nature genetics

Genome-wide association study of primary sclerosing cholangitis identifies new risk loci and quantifies the genetic relationship with inflammatory bowel disease

Sun-Gou Ji[1,59], Brian D Juran[2,59], Sören Mucha[3], Trine Folseraas[4–6], Luke Jostins[7,8], Espen Melum[4,5], Natsuhiko Kumasaka[1], Elizabeth J Atkinson[9], Erik M Schlicht[2], Jimmy Z Liu[1], Tejas Shah[1], Javier Gutierrez-Achury[1], Kirsten M Boberg[4,6,10], Annika Bergquist[11], Severine Vermeire[12,13], Bertus Eksteen[14], Peter R Durie[15], Martti Farkkila[16], Tobias Müller[17], Christoph Schramm[18], Martina Sterneck[19], Tobias J Weismüller[20–22], Daniel N Gotthardt[23], David Ellinghaus[3], Felix Braun[24], Andreas Teufel[25], Mattias Laudes[26], Wolfgang Lieb[27], Gunnar Jacobs[27], Ulrich Beuers[28], Rinse K Weersma[29], Cisca Wijmenga[30], Hanns-Ulrich Marschall[31], Piotr Milkiewicz[32], Albert Pares[33], Kimmo Kontula[34], Olivier Chazouillères[35], Pietro Invernizzi[36], Elizabeth Goode[37], Kelly Spiess[37], Carmel Moore[38,39], Jennifer Sambrook[39,40], Willem H Ouwehand[1,38,40,41], David J Roberts[38,42,43], John Danesh[1,38,39], Annarosa Floreani[44], Aliya F Gulamhusein[2], John E Eaton[2], Stefan Schreiber[3,45], Catalina Coltescu[46], Christopher L Bowlus[47], Velimir A Luketic[48], Joseph A Odin[49], Kapil B Chopra[50], Kris V Kowdley[51], Naga Chalasani[52], Michael P Manns[20,21], Brijesh Srivastava[37], George Mells[37,53], Richard N Sandford[37], Graeme Alexander[54], Daniel J Gaffney[1], Roger W Chapman[55], Gideon M Hirschfield[56,57], Mariza de Andrade[9] The UK-PSC Consortium[58], The International IBD Genetics Consortium[58], The International PSC Study Group[58], Simon M Rushbrook[37], Andre Franke[3], Tom H Karlsen[4–6,10], Konstantinos N Lazaridis[2,60] & Carl A Anderson[1,60]

Primary sclerosing cholangitis (PSC) is a rare progressive disorder leading to bile duct destruction; ~75% of patients have comorbid inflammatory bowel disease (IBD). We undertook the largest genome-wide association study of PSC (4,796 cases and 19,955 population controls) and identified four new genome-wide significant loci. The most associated SNP at one locus affects splicing and expression of *UBASH3A*, with the protective allele (C) predicted to cause nonstop-mediated mRNA decay and lower expression of UBASH3A. Further analyses based on common variants suggested that the genome-wide genetic correlation (r_G) between PSC and ulcerative colitis (UC) (r_G = 0.29) was significantly greater than that between PSC and Crohn's disease (CD) (r_G = 0.04) (P = 2.55 × 10^{-15}). UC and CD were genetically more similar to each other (r_G = 0.56) than either was to PSC ($P < 1.0 \times 10^{-15}$). Our study represents a substantial advance in understanding of the genetics of PSC.

PSC affects around 1 in 10,000 individuals of European ancestry, and is characterized by chronic inflammation and stricturing fibrosis of the biliary tree[1]. There remains no effective medical therapy, and the majority of patients require orthotopic liver transplantation owing to the progressive nature of the disease[2]. PSC is highly comorbid with IBD, which is ultimately diagnosed in around 75% of patients. The clinical presentation of IBD in PSC is most often consistent with UC (~80%), but CD (~15%) and indeterminate forms of IBD (~5%) occur in some patients. Time of disease onset and expression of the IBD phenotype in PSC varies, with an overall trend toward IBD preceding PSC, and milder but more extensive intestinal inflammation (pancolitis) compared to classical UC or CD[3,4]. This tendency, along with other clinical and epidemiological differences, has led to the proposal that IBD in the context of PSC (PSC-IBD) should be considered a disease entity separate from both UC and CD. Elevated risk of PSC and UC in first-degree relatives of PSC patients indicates a strong genetic component to PSC susceptibility, and suggests the presence of shared genetic risk factors between PSC and UC[5,6]. However, the genetic relationship between PSC and both UC and CD remains poorly defined because the low prevalence of PSC has precluded familial studies. Large-scale association studies have identified 16 loci, including the *HLA* locus, underlying PSC risk[7–12]. Here we undertook the largest genome-wide association study (GWAS) of PSC to date to identify new PSC risk loci and enable us, to estimate the r_G between PSC and the common forms of IBD.

Following quality control (**Supplementary Figs. 1–3** and **Supplementary Tables 1** and **2**) and imputation using reference haplotypes from the 1000 Genomes (Phase III) and UK10K

A full list of affiliations appears at the end of the paper.

Received 9 February; accepted 18 November; published online 19 December 2016; doi:10.1038/ng.3745

Table 1 Association summary statistics across four newly associated PSC risk loci

SNP	Chromosome and position (bp)	Risk allele	RAF	OR	95% CI	*P* value			Candidate causal gene
						GWAS	Replication	Combined	
rs80060485	3:71153890	C	0.07	1.44	1.32-1.58	8.54×10^{-9}	4.67×10^{-8}	2.62×10^{-15}	*FOXP1*
rs663743	11:64107735	G	0.66	1.20	1.14-1.26	8.42×10^{-8}	4.44×10^{-7}	2.24×10^{-13}	*CCDC88B*
rs725613	16:11169683	T	0.65	1.20	1.14-1.26	5.50×10^{-10}	9.52×10^{-5}	3.59×10^{-13}	*CLEC16A*
rs1893592	21:43855067	A	0.73	1.22	1.15-1.29	1.90×10^{-7}	2.42×10^{-6}	2.19×10^{-12}	*UBASH3A*

Base pair coordinates from build 37. RAF, risk allele frequency in replication controls; OR, odds ratio in the GWAS and replication meta-analysis (combined); 95% CI, 95% confidence interval of OR estimates. Detailed association results, including those for the 15 loci previously associated with PSC, are given in **Supplementary Table 4**.

projects[13,14], we tested 7,891,602 SNPs for association in a sample of 2,871 PSC cases and 12,019 population controls using a linear mixed model to account for population stratification (Online Methods and **Supplementary Tables 1** and **2**). Genome-wide summary statistics are available from the International PSC Study Group website (see URLs). We tested 40 SNPs for association in an independent cohort of 1,925 PSC cases, and 7,936 population controls (Online Methods and **Supplementary Table 3**), including 24 SNPs with $P < 5 \times 10^{-6}$ in the GWAS that are located outside of known PSC loci. We used an inverse-variance weighted fixed effects meta-analysis, implemented in METAL[15], to test the evidence of association across the GWAS and replication cohorts combined, and identified four new genome-wide significant loci with $P < 5.26 \times 10^{-3}$ in the replication study and $P < 5 \times 10^{-8}$ in the combined meta-analysis (**Table 1**, **Supplementary Fig. 4** and **Supplementary Table 4**). One of the newly associated loci, tagged by rs80060485 (3:g.71153890T>C) in *FOXP1*, was associated with immune-mediated disease for the first time, to our knowledge. The three other newly associated PSC loci (implicating *CCDC88B*, *CLEC16A* and *UBASH3A*) are in high linkage disequilibrium (LD), defined as ($r^2 > 0.8$) with variants significantly associated with other immune-mediated diseases (**Supplementary Table 5**). We found consistent evidence of association at 15 of the 16 previously established PSC loci and thus consider 19 regions of the genome to be associated with PSC risk (**Supplementary Fig. 4** and **Supplementary Table 4**).

