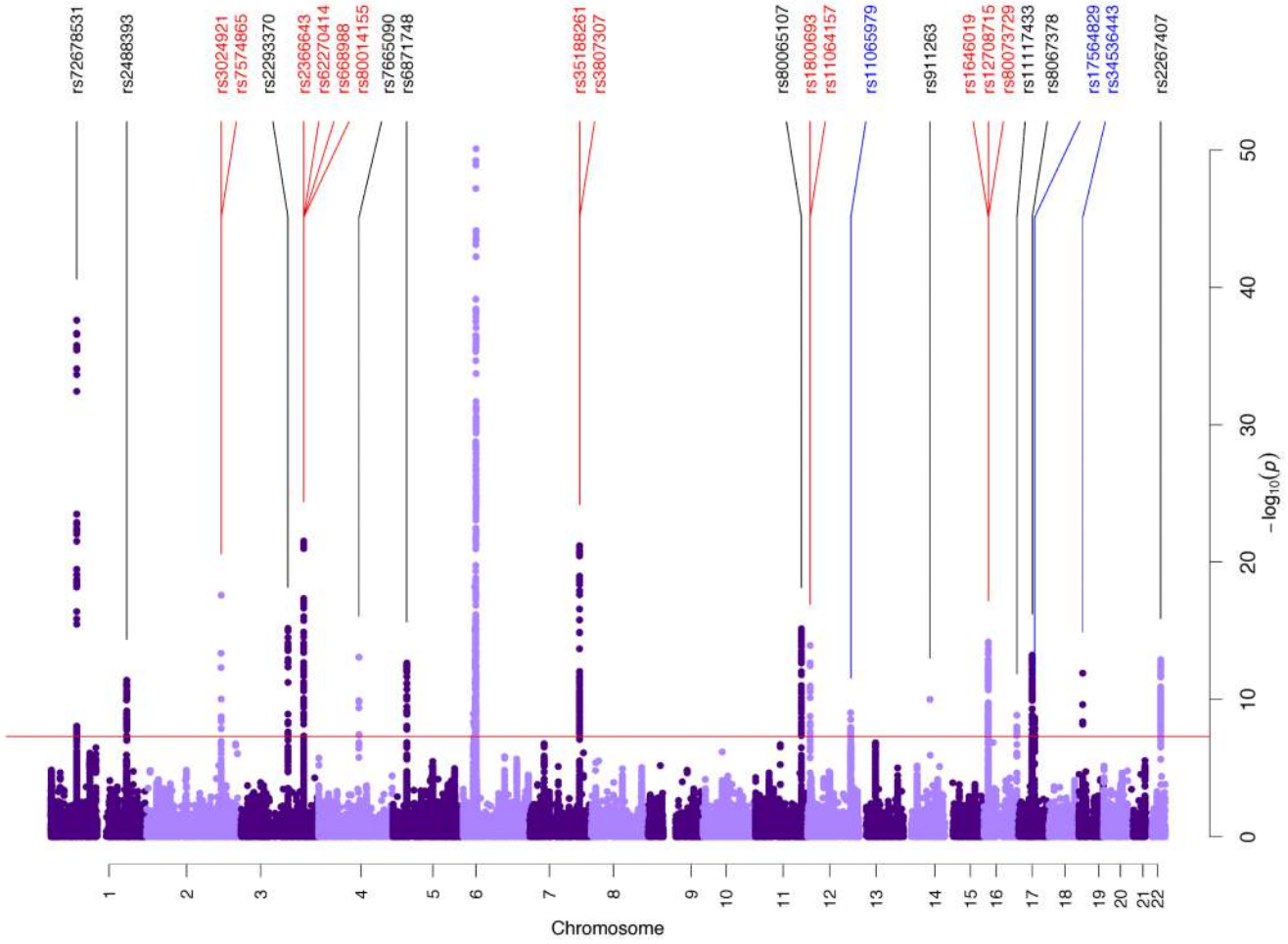

Figure 1. Manhattan plot and list of genome-wide significant PBC risk loci across Immunochip

Novel risk loci are highlighted in blue. Loci with more than one independent signal are highlighted in red. The vertical red line indicates the genome-wide significance threshold of $P=5\times10^{-8}$. The peak on chromosome 6 is the HLA region.

