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 (GWAS) 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 \textit{UBASH3A}, with the protective allele (C) predicted to cause nonstop-mediated mRNA decay and lower expression of \textit{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) ($r_G = 0.04$) ($P = 2.55 \times 10^{-15}$). UC and CD were genetically more similar to each other ($r_G = 0.56$) than either was to PSC ($r_G < 1.0 \times 10^{-15}$). 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\textsuperscript{1}. There remains no effective medical therapy, and the majority of patients require orthotopic liver transplantation owing to the progressive nature of the disease\textsuperscript{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 (pan-colitis) compared to classical UC or CD\textsuperscript{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\textsuperscript{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\textsuperscript{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
projects13,14, 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 METAL15, 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 \( P < 5 \times 10^{-10} \) 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 \( UBASH3A \), 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 SIFT16 and PolyPhen2 (ref. 17), the Genome Wide Annotation of Variants (GWA) online tool18, 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 \( UBASH3A \) gene (rs11203203, 21:g.43836186G>A) that is in low LD \( (r^2 = 0.12) \) with rs1893592 has been associated with vitiligo24 and type-1 diabetes25, 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 ancestry26. Across each of the 18 non-HLA PSC risk loci, we used a Bayesian test of colocalization27 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 had 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 \( PRKDK2 \)), 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 arthritis28,29. 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 \( (P_{\text{empirical}} < 3.00 \times 10^{-4}) \) 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

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

| SNP          | Chromosome and position (bp) | Risk allele | RAF  | OR   | 95% CI        | GWAS      | Replication | Combined      | P value | Candidate causal gene |
|--------------|------------------------------|-------------|------|------|---------------|-----------|-------------|---------------|---------|-----------------------|
| 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^{-6} | 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 |
heterogeneity varies across SNPs. A failure to detect significant heterogeneity of odds does not necessarily indicate that effect sizes are equivalent because power to detect 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 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. Heterogeneity of odds tests were carried out using Cochran’s Q test. UC and CD odds ratios were obtained from the latest association studies conducted by the International IBD Genetics Consortium. 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.

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 (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 ($r_G = 0.04$) ($P = 2.55 \times 10^{-15}$) (Figure 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), 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 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. 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.uk10k.org; http://www.genome.gov/gwastudies/; https://www.immunobase.org/; http://eqtl.uchicago.edu/cgi-bin/gbrowse/eqtl/; http://www.genetixportal.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/phph2/; http://www.broadinstitute.org/mpg/grail/; https://data.broadinstitute.org/alkesgroup/LDSCORE.
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 (HRS) conducted by the University of Michigan, funded by the National Institute on Aging (grant numbers U01AG099740, R2CA036495 and R2CA039292) and accessed via dbGaP; Popgen 2.0, supported by a grant from the German Ministry for Education and Research (01EL1103); 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. The study was supported by the ISPCM, the National Institute of Diabetes and Digestive and Kidney Diseases (NIDDK RO1DK084960, KNL), the Wellcome Trust (098759/21/Z: LJ); 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 (BMBF) within the framework of the eMed research and funding concept (SysInflame grant 01ZX1306A) and the Chris M. Carlos and Catharine Nicole Joeksch 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 YAEI- founda- tion 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. Dahlén Solidol, B. Wümbeth, J. Andreas and L. Wenche Torbjörnen 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 Investigacion Biomedica en Red de Enfermedades Hepáticas y Digestivas (CIBERehd), grant PI10/1318 Instituto de Salud Carlos III, Ministerio de Ciencia e Innovación, and grant PI12/01448, from Ministerio de Economía y Competitividad, 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 grant1918.66.620). 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.I., N.K., T.S., J.G.-A. and C.A.A. performed statistical data analysis. S.-G.J., B.D.I., 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. coordinated the project and supervised the analyses. S.-G.J., B.D.I., 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.G.J., U.B., R.K.W., C.W., H.-U.M., P.M., A.K., O.C., P.L., E.G., K.S., C.M., J.S., W.H.O., D.J.R., J.D., A.P., A.F.G., J.E.E., S.S., C.C., L.B., V.A.L., J.A.O., K.B.C., K.Y.K., N.C., R.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 collections. All authors read and approved the final version of the manuscript.