We evaluated all SNPs in high LD ($r^2 > 0.8$) with the most associated SNP at each PSC locus for potential function using SIFT[16] and PolyPhen 2 (ref. 17), the Genome Wide Annotation of Variants (GWAVA) online tool[18], and a number of expression quantitative trait locus (eQTL) databases (Online Methods and **Supplementary Tables 6–8**). One of the new PSC risk variants (rs1893592, 21:g.43855067A>C) was the most strongly associated eQTL of *UBASH3A*, a gene involved in regulation of T-cell signaling, in two whole blood-based analyses[19,20] and a B-cell-only study[21]. The SNP is located three bases downstream of the 10th exon of *UBASH3A*, in the splice consensus sequence, and has been reported as a splice QTL in a recent RNA sequencing study[19]. The C allele, which is associated with reduced risk of PSC and has a frequency of 27.8% in our controls, disrupts the conserved 5′ splice donor sequence at this position in vertebrate introns, which is typically A (71% of sites) or G (24% of sites)[22]. The predicted consequence of this change is partial retention of the downstream intron, possibly leading to nonstop-mediated decay. Reanalysis of the gEUVADIS RNA-seq data[23] revealed that this SNP was the most strongly associated with increased intron expression ($P = 2 \times 10^{-16}$; **Supplementary Fig. 5**), with the PSC protective allele causing intron 10 to be retained in the *UBASH3A* mRNA. Further work is required to determine whether carrying the C allele at this SNP decreases UBASH3A protein levels and whether this is the causal mechanism behind the reduced risk of PSC, celiac disease and rheumatoid arthritis (**Supplementary Table 5**). In addition, another variant within the *UBASH3A* gene (rs11203203, 21:g.43836186G>A) that is in low LD ($r^2 = 0.12$) with rs1893592 has been associated with vitiligo[24] and type-1 diabetes[25], further supporting the role of UBASH3A in immune-mediated disorders. We could not identify any current drugs targeting UBASH3A (**Supplementary Note**).

To enable us to address the genetic relationship between PSC and IBD, we obtained association summary statistics from the International IBD Genetics Consortium for 20,550 CD cases, 17,647 UC cases and 48,485 controls of European ancestry[26]. Across each of the 18 non-HLA PSC risk loci, we used a Bayesian test of colocalization[27] to identify loci with strong evidence (posterior probability > 0.8) of either shared or independent causal variants between pairs of traits (Online Methods and **Supplementary Table 9**). Four of the 18 PSC risk loci have not been associated at genome-wide significance with IBD (*BCL2L11*, *FOXP1*, *SIK2* and *UBASH3A*) although the lead SNPs at two of these loci (rs72837826–*BCL2L11* and rs1893592–*UBASH3A*) did demonstrate strong evidence for colocalization (posterior probability > 0.8) and suggestive evidence of association ($P < 10^{-4}$) in the UC cohort (**Supplementary Tables 9** and **10**). Of the 14 PSC loci that had been previously associated with IBD (UC, CD or both), four demonstrated strong evidence that the causal variant is independent from that in UC and CD (*IL2RA*, *CCDC88B*, *CLEC16A* and *PRKD2*), a finding supported by the low LD ($r^2 < 0.2$) between the lead SNPs in PSC and UC or CD at these loci (**Supplementary Tables 9** and **10**). Thus, even for highly comorbid diseases, significant association to the same region of the genome will not always be driven by a shared causal variant. This supports similar observations for other related phenotypes such as psoriasis versus psoriatic arthritis[28,29]. Six of the 14 loci associated with PSC and IBD displayed strong evidence of a shared causal variant with UC, CD or both (*MST1*, *IL21*, *HDAC7*, *SH2B3*, *CD226* and *PSMG1*) (**Fig. 1** and **Supplementary Tables 9** and **10**). We further tested these six SNPs for evidence of heterogeneity of effect using Cochran's *Q* test (Online Methods). Four showed significantly increased effect size in PSC relative to both UC and CD (*MST1*, *IL21*, *SH2B3* and *CD226*) ($P < 2.78 \times 10^{-3}$), with an additional locus (*PSMG1*) showing significantly increased effect size relative to CD only (**Fig. 1**). Simulation studies showed that the observed heterogeneity of effect is unlikely to be driven by the large difference in sample size between the PSC and UC cohorts ($P_{empirical} < 3.00 \times 10^{-4}$ at all four SNPs) (**Supplementary Note**). We did not detect evidence of heterogeneity of effect between PSC patients expressing different IBD phenotypes (PSC-UC, PSC-CD or PSC-NoIBD) (**Supplementary Fig. 6**). However, our power to detect significant heterogeneity of effect between these PSC subphenotypes was limited by sample size (**Supplementary Table 11**).

Although the much larger size of the UC and CD cohorts gave us power to investigate the effects of PSC risk SNPs in IBD, the PSC cohort was underpowered to do the reverse. Thus, to clarify the pairwise genetic correlation between PSC, UC and CD, we obtained

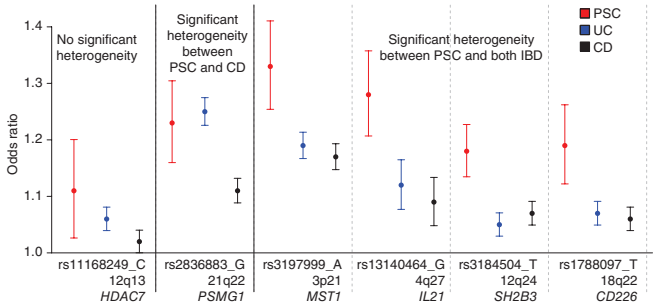

Figure 1 Odds ratios (and their 95% confidence intervals) for PSC, UC and CD across the six PSC-associated SNPs demonstrating strong evidence for a shared causal variant (maximum posterior probability > 0.8). PSC odds ratios were taken from the GWAS and replication meta-analysis. UC and CD odds ratios were obtained from the latest association studies conducted by the International IBD Genetics Consortium[26]. Heterogeneity of odds tests were carried out using Cochran's *Q* test. A failure to detect significant heterogeneity of odds does not necessarily indicate that effect sizes are equivalent because power to detect heterogeneity varies across SNPs.