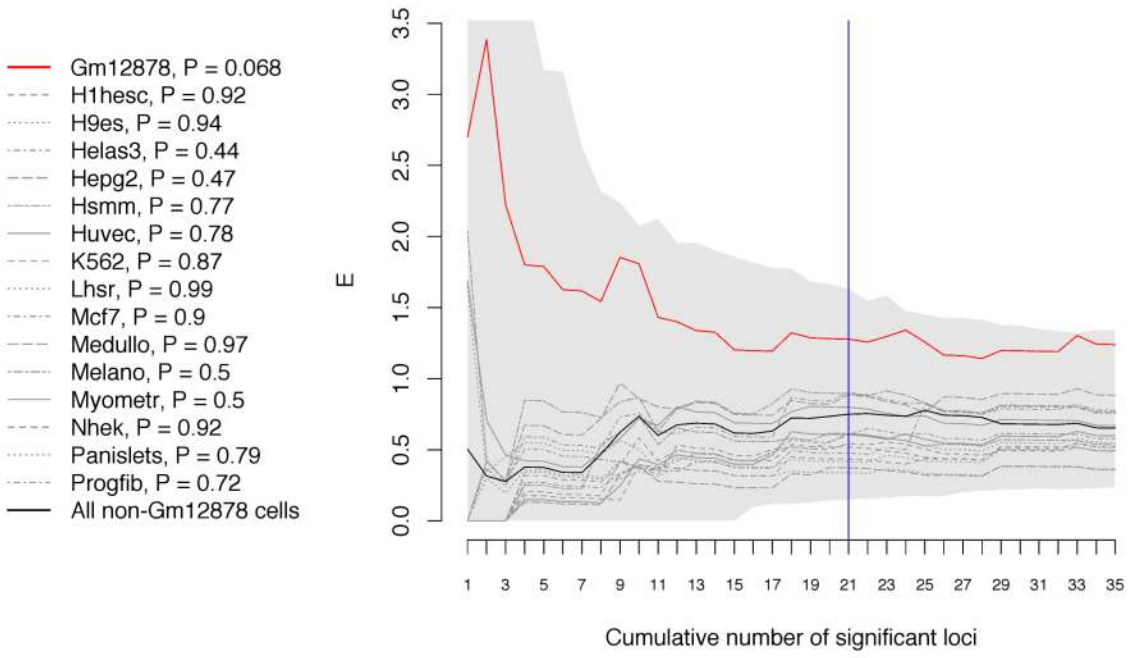

Figure 2. Enrichment of DNase-seq peaks among PBC risk loci in Gm12878 compared to other ENCODE cell lines

The relative enrichment (*E*) of SNPs within DNase-seq peaks was calculated across the 35 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.

Europe PMC Funders Author Manuscripts

Table 1

PBC risk loci at genome-wide significance

Chr	SNP[a]	RA[b]	RAF[c]	*p* [d]	OR (95% CI)	LD region[e] (size)	Nearby gene(s)[f]	Functional annotation[g]
1p31	rs72678531	G	0.17	2.47×10^{-38}	1.61 (1.49-1.73)	67,560,940-67,592,782 (31,842)	*Il12RB2*	OC
1q31	rs2488393	A	0.21	4.29×10^{-12}	1.28 (1.19-1.37)	195,609,003-196,047,821 (438,818)	*DENND1B*	
2q32	rs3024921	A	0.06	2.59×10^{-18}	1.62 (1.45-1.80)	191,651,517-191,651,517 (0)	*STAT1,STAT4*	
2q32	Second signal rs7574865	A	0.22	1.38×10^{-13}	1.31 (1.22-1.40)	191,651,987-191,681,279 (29,292)	*STAT1,STAT4*	
3q13	rs2293370	G	0.8	6.84×10^{-16}	1.39 (1.29-1.52)	120,598,840-120,734,898 (136,058)	*TMEM39A,POGLUT1,TIMMDC1,CD80*	NS
3q25	rs2366643	A	0.57	3.92×10^{-22}	1.35 (1.27-1.44)	161,202,965-161,219,770 (16,805)	*IL12A*	OC
3q25	Second signal rs62270414	G	0.15	5.74×10^{-17}	1.41 (1.30-1.53)	161,122,353-161,174,976 (52,623)	*IL12A*	OC
3q25	Third signal rs668998	G	0.43	4.73×10^{-9}	1.26 (1.17-1.36)	161,192,695-161,198,245 (5,550)	*IL12A*	OC
3q25	Fourth signal rs80014155	A	0.004	2.64×10^{-11}	3.44 (2.39-4.94)	161,108,087-161,176,747 (68,660)	*IL12A*	
4q24	rs7665090	G	0.52	8.48×10^{-14}	1.26 (1.19-1.34)	103,770,651-103,770,651 (0)	*MANBA,NFKB1*	NS,eQTL
5p13	rs6871748	A	0.72	2.26×10^{-13}	1.3 (1.21-1.4)	35,885,906-35,921,739 (35,833)	*IL7R,CAPSL,SPEF2,UGT3A1*	NS
7q32	rs35188261	A	0.17	6.52×10^{-22}	1.52 (1.39-1.63)	128,372,852-128,499,110 (126,258)	*IRF5,TNP03*	OC
7q32	Second signal rs3807307	G	0.47	4.12×10^{-9}	1.22 (1.14-1.30)	128,361,203-128,367,916 (6,713)	*IRF5,TNP03*	OC,eQTL
11q23	rs80065107	A	0.79	7.20×10^{-16}	1.39 (1.28-1.5)	118,115,759-118,248,982 (133,223)	*DDX6*	OC