COMPETING FINANCIAL INTERESTS

The authors declare no competing financial interests.

Reprints and permissions information is available online at http://www.nature.com/reprints/index.html.
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).

© 2017 Nature America, Inc., part of Springer Nature. All rights reserved.
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 3,302 population controls from Scandinavia and Germany were ascertained and genotyped using the Affymetrix Genome-Wide Human SNP Array 6.0 (Affymetrix) at three different centers7. 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 criteria35, with exclusion of secondary causes of sclerosing cholangitis. Commonly accepted clinical, radiological, endoscopic and histological criteria were also used for diagnosis and classification of IBD36. Genetic data from 12,595 individuals genotyped on the Illumina HumanOmni2.5-4v1 array (Omni2.5-4) as part of the For the Institute of Molecular Medicine (MMIF) genotyping project and Phenotypes (dbGaP37). 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 (AffySF, n = 2,205; AffyHS, n = 1,256; AffyAY, n = 472; IlluminaWTS, n = 1,227; IlluminaMAYO, n = 1,206; IlluminaCIDR, n = 12,595). Initially, SNPs out of Hardy-Weinberg equilibrium (HWE, P < 5 × 10−6) within PSC associated loci. The previously reported lead SNP (rs734874, P = 1.2 × 10−16) was removed from our analysis as it was poorly genotyped in our study. A total of 912,111 SNPs remained for analysis across the GWAS cohort.

Association analysis. A linear mixed model implemented in the MMM software 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 × 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 (r2 < 0.8) of previously reported SNP. In addition, 24 SNPs outside of established PSC risk loci with P < 5 × 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 Biobank43 and 1000 from the INTERVAL study44. 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 analysis. Poor genotype clusters were removed from the discovery and validation (P < 5 × 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 (r2 < 0.8) of previously reported SNP. In addition, 24 SNPs outside of established PSC risk loci with P < 5 × 10−6 were also included in the replication experiment. All cluster plots were manually inspected before SNP selection.

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 SNPTESs v2. Inverse-variance-weighted fixed effects meta-analyses implemented in METAL15 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, 2017 Nature America, Inc., part of Springer Nature. All rights reserved.

NATURE GENETICS
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\(^{46}\) 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\(^{25}\), purified B cells and monocytes from 283 individuals\(^{21}\), activated monocytes from 432 individuals\(^{47}\), and activated monocyte-derived dendritic cells from diverse populations\(^{24}\). 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 \((1 \times 10^{-6})\) and (ii) causally associated to both diseases \((1 \times 10^{-3})\). This prior probability of colocalization was more conservative in declaring distinct causal variants compared to a recent colocalization analysis across six immunemediated 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\(^{35}\). 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\(^{50}\), again implemented in GCTA\(^{31}\), was used to estimate the additive covariance component and estimate \(r_{CD}\) 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\(^{52}\). 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 pleotropic 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

\[
P(L_1 > T_1 \cap L_2 > T_2) = \frac{P(L_1 > T_1, L_2 > T_2)}{P(L_2 > T_2)} = F((-I_1 - L_1)\mu = (0, 0), \Sigma) = \left(1 - \frac{h_1 h_2 r_{g}}{K_2} \right)
\]

where \(K_i\) is the prevalence of disease \(i\), \(T_i = \Phi^{-1}(1 - K_i)\) is the liability threshold of disease \(i\), \(h^2_i\) is the heritability of disease \(i\), and \(r_{g}\) 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. et al. Primary sclerosing cholangitis: a review of its clinical features, cholangiography, and hepatic histology. Gut 21, 870–877 (1980).
36. Yinam, 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 in related individuals. Nat. Methods 5, e1000529 (2009).
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 for the next generation of genome-wide association 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 INTERWAL trial to determine whether intervals between blood donations can be safely and adequately 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).