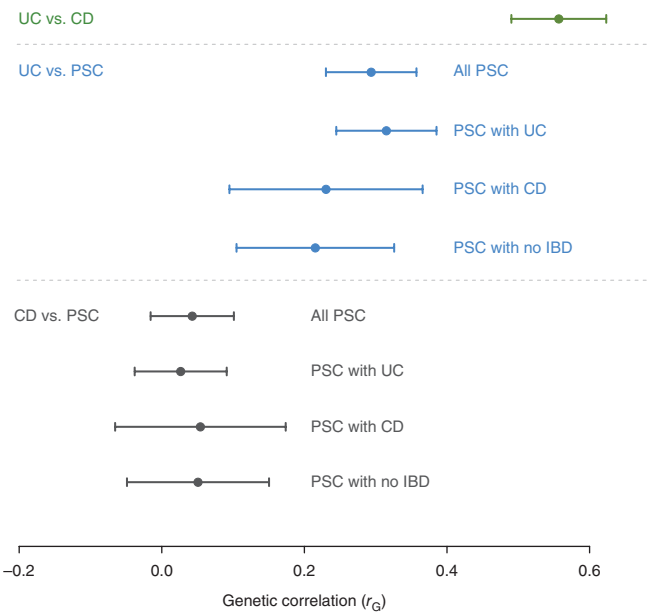

Figure 2 Genome-wide genetic correlation between PSC (and its subphenotypes), CD and UC. Genetic correlations (and their 95% confidence intervals) were calculated using a bivariate extension of the linear mixed model[30] implemented in GCTA (Online Methods). PSC had a lower genetic correlation with both CD and UC than the two inflammatory bowel diseases had to each other. PSC was genetically more correlated to UC than it was to CD, and this was consistent across the PSC subphenotypes.

genome-wide individual level genotype data from the International IBD Genetics Consortium for 6,247 CD cases, 6,686 UC cases and 34,393 population controls of European descent[26], and used genome-wide complex trait analysis (GCTA) to estimate r_G using a bivariate linear mixed model[30,31] (Online Methods and **Supplementary Note**). This analysis quantified the SNP heritability (h^2_{SNP}) of PSC as 0.148 (95% CI: 0.135–0.161), and showed that in the context of common genetic variation, PSC is significantly more related to UC ($r_G = 0.29$) than CD ($r_G = 0.04$) ($P = 2.55 \times 10^{-15}$) (**Fig. 2**), consistent with the clinical phenotype most often observed in PSC-IBD patients. Moreover, the genetic correlation between UC and CD ($r_G = 0.56$) was significantly greater than that between PSC and either UC or CD ($P < 1.0 \times 10^{-15}$). Owing to a lack of data regarding the PSC status of individuals in the UC and CD cohorts, we could not remove the ~5% of patients we would expect to have comorbid PSC. This suggests that, although our estimates of the r_G between PSC and both UC and CD may seem surprisingly low, these are likely slight overestimates of the true genetic correlation between the diseases. We validated the GCTA co-heritability estimates using a summary-statistics-based genetic correlation analysis (LD score regression[32]), and found support for the reported genetic relationships (i.e., $r_{G\ CD\ vs.\ UC} = 0.68 > r_{G\ PSC\ vs.\ UC} = 0.39 > r_{G\ PSC\ vs.\ CD} = 0.09$) (**Supplementary Fig. 7**). The low r_G between PSC and the IBDs is also supported by known differences in HLA risk alleles[11,33] and our discovery that PSC has both independent causal variants and shared causal variants of heterogeneous effect size compared to both UC and CD. The analyses presented in this study, based on common genetic variants (MAF > 1%), suggest functional studies in both the biliary tree and intestinal tract are required if we are to understand the biological consequences of PSC-associated genetic variants, whether or not they are shared with IBD.

Although it is clear that a substantial component of the genetic architecture of PSC is not shared with either CD or UC, our data also showed that shared genetic risk factors do certainly exist and likely have some role in disease comorbidity. However, under a purely additive genetic liability threshold model, the genetic covariance between the two diseases would need to be greater than 0.76 to fully explain the fact that 60% of PSC cases have comorbid UC (**Supplementary Fig. 8**). In contrast, the observed genetic correlation ($r_G = 0.29$) would generate a PSC-UC comorbidity rate of only 1.6% under this model. This demonstrates that the observed extent of comorbidity between PSC and UC is not fully explained by shared additive genetic effects of common variants and that other factors must play a role, such as shared environmental effects or shared rare variants not captured by our GWAS and imputation data.

In summary, we performed the largest genome-wide association study of PSC to date and identified four new PSC risk loci. We now consider 23 regions of the genome to be associated with disease risk, including four loci only recently associated with PSC in a cross-disease meta-analysis[34]. One of our new associations suggests that decreased *UBASH3A* is associated with a lower risk of PSC through a common nonstop-mediated mRNA decay variant. We also showed that, even for highly comorbid phenotypes such as PSC and IBD, significant association to the same region of the genome will not always be driven by a common causal variant. Furthermore, by conducting genome-wide comparisons with CD and UC, we showed that the comorbid gastrointestinal inflammation seen in the majority of PSC patients cannot be fully explained by shared genetic risk. Thus, the biliary and intestinal inflammation seen specifically in PSC should be studied to advance our understanding of the disease and improve clinical outcome for patients with this devastating disorder.

URLs. http://www.ipscsg.org; http://ous-research.no/nopsc/; http://www.ibdgenetics.org; http://www.internationalgenome.org/; http://www.uk10k.org; http://www.genome.gov/gwastudies/; https://www.immunobase.org/; http://eqtl.uchicago.edu/cgi-bin/gbrowse/eqtl/; http://www.gtexportal.org/; http://hrsonline.isr.umich.edu/; https://www.sanger.ac.uk/sanger/StatGen_Gwava; http://sift.jcvi.org/www/SIFT_chr_coords_submit.html; http://genetics.bwh.harvard.edu/pph2/; http://www.broadinstitute.org/mpg/grail/; https://data.broadinstitute.org/alkesgroup/LDSCORE.

METHODS

Methods, including statements of data availability and any associated accession codes and references, are available in the online version of the paper.

Note: Any Supplementary Information and Source Data files are available in the online version of the paper.