Europe PMC Funders Author Manuscripts

Chr	SNP[a]	RA[b]	RAF[c]	p[d]	OR (95% CI)	LD region[e] (size)	Nearby gene(s)[f]	Functional annotation[g]
12p13	rs1800693	G	0.4	1.18×10^{-14}	1.27 (1.19-1.34)	6,310,270-6,323,072 (12,802)	*TNFRSF1A,LTBR,SCNN1A*	OC
12p13	Second signal rs11064157	A	0.25	1.69×10^{-9}	1.23 (1.15-1.32)	6,362,910-6,362,910 (0)	*TNFRSF1A,LTBR,SCNN1A*	OC
12q24	**rs11065979**	A	0.44	2.87×10^{-9}	1.2 (1.13-1.27)	110,368,991-111,095,097 (726,106)	*ATXN2,BRAP,SH2B3*	NS
14q24	rs911263	A	0.71	9.95×10^{-11}	1.26 (1.17-1.35)	67,823,346-67,823,346 (0)	*RAD51B*	
16p13	rs1646019	G	0.71	6.72×10^{-15}	1.31 (1.23-1.41)	11,254,549-11,273,001 (18,452)	*SOCS1,CLEC16A,PRM1,PRM2*	OC,eQTL
16p13	Second signal rs12708715	C	0.68	2.19×10^{-13}	1.29 (1.21-1.38)	10,999,820-11,117,948 (118,128)	*SOCS1,CLEC16A,PRM1,PRM2*	OC,eQTL
16p13	Third signal rs80073729	A	0.004	2.69×10^{-8}	2.96 (2.02-4.33)	11,281,298-11,281,298 (0)	*SOCS1,CLEC16A,PRM1,PRM2*	OC
16q24	rs11117433	G	0.77	1.41×10^{-9}	1.26 (1.17-1.36)	84,577,017-84,577,017 (0)	*IRF8*	OC
17q12	rs8067378	G	0.52	6.05×10^{-14}	1.26 (1.19-1.34)	35,158,633-35,336,333 (177,700)	*ORMDL3,ZPBP2,GSDMB,IKZF3*	NS,OC,eQTL
17q21	**rs17564829**	G	0.24	2.15×10^{-9}	1.25 (1.16-1.35)	41,047,160-42,211,804 (1,164,644)	*CRHR1,MAPT*	NS,OC,eQTL
19p12	**rs34536443**	G	0.95	1.23×10^{-12}	1.91 (1.59-2.28)	10,324,118-10,324,118 (0)	*TYK2*	NS
22q13	rs2267407	A	0.23	1.29×10^{-13}	1.29 (1.21-1.38)	38,076,996-38,086,596 (9,600)	*SYNGR1,PDGFB,RPL3*	OC,eQTL

Non HLA PBC risk loci meeting genome-wide significance ($P < 5 \times 10^{-8}$) are shown.

[a] Most significant SNP in the locus. Novel associations are highlighted in bold.

[b] Risk allele.

[c] Risk allele frequency in controls.

[d] For primary signals, p-values were obtained from Cochran-Armitage tests for trend. For second, third and fourth association signals, p-values were obtained using logistic regression conditioning on the previous independent SNPs.

Europe PMC Funders Author Manuscripts

[e] Regions of high linkage disequilibrium defined by SNPs with r2 > 0.8. See Supplementary Figure 10 for regional locus plots.

[f] RefSeq UCSC hg18 track.

[g] Denotes if there are SNPs with $r^2 > 0.8$ with the hit SNP that lie within OC (open chromatin peaks), are non-synonymous (NS) or are expression quantitative trait locus (eQTL). Full list of SNPs is given in Supplementary Table 6–8.