ACKNOWLEDGMENTS

We thank the patients and healthy controls for their participation, and are grateful to the physicians, scientists and nursing staff who recruited individuals whose data is used in our study. We acknowledge the use of DNA or genotype data from a number of sources, including: the Health and Retirement Study (HSR) conducted by the University of Michigan, funded by the National Institute on Aging (grant numbers U01AG009740, RC2AG036495 and RC4AG039029) and accessed via dbGaP; Popgen 2.0, supported by a grant from the German Ministry for Education and Research (01EY1103); The Mayo Clinic Biobank, supported by the Mayo Clinic Center for Individualized Medicine; the INTERVAL study, undertaken by the University of Cambridge with funding from the National Health Service Blood and Transplant (NHSBT) (the views expressed in this publication are those of the authors and not necessarily those of the NHSBT); the FOCUS biobank. We thank the investigators of the 1000 Genomes and UK10K projects for generating and sharing the population haplotypes and Jie Huang for advice regarding imputation. We thank all members of the International IBD Genetics Consortium for sharing genetic data vital to the success of our study. This study was supported by NoPSC, the National Institute of Diabetes and Digestive and Kidney Diseases (NIDDK RO1DK084960, KNL), the Wellcome Trust (098759/Z/12/Z: L.J.; 098051: S.-G.J., J.Z.L., T.S., J.G.-A., N.K., D.J.G. and C.A.A.), the Kwanjeong Educational Foundation (S.-G.J.), the German Federal Ministry of Education and Research (B.M.B.F.) within the framework of the e:Med research and funding concept (SysInflame grant 01ZX1306A) and the Chris M. Carlos and Catharine Nicole Jockisch Carlos Endowment in PSC. This project received infrastructure support from the DFG Excellence Cluster 306 "Inflammation at Interfaces" and the PopGen Biobank (Kiel, Germany), an endowment professorship (A.F.) by the Foundation for Experimental Medicine (Zurich, Switzerland). The recruitment of patients in Hamburg was supported by the YAEL-Foundation and the DFG (SFB841). B.A. Lie and the Norwegian Bone Marrow Donor Registry at Oslo University Hospital, Rikshospitalet in Oslo are acknowledged for sharing the healthy Norwegian controls. Participants in the INTERVAL randomized controlled trial were recruited with the active collaboration of NHS Blood and Transplant England (http://www.nhsbt.nhs.uk), which has supported field work and other elements of the trial. DNA extraction and genotyping was funded by the National Institute of Health Research (NIHR), the NIHR BioResource (http://bioresource.nihr.ac.uk/) and the NIHR Cambridge Biomedical Research Centre (http://www.cambridge-brc.org.uk). The academic coordinating centre for INTERVAL was supported by core funding from: NIHR Blood and Transplant Research Unit in Donor Health and Genomics, UK Medical Research Council (G0800270), British Heart Foundation (SP/09/002), and NIHR Research Cambridge Biomedical Research Centre. We thank K. Cloppenborg-Schmidt, I. Urbach, I. Pauselis, T. Wesse, T. Henke, R. Vogler, V. Pelkonen, K. Holm, H. Dahlen Sollid, B. Woldseth, J. Andreas and L. Wenche Torbjørnsen for expert help. R.K.W. is supported by a clinical fellowship grant (90.700.281) from the Netherlands Organization for Scientific Research. B.E. receives support from Medical Research Council, United Kingdom. T.M. and D.G. are supported by Deutsche Forschungsgemeinschaft, Grant. A.P. is supported by Centro de Investigación Biomédica en Red de Enfermedades Hepáticas y Digestivas (CIBERehd), grant PI071318 Instituto de Salud Carlos III, Ministerio de Ciencia e Innovación, and grant PI12/01448, from Ministerio de Economía y Competitvidad, Spain. P.R.D. is supported by Canadian Institutes of Health research (CIHR) and Genome Canada. C.W. is supported by grants from the Celiac Disease Consortium (BSIK03009) and Netherlands Organization for Scientific Research (NWO, VICI grant918.66.620). We acknowledge members of the International PSC Study Group, the NIDDK Inflammatory Bowel Disease Genetics Consortium (IBDGC), and the UK-PSC Consortium for their participation. We thank J. Rud for secretarial support.

AUTHOR CONTRIBUTIONS

S.-G.J., B.D.J., N.K., T.S., J.G.-A. and C.A.A. performed statistical data analysis. S.-G.J., B.D.J., S.M., T.F., E.M., E.J.A. and C.A.A. performed initial quality control and sample identification. L.J., J.Z.L., D.J.G., M.d.A. and C.A.A. provided statistical and analytical advice. T.H.K., K.N.L. and C.A.A. coordinated the project and supervised the analyses. S.-G.J., B.D.J., T.H.K., K.N.L. and C.A.A. drafted of the manuscript. E.M.S., K.M.B., A.B., S.V., B.E., P.R.D., M.F., T.M., C.S., M.S., T.J.W., D.N.G., D.E., F.B., A.T., M.L., W.L., G.J., U.B., R.K.W., C.W., H.-U.M., P.M., A.P., K.K., O.C., P.I., E.G., K.S., C.M., J.S., W.H.O., D.J.R., J.D., A.F., A.F.G., J.E.E., S.S., C.C., C.L.B., V.A.L., J.A.O., K.B.C., K.V.K., N.C., M.P.M., B.S., G.M., R.N.S., G.A., R.W.C., G.M.H., S.M.R., A.F., K.N.L., C.A.A., The UK-PSC Consortium, The International IBD Genetics Consortium, and The International PSC Study Group collected the samples, performed clinical ascertainment or coordinated sample logistics. All authors read and approved the final version of the manuscript.

COMPETING FINANCIAL INTERESTS

The authors declare no competing financial interests.

1. Boonstra, K. *et al.* Primary sclerosing cholangitis is associated with a distinct phenotype of inflammatory bowel disease. *Inflamm. Bowel Dis.* **18**, 2270–2276 (2012).
2. Tischendorf, J.J.W., Hecker, H., Krüger, M., Manns, M.P. & Meier, P.N. Characterization, outcome, and prognosis in 273 patients with primary sclerosing cholangitis: a single center study. *Am. J. Gastroenterol.* **102**, 107–114 (2007).
3. Karlsen, T.H. & Kaser, A. Deciphering the genetic predisposition to primary sclerosing cholangitis. *Semin. Liver Dis.* **31**, 188–207 (2011).
4. Karlsen, T.H., Schrumpf, E. & Boberg, K.M. Update on primary sclerosing cholangitis. *Dig. Liver Dis.* **42**, 390–400 (2010).
5. Bergquist, A. *et al.* Increased risk of primary sclerosing cholangitis and ulcerative colitis in first-degree relatives of patients with primary sclerosing cholangitis. *Clin. Gastroenterol. Hepatol.* **6**, 939–943 (2008).
6. de Vries, A.B., Janse, M., Blokzijl, H. & Weersma, R.K. Distinctive inflammatory bowel disease phenotype in primary sclerosing cholangitis. *World J. Gastroenterol.* **21**, 1956–1971 (2015).
7. Melum, E. *et al.* Genome-wide association analysis in primary sclerosing cholangitis identifies two non-HLA susceptibility loci. *Nat. Genet.* **43**, 17–19 (2011).
8. Ellinghaus, D. *et al.* Genome-wide association analysis in primary sclerosing cholangitis and ulcerative colitis identifies risk loci at *GPR35* and *TCF4*. *Hepatology* **58**, 1074–1083 (2013).
9. Folseraas, T. *et al.* Extended analysis of a genome-wide association study in primary sclerosing cholangitis detects multiple novel risk loci. *J. Hepatol.* **57**, 366–375 (2012).
10. Karlsen, T.H. *et al.* Genome-wide association analysis in primary sclerosing cholangitis. *Gastroenterology* **138**, 1102–1111 (2010).
11. Liu, J.Z. *et al.* Dense genotyping of immune-related disease regions identifies nine new risk loci for primary sclerosing cholangitis. *Nat. Genet.* **45**, 670–675 (2013).
12. Srivastava, B. *et al.* Fine mapping and replication of genetic risk loci in primary sclerosing cholangitis. *Scand. J. Gastroenterol.* **47**, 820–826 (2012).
13. 1000 Genomes Project Consortium. A global reference for human genetic variation. *Nature* **526**, 68–74 (2015).
14. UK 10K Consortium. The UK10K project identifies rare variants in health and disease. *Nature* **526**, 82–90 (2015).
15. Willer, C.J., Li, Y. & Abecasis, G.R. METAL: fast and efficient meta-analysis of genomewide association scans. *Bioinformatics* **26**, 2190–2191 (2010).
16. Kumar, P., Henikoff, S. & Ng, P.C. Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm. *Nat. Protoc.* **4**, 1073–1081 (2009).
17. Adzhubei, I.A. *et al.* A method and server for predicting damaging missense mutations. *Nat. Methods* **7**, 248–249 (2010).
18. Ritchie, G.R.S., Dunham, I., Zeggini, E. & Flicek, P. Functional annotation of noncoding sequence variants. *Nat. Methods* **11**, 294–296 (2014).
19. Battle, A. *et al.* Characterizing the genetic basis of transcriptome diversity through RNA-sequencing of 922 individuals. *Genome Res.* **24**, 14–24 (2014).
20. Westra, H.J. *et al.* Systematic identification of *trans* eQTLs as putative drivers of known disease associations. *Nat. Genet.* **45**, 1238–1243 (2013).
21. Fairfax, B.P. *et al.* Genetics of gene expression in primary immune cells identifies cell type–specific master regulators and roles of HLA alleles. *Nat. Genet.* **44**, 502–510 (2012).
22. Zhang, M.Q. Statistical features of human exons and their flanking regions. *Hum. Mol. Genet.* **7**, 919–932 (1998).
23. Lappalainen, T. *et al.* Transcriptome and genome sequencing uncovers functional variation in humans. *Nature* **501**, 506–511 (2013).
24. Jin, Y. *et al.* Genome-wide association analyses identify 13 new susceptibility loci for generalized vitiligo. *Nat. Genet.* **44**, 676–680 (2012).
25. Barrett, J.C. *et al.* Genome-wide association study and meta-analysis find that over 40 loci affect risk of type 1 diabetes. *Nat. Genet.* **41**, 703–707 (2009).
26. Liu, J.Z. *et al.* Association analyses identify 38 susceptibility loci for inflammatory bowel disease and highlight shared genetic risk across populations. *Nat. Genet.* **47**, 979–986 (2015).

27. Giambartolomei, C. *et al.* Bayesian test for colocalisation between pairs of genetic association studies using summary statistics. *PLoS Genet.* **10**, e1004383 (2014).
28. Stuart, P.E. *et al.* Genome-wide association analysis of psoriatic arthritis and cutaneous psoriasis reveals differences in their genetic architecture. *Am. J. Hum. Genet.* **97**, 816–836 (2015).
29. Bowes, J. *et al.* Dense genotyping of immune-related susceptibility loci reveals new insights into the genetics of psoriatic arthritis. *Nat. Commun.* **6**, 6046 (2015).
30. Lee, S.H., Yang, J., Goddard, M.E., Visscher, P.M. & Wray, N.R. Estimation of pleiotropy between complex diseases using single-nucleotide polymorphism–derived genomic relationships and restricted maximum likelihood. *Bioinformatics* **28**, 2540–2542 (2012).
31. Yang, J., Lee, S.H., Goddard, M.E. & Visscher, P.M. GCTA: a tool for genome-wide complex trait analysis. *Am. J. Hum. Genet.* **88**, 76–82 (2011).
32. Bulik-Sullivan, B.K. *et al.* LD Score regression distinguishes confounding from polygenicity in genome-wide association studies. *Nat. Genet.* **47**, 291–295 (2015).
33. Goyette, P. *et al.* High-density mapping of the MHC identifies a shared role for HLA-DRB1*01:03 in inflammatory bowel diseases and heterozygous advantage in ulcerative colitis. *Nat. Genet.* **47**, 172–179 (2015).
34. Ellinghaus, D. *et al.* Analysis of five chronic inflammatory diseases identifies 27 new associations and highlights disease-specific patterns at shared loci. *Nat. Genet.* **48**, 510–518 (2016).

[1]Wellcome Trust Sanger Institute, Wellcome Trust Genome Campus, Hinxton, UK. [2]Center for Basic Research in Digestive Diseases, Division of Gastroenterology and Hepatology, Mayo Clinic College of Medicine, Rochester, Minnesota, USA. [3]Institute of Clinical Molecular Biology, Christian Albrechts University of Kiel, Kiel, Germany. [4]Norwegian PSC Research Center, Division of Cancer Medicine, Surgery and Transplantation, Oslo University Hospital, Rikshospitalet, Oslo, Norway. [5]Research Institute of Internal Medicine, Oslo University Hospital, Rikshospitalet, Oslo, Norway. [6]Institute of Clinical Medicine, University of Oslo, Oslo, Norway. [7]Wellcome Trust Centre for Human Genetics, University of Oxford, Oxford, UK. [8]Christ Church, University of Oxford, St Aldates, Oxford, UK. [9]Division of Biomedical Statistics and Informatics, Mayo Clinic, Rochester, Minnesota, USA. [10]Section of Gastroenterology, Department of Transplantation Medicine, Division of Cancer, Surgery and Transplantation, Oslo University Hospital, Rikshospitalet, Oslo, Norway. [11]Department of Gastroenterology and Hepatology, Karolinska University Hospital Huddinge, Karolinska Institutet, Stockholm, Sweden. [12]Department of Clinical and Experimental Medicine, Katholieke Universiteit Leuven, Leuven, Belgium. [13]Department of Gastroenterology, University Hospital Leuven, Leuven, Belgium. [14]Snyder Institute for Chronic Diseases, Department of Medicine, University of Calgary, Calgary, Alberta, Canada. [15]Physiology and Experimental Medicine, Research Institute, Hospital for Sick Children, Toronto, Ontario, Canada. [16]Helsinki University and Helsinki University Hospital, Clinic of Gastroenterology, Helsinki, Finland. [17]Department of Internal Medicine, Hepatology and Gastroenterology, Charité Universitätsmedizin Berlin, Berlin, Germany. [18]1st Department of Medicine, University Medical Center Hamburg-Eppendorf, Hamburg, Germany. [19]Department of Hepatobiliary Surgery and Transplantation, University Medical Center Hamburg-Eppendorf, Hamburg, Germany. [20]Department of Gastroenterology, Hepatology and Endocrinology, Hannover Medical School, Hannover, Germany. [21]Integrated Research and Treatment Center–Transplantation (IFB-tx), Hannover Medical School, Hannover, Germany. [22]Department of Internal Medicine 1, University Hospital of Bonn, Bonn, Germany. [23]Department of Medicine, University Hospital of Heidelberg, Heidelberg, Germany. [24]Department of General, Visceral, Thoracic, Transplantation and Pediatric Surgery, University Medical Centre Schleswig-Holstein, Campus Kiel, Kiel, Germany. [25]Department of Medicine I, University Medical Center, Regensburg, Germany. [26]Clinic of Internal Medicine I, University Hospital Schleswig-Holstein, Kiel, Germany. [27]Institute of Epidemiology and Biobank PopGen, University Hospital Schleswig-Holstein, Kiel, Germany. [28]Department of Gastroenterology and Hepatology, Academic Medical Center, University of Amsterdam, Amsterdam, the Netherlands. [29]Department of Gastroenterology and Hepatology, University of Groningen, University Medical Centre Groningen, Groningen, the Netherlands. [30]Department of Genetics, University of Groningen, University Medical Centre Groningen, Groningen, the Netherlands. [31]Department of Molecular and Clinical Medicine, Institute of Medicine, Sahlgrenska Academy, University of Gothenburg, Gothenburg, Sweden. [32]Liver and Internal Medicine Unit, Department of General, Transplant and Liver Surgery, Medical University of Warsaw, Warsaw, Poland. [33]Liver Unit, Hospital Clínic, IDIBAPS, CIBERehd, University of Barcelona, Barcelona, Spain. [34]Department of Medicine, University of Helsinki, Helsinki, Finland. [35]AP-HP Hôpital Saint Antoine, Department of Hepatology, UPMC University Paris 6, Paris, France. [36]Center for Autoimmune Liver Diseases, Humanitas Clinical and Research Center, Rozzano, Milan, Italy. [37]Academic Department of Medical Genetics, University of Cambridge, Cambridge, UK. [38]NIHR Blood and Transplant Research Unit in Donor Health and Genomics, Department of Public Health and Primary Care, University of Cambridge, Cambridge, UK. [39]INTERVAL Coordinating Centre, Department of Public Health and Primary Care, University of Cambridge, Cambridge, UK. [40]Department of Hematology, University of Cambridge, Cambridge, UK. [41]NHS Blood and Transplant, Cambridge, UK. [42]NHS Blood and Transplant–Oxford Centre, John Radcliffe Hospital, Oxford, UK. [43]Radcliffe Department of Medicine, University of Oxford, John Radcliffe Hospital, Oxford, UK. [44]Department of Surgical, Oncological and Gastroenterological Sciences, University of Padova, Padova, Italy. [45]Department for General Internal Medicine, University Hospital Schleswig-Holstein Campus Kiel, Kiel, Germany. [46]Liver Centre, Toronto Western Hospital, Toronto, Ontario, Canada. [47]Division of Gastroenterology and Hepatology, University of California at Davis, Davis, California, USA. [48]Gastroenterology and Hepatology Section, Virginia Commonwealth University, Richmond, Virginia, USA. [49]Department of Medicine, Mount Sinai School of Medicine, New York, New York, USA. [50]Division of Gastroenterology, Hepatology and Nutrition, University of Pittsburgh, Pittsburgh, Pennsylvania, USA. [51]Liver Care Network and Organ Care Research, Swedish Medical Center, Seattle, Washington, USA. [52]Indiana University School of Medicine, Indianapolis, Indiana, USA. [53]Division of Gastroenterology and Hepatology, Addenbrooke's Hospital, Cambridge, UK. [54]Department of Medicine, Division of Hepatology, University of Cambridge, Cambridge, UK. [55]Department of Translational Gastroenterology, Oxford University Hospitals NHS Trust, Oxford, UK. [56]Centre for Liver Research, NIHR Biomedical Research Unit, University of Birmingham, Birmingham, UK. [57]University of Toronto and Liver Center, Toronto Western Hospital, Toronto, Ontario, Canada. [58]A list of members and affiliations appears in the **Supplementary Note**. [59]These authors contributed equally to this work. [60]These authors jointly directed this work. Correspondence should be addressed to C.A.A. (carl.anderson@sanger.ac.uk) or K.N.L. (lazaridis.konstantinos@mayo.edu).

ONLINE METHODS

Ethical approval. The ethics committees or institutional review boards of all participating centers approved the studies and the recruitment of participants. Written informed consent was obtained from all participants.

GWAS cohort. *Cohorts and genotyping.* 731 PSC cases and 3,202 population controls from Scandinavia and Germany were ascertained and genotyped using the Affymetrix Genome-Wide Human SNP Array 6.0 (Affymetrix) at three different centers[7]. A cohort of 1,227 UK PSC cases was recruited from across more than 150 UK National Health Service Trusts or Health Boards, including all transplant centers in the UK, by the UK-PSC consortium. A cohort of 904 US PSC patients were enrolled in the PSC Resource of Genetic Risk, Environment and Synergy Studies (PROGRESS), a multicenter collaboration between eight academic research institutions across the US and Canada. PROGRESS ascertained additional DNA samples from established PSC cohorts from Canada ($n = 259$) and Poland ($n = 43$). The UK and US GWAS cohorts were genotyped using the Illumina HumanOmni2.5-8 BeadChip (Illumina) and called using the GenCall algorithm implemented in GenomeStudio. UK samples were genotyped at the Wellcome Trust Sanger Institute (Hinxton, UK) and the US samples at the Mayo Clinic Medical Genome Facility (Rochester, Minnesota, USA). A diagnosis of PSC was based on standard clinical, biochemical, cholangiographic and histological criteria[35], with exclusion of secondary causes of sclerosing cholangitis. Commonly accepted clinical, radiological, endoscopic and histological criteria were also used for diagnosis and classification of IBD[36]. Genetic data from 12,595 individuals genotyped on the Illumina HumanOmni2.5-4v1 array (Omni2.5-4) as part of The University of Michigan Health Retirement Study were downloaded from the Database of Genotypes and Phenotypes (dbGaP[37]). Genotyping was performed at the Center for Inherited Disease Research (CIDR) and genotypes called using GenomeStudio version 2011.e, (see the Health Retirement Study (HRS) website for more details).

Quality control. All SNPs were aligned to NCBI build 37 (hg19). Genotype data were quality=controlled independently across six batches defined by genotyping center ($Affy^{SF}$, $n = 2,205$; $Affy^{HZ}$, $n = 1,256$; $Affy^{AB}$, $n = 472$; $Illumina^{WTSI}$, $n = 1,227$; $Illumina^{MAYO}$, $n = 1,206$; $Illumina^{CIDR}$, $n = 12,595$). Initially, SNPs out of Hardy-Weinberg equilibrium (HWE, $P < 1 \times 10^{-6}$) in controls (excluding those in the HLA region) or with a call rate less than 80% were removed. SNPs failing in at least one batch were removed from all cohorts genotyped using the same chip. For sample quality control (QC), individuals whose sex determined using the X chromosome homozygosity rate (F) and Y chromosome call rate differed from that in our patient database (or could not be genetically determined, F or Y-chromosome call rate between 0.3 and 0.7) were removed. Next, Abberant[38] was used to identify samples with outlying heterozygosity or genotype call rate. Samples with a call rate less than 90% for an individual chromosome were also removed. A set of 82,085 independent SNPs (pairwise $r^2 < 0.2$) genotyped on all arrays was identified for the purpose of estimating sample relatedness and ancestry, excluding SNPs that (i) were in regions of high linkage disequilibrium, (ii) had a MAF < 10% or (iii) were A/T or C/G SNPs. Pairwise identity by descent was estimated for all individuals in the study using PLINK, and the sample with the lowest genotype call rate was removed for all pairs with IBD > 0.9. Both samples were excluded if case/control status was discordant between duplicates. To maximize power to detect association, related samples (0.1875 < IBD < 0.9) were retained and a mixed model used for association testing. Sample ancestry was inferred via principal components analysis implemented in EIGENSTRAT[39]. Population principal components were calculated using genotype data from the CEU, YRI and CHB/JPT samples from the 1000 Genomes Project. Factor loadings from these principal components were then used to project these principal components for our cases and controls. Samples of non-European ancestry were identified using Aberrant[38]. The number of samples failing each QC step is shown in **Supplementary Table 1**. In total, 2,871 cases and 12,019 controls passed sample QC. Next, a more thorough marker QC was conducted within batches by excluding, genotyping platform-wide, SNPs with (i) different probe sequences on the Omni2.5-4 and Omni2.5-8 array, (ii) a call rate < 98%, (iii) MAF<1%, (iv) significant evidence of deviation from HWE ($P < 1 \times 10^{-5}$) in controls and (v) a significant difference in call rate between cases and controls ($P < 1 \times 10^{-5}$), in at least one of the genotyping batches. Outside of the HLA region, markers only present on one of the two Illumina arrays were also removed. After SNP QC, 1,207,121 Omni2.5-4 SNPs, 1,215,097 Omni2.5-8 SNPs and 528,496 Affymetrix 6 SNPs were available.

Genotype imputation. Only 322,807 SNPs feature on both the Affy6 and Omni2.5 arrays so the samples genotyped on these arrays were phased and imputed separately. For computational efficiency, the genome was split into 3-Mb batches, and those spanning the centromere were split and joined to the last complete batch either side of the centromere. Batches of less than 200 SNPs were merged with an adjacent batch. Pre-phasing was performed using the SHAPEIT2 algorithm[40] and imputation using the IMPUTE2 algorithm[41]. We used a combined reference panel of the 1000 Genomes Phase 1 integrated version 3 and the UK10K cohort, consisting of 4,873 individuals and 42,359,694 SNPs (k_hap = 2,000, Ne = 20,000). Post-imputation, SNPs with a posterior probability less than 0.9 or info score less than 0.5 were removed. The QC steps outlined above for directly genotyped SNPs were applied to the imputed genotype data. SNPs with $r^2 < 0.8$ between directly genotyped and imputed genotypes were removed and phasing and imputation repeated. Following QC (as outlined above), a total of 7,891,602 SNPs available for association testing across 2,871 PSC cases and 12,019 population controls (**Supplementary Table 2**).

Association analysis. A linear mixed model implemented in the MMM software[42] was used to test association between genetic variants and case/control status. To reduce compute time the relationship matrix was constructed using the 82,085 quasi-independent SNPs previously used in the principal-component analysis (PCA). To prevent the association analyses being biased by informed missingness across our genotyping batches, linear mixed model association tests were conducted across three different batches of directly-genotyped and imputed SNPs, defined on their availability for only the Omni2.5 genotyped samples ($n = 2,015,514$), only the Affy6 genotyped samples ($n = 114,935$), or across all genotyped samples ($n = 5,761,153$).

Stepwise conditional regression analysis (excluding the extended MHC region) was undertaken in MMM to identify independent association signals ($P < 5.0 \times 10^{-6}$) within PSC associated loci. The previously reported lead SNP within each of the 15 known PSC loci was selected for replication, though we also took forward the most associated SNP in our study if it was a poor tag ($r^2 < 0.8$) of previously reported SNP. In addition, 24 SNPs outside of established PSC risk loci with $P < 5 \times 10^{-6}$ were also included in the replication experiment. All cluster plots were manually inspected before SNP selection.

Validation and replication cohorts. *Cohorts and genotyping.* An independent replication cohort of 2,011 PSC cases from Europe and North America was ascertained following the diagnostic criteria outlined above. A total of 8,784 population controls of European descent were ascertained, including 515 from the Mayo Clinic Biobank[43] and 1000 from the INTERVAL study[44]. British and Canadian samples were genotyped at the Wellcome Trust Sanger Institute in Cambridge, UK ($n = 2,366$) and all other samples at the Institute of Clinical Molecular Biology in Kiel, Germany ($n = 11,152$) using the same Agena Biosciences iPLEX design. To reduce the risk of false-positive associations being driven by imputation errors we undertook a substantial validation experiment, genotyping the 40 SNPs in our replication experiment across 2,723 cases in the GWAS study.

Quality control. Two SNPs yielded poor genotype clusters and were removed from further study. Four SNPs with a call rate less than 95% or Hardy Weinberg equilibrium $P < 1.25 \times 10^{-3}$ (Bonferroni correction for 40 SNPs) in controls were excluded (**Supplementary Table 12**). Samples with a call rate less than 92%, or where the genetically determined sex differed from that in our patient database, were removed. The sample with the lowest call rate in duplicate pairs was removed from duplicate pairs (IBS > 0.9) (**Supplementary Table 3**). Post-QC, one SNP had an r^2 less than 0.90 between the discovery and validation genotyping and, following manual inspection of cluster plots, was removed from the replication study.

Replication and combined association analyses. For the replication analysis, logistic regression tests of association were performed separately for samples from six geographic regions (**Supplementary Table 3**) using SNPTEST v2. Inverse-variance-weighted fixed effects meta-analyses implemented in METAL[15] were then used to (i) test for association across all replication samples and (ii) test the evidence of association across the GWAS and replication cohorts combined. To classify a region as newly associated with PSC,

doi:10.1038/ng.3745

we required both significant evidence of association in the replication cohort ($P < 5.26 \times 10^{-3}$, Bonferroni correction for 19 one-tailed tests) and genome-wide significance ($P < 5 \times 10^{-8}$) in the combined meta-analysis.

Candidate gene prioritization. *Functional annotation.* All SNPs in high LD ($r^2 > 0.8$) with lead SNPs at PSC associated loci were annotated for potential function using the Genome Wide Annotation of Variants (GWAVA) online tool[18]. In addition, all coding SNPs from this set were also annotated using SIFT[16] and PolyPhen2 (ref. 17).

Pathway analysis. To quantify the functional relationship between genes within PSC risk loci, we conducted a GRAIL pathway analysis. GRAIL evaluates the degree of functional connectivity between genes based on the extent they co-feature in published abstracts (we used all PubMed abstracts before 2006 to avoid biasing our analysis due to results from large-scale GWASs). All PSC associated loci were included in the analysis and only genes with GRAIL $P < 0.05$ and edges with a score of > 0.5 were included in the connectivity map.

Expression quantitative trait loci. eQTL analysis focused on published *cis*-eQTLs owing to the lower reproducibility caused by smaller effect sizes and context-specificity of *trans*-eQTL[45]. Eight eQTL data sets were included in the analysis: eQTL data from 12 studies collated in the Chicago eQTL browser, eQTL results from 1,421 samples of 13 different tissue types by the genotype-tissue expression (GTEx) project[46], 462 lymphoblastoid cell lines[23], 922 whole blood samples[19], 8,086 whole blood samples[20], purified B cells and monocytes from 283 individuals[21], activated monocytes from 432 individuals[47], and activated monocyte-derived dendritic cells from diverse populations[48]. The most significant variant-gene associations were extracted from each eQTL data set and were reported as overlapping if that variant was in high LD ($r^2 > 0.8$) with any of the lead SNPs in the PSC GWAS meta-analysis.

Modelling PSC and IBD genetic risk. Association summary statistics from the European arm of the latest International IBD Genetics Consortium study[26] were downloaded. Where available, we used results from their combined GWAS plus Immunochip follow-up study and otherwise used those from the GWAS analysis. Definition of the 231 significantly associated loci as CD, UC or both (IBD) was taken from Liu *et al.*[26]. Owing to the limited availability of relevant subphenotype data within the IIBDGC data, we could not identify the 3–5% of IBD cases that we expect to have PSC. Including these individuals as IBD cases in our comparisons lowered our power to detect differences between the two diseases.

Causal variant co-localization analysis. To identify causal variants in disease-associated loci that are shared between diseases, we used a summary-statistic-based Bayesian test of colocalization (COLOC), implemented in R[27]. Briefly, COLOC generates posterior probabilities for five different hypotheses: (i) no association to either disease, (ii) association to disease 1 but not disease 2, (iii) association to disease 2 but not disease 1, (iv) association to both disease 1 and 2 but independent causal variants and (v) association to both disease 1 and 2 with a common causal variant. Only SNPs present in all the cohorts (PSC, CD, UC and IBD) were included in the analysis and associated regions were defined as 1-Mb regions with the most associated SNP at the center. Within each region, we calculated the r^2 between the PSC lead SNP and the SNP most associated with each of the other three diseases. Default priors were used for the probability of a SNP being (i) associated to an individual disease (1×10^{-4}) and (ii) causally associated to both diseases (1×10^{-5}). This prior probability of colocalization was more conservative in declaring distinct causal variants compared to a recent colocalization analysis across six immune-mediated disorders[49].

Heterogeneity of effects analysis. A formal heterogeneity of odds test was performed between PSC and IBD using the Cochran's Q test implemented in METAL[16] for all 18 PSC risk loci. The odds ratios and standard errors were obtained from our current PSC GWAS and the IIBDGC analysis[26]. A locus was declared to have significant heterogeneity of effects based on a threshold of $P = 2.78 \times 10^{-3}$ to account for multiple testing (Bonferroni correction applied to 5% significance threshold, $n = 18$ tests). To test whether the significant heterogeneity of effects were due to an overestimation of effect sizes in the smaller PSC cohort, we undertook a simulation study which demonstrated that the observed degree of heterogeneity is unlikely to occur by chance (**Supplementary Note**).

Genetic correlation analysis. Genome-wide SNP data from 12,933 IBD cases and 34,393 population controls of European descent was made available to us by the International IBD Genetics Consortium (IIBDGC). The quality control and imputation of these data using 1000 Genomes haplotypes has been previously described[26]. See **Supplementary Note** for details of the SNP and sample quality control (**Supplementary Table 13**) undertaken across the IIBDGC and PSC data to ensure compatibility and remove duplicated individuals. Individual level genotype data for PSC, CD, UC and IBD were used to estimate the proportion of variance in liability explained by SNPs genome-wide under a multiplicative model using the linear mixed model based restricted maximum likelihood (REML) method implemented in the GCTA software[31,50,51]. Ancestry principal components were calculated using genotype data from the 1000 Genomes project and were projected for all our cases and controls. The first 20 principal components were included as covariates in the linear mixed model. We assumed a prevalence of 0.0001 for PSC, 0.005 for CD and 0.0025 for UC. A bivariate extension of the linear mixed model[30], again implemented in GCTA[31], was used to estimate the additive covariance component and estimate r_G between PSC and either CD, UC or IBD.

In addition, we undertook an alternative genetic correlation analysis that used summary statistics and LD score regression[32]. Of the 7,458,430 SNPs that were shared between PSC and both IBDs, 1,102,210 HapMap3 SNPs were selected for the analysis as recommended. Then, pre-computed LD scores from the 1000 Genomes European data were used to run LD score regression to estimate genetic correlation.

Calculating comorbidity under a purely pleiotropic genetic model. Under a bivariate liability threshold model, where all disease risk is explained by additive genetics, the probability that an individual has disease 1, given that he has disease 2, is given by

$$P(L_1 > T_1 | L_2 > T_2) = \frac{P(L_1 > T_1, L_2 > T_2)}{P(L_2 > T_2)} = \frac{F((-L_1, -L_2) | \mu = (0,0), \mathbf{\Sigma} = \begin{pmatrix} 1 & h_1 h_2 r_g \\ h_1 h_2 r_g & 1 \end{pmatrix}}{K_2}$$

where K_i is the prevalence of disease i, $T_i = \Phi^{-1}(1 - K_i)$ is the liability threshold of disease i, h_i^2 is the heritability of disease i, r_g is the genetic correlation and $F(.)$ is the multivariate cumulative distribution function for normal distribution.

Data availability. Genome-wide summary statistics are available at http://www.ipscsg.org/. The University of Michigan HRS data are available from dbGaP under accession number phs000428.

35. Chapman, R.W.G. *et al.* Primary sclerosing cholangitis: a review of its clinical features, cholangiography, and hepatic histology. *Gut* **21**, 870–877 (1980).
36. Yimam, K.K. & Bowlus, C.L. Diagnosis and classification of primary sclerosing cholangitis. *Autoimmun. Rev.* **13**, 445–450 (2014).
37. Mailman, M.D. *et al.* The NCBI dbGaP database of genotypes and phenotypes. *Nat. Genet.* **39**, 1181–1186 (2007).
38. Bellenguez, C., Strange, A., Freeman, C., Donnelly, P. & Spencer, C.C. A robust clustering algorithm for identifying problematic samples in genome-wide association studies. *Bioinformatics* **28**, 134–135 (2012).
39. Price, A.L. *et al.* Principal components analysis corrects for stratification in genome-wide association studies. *Nat. Genet.* **38**, 904–909 (2006).
40. Delaneau, O., Zagury, J.F. & Marchini, J. Improved whole-chromosome phasing for disease and population genetic studies. *Nat. Methods* **10**, 5–6 (2013).
41. Howie, B.N., Donnelly, P. & Marchini, J. A flexible and accurate genotype imputation method for the next generation of genome-wide association studies. *PLoS Genet.* **5**, e1000529 (2009).
42. Pirinen, M., Donnelly, P. & Spencer, C.C.A. Efficient computation with a linear mixed model on large-scale data sets with applications to genetic studies. *Ann. Appl. Stat.* **7**, 369–390 (2013).
43. Olson, J.E. *et al.* The Mayo Clinic Biobank: a building block for individualized medicine. *Mayo Clin. Proc.* **88**, 952–962 (2013).
44. Moore, C. *et al.* The INTERVAL trial to determine whether intervals between blood donations can be safely and acceptably decreased to optimise blood supply: study protocol for a randomised controlled trial. *Trials* **15**, 363 (2014).

45. Gaffney, D.J. Global properties and functional complexity of human gene regulatory variation. *PLoS Genet.* **9**, e1003501 (2013).
46. GTEx Consortium. The Genotype-Tissue Expression (GTEx) project. *Nat. Genet.* **45**, 580–585 (2013).
47. Fairfax, B.P. *et al.* Innate immune activity conditions the effect of regulatory variants upon monocyte gene expression. *Science* **343**, 1246949 (2014).
48. Lee, M.N. *et al.* Common genetic variants modulate pathogen-sensing responses in human dendritic cells. *Science* **343**, 1246980 (2014).
49. Fortune, M.D. *et al.* Statistical colocalization of genetic risk variants for related autoimmune diseases in the context of common controls. *Nat. Genet.* **47**, 839–846 (2015).
50. Lee, S.H., Wray, N.R., Goddard, M.E. & Visscher, P.M. Estimating missing heritability for disease from genome-wide association studies. *Am. J. Hum. Genet.* **88**, 294–305 (2011).
51. Yang, J. *et al.* Common SNPs explain a large proportion of the heritability for human height. *Nat. Genet.* **42**, 565–569 (2010).

doi:10.1038/ng.3745