Supplementary Material for:

Genome-wide association analysis in primary sclerosing cholangitis identifies two non-HLA susceptibility loci

Espen Melum[*], Andre Franke[*], Christoph Schramm, Tobias J. Weismüller, Daniel Nils Gotthardt, Felix A. Offner, Brian D. Juran, Jon K. Laerdahl, Verena Labi, Einar Björnsson, Rinse K. Weersma, Liesbet Henckaerts, Andreas Teufel, Christian Rust, Eva Ellinghaus, Tobias Balschun, Kirsten Muri Boberg, David Ellinghaus, Annika Bergquist, Peter Sauer, Euijung Ryu, Johannes Roksund Hov, Jochen Wedemeyer, Björn Lindkvist, Michael Wittig, Robert J. Porte, Kristian Holm, Christian Gieger, H.-Erich Wichmann, Pieter Stokkers, Cyriel Y. Ponsioen, Heiko Runz, Adolf Stiehl, Cisca Wijmenga, Martina Sterneck, Severine Vermeire, Ulrich Beuers, Andreas Villunger, Erik Schrumpf, Konstantinos N. Lazaridis, Michael P. Manns, Stefan Schreiber[§] and Tom H. Karlsen[§]

* These authors contributed equally to this work
§ These authors jointly directed this work

Nature Genetics: doi: 10.1038/ng.728

Supplementary Table 1. Allele frequencies and results from the association analyses for SNPs genotyped in the replication cohort.

					Genome-wide analysis				Replication analysis				
					Allele frequencies (Cases/Controls)				Allele frequencies (Cases/Controls)				
Chromosome	SNP	Position	Locus	Alleles	Scandinavia (332/262)	Germany (383/2700)	*P*-value[†]	OR (95% CI) [†]	Scandinavia (259/729)	Central Europe (498/891)	United States (268/554)	*P*-value[‡]	OR (95% CI)
1	rs12735793	29,509,563	*PTPRU*	A/G	0.09/0.02	0.11/0.08	1.6E-06	1.70 (1.37 - 2.11)	0.07/0.08	0.11/0.10	0.12/0.11	0.57	1.05 (0.88 - 1.26)
1	rs896403	164,442,667	1q24	T/C	0.17/0.28	0.18/0.21	1.5E-06	0.67 (0.57 - 0.79)	0.21/0.20	0.21/0.22	0.21/0.22	0.62	0.97 (0.85 - 1.10)
1	rs4391646	164,444,008	1q24	C/T	0.15/0.26	0.17/0.20	8.8E-07	0.65 (0.55 - 0.77)	0.19/0.18	0.17/0.20	0.19/0.21	0.15	0.90 (0.79 - 1.04)
1	rs6678400	164,444,384	1q24	T/G	0.15/0.26	0.17/0.20	9.5E-07	0.65 (0.55 - 0.77)	0.20/0.18	0.17/0.20	0.19/0.21	0.19	0.91 (0.79 - 1.05)
1	rs12043426	229,308,447	1q42	C/A	0.04/0.02	0.05/0.02	3.5E-06	2.23 (1.59 - 3.14)	0.04/0.04	0.03/0.03	0.02/0.03	0.88	1.02 (0.75 - 1.39)
2	rs6720394	111,705,843	*BCL2L11*	G/T	0.17/0.13	0.15/0.11	5.2E-06	1.60 (1.31 - 1.96)	0.14/0.12	0.13/0.11	0.16/0.10	0.0016	1.29 (1.10 - 1.51)
3	rs3197999	49,696,536	*MST1*	A/G	0.36/0.32	0.40/0.28	1.4E-09	1.51 (1.32 - 1.72)	0.35/0.31	0.33/0.25	0.38/0.30	1.5E-08	1.39 (1.24 - 1.56)
7	rs6971637	76,892,656	*PION*	T/C	0.03/0.02	0.03/0.02	1.6E-07	5.26 (2.83 - 9.78)	0.01/0.01	0.01/0.01	0.01/0.01	0.53	0.82 (0.44 - 1.52)
7	rs7791854	76,895,158	*PION*	T/C	0.02/0.01	0.02/0.01	1.0E-07	9.97 (4.28 - 23.24)	0.00/0.00	0.01/0.01	0.01/0.01	0.87	0.94 (0.44 - 1.99)
7	rs3807746	77,554,557	*MAGI2*	T/A	0.10/0.04	0.10/0.06	2.1E-06	1.73 (1.38 - 2.18)	0.08/0.07	0.08/0.08	0.07/0.09	0.73	0.97 (0.79 - 1.18)
7	rs6973565	131,593,184	*PLXNA4*	A/T	0.12/0.20	0.11/0.15	1.4E-06	0.62 (0.51 - 0.75)	0.14/0.16	0.15/0.16	0.16/0.15	0.69	0.97 (0.84 - 1.13)
7	rs13231950	131,647,489	*PLXNA4*	C/T	0.15/0.25	0.14/0.18	4.3E-06	0.65 (0.55 - 0.78)	0.17/0.19	0.20/0.17	0.17/0.17	0.41	1.06 (0.92 - 1.22)
8	rs7462577	135,977,697	8q24	A/G	0.24/0.17	0.31/0.25	1.9E-06	1.41 (1.22 - 1.62)	0.21/0.22	0.28/0.25	0.23/0.23	0.41	1.05 (0.93 - 1.19)
9	rs7038037	115,164,146	*BSPRY*	C/T	0.54/0.46	0.55/0.48	2.7E-06	1.36 (1.20 - 1.55)	0.49/0.51	0.50/0.49	0.49/0.49	0.71	0.98 (0.88 - 1.09)
10	rs706778	6,138,955	*IL2RA*	T/C	0.49/0.39	0.46/0.41	5.4E-06	1.35 (1.18 - 1.53)	0.52/0.40	0.43/0.42	0.46/0.44	0.15*	1.22 (0.93 - 1.59)*
10	rs3134883	6,140,731	*IL2RA*	A/G	0.39/0.32	0.38/0.31	8.5E-07	1.40 (1.22 - 1.59)	0.44/0.31	0.33/0.32	0.33/0.32	0.14*	1.24 (0.93 - 1.66)*
10	rs4147359	6,148,445	*IL2RA*	A/G	0.43/0.34	0.41/0.34	4.7E-06	1.36 (1.19 - 1.55)	0.46/0.35	0.36/0.35	0.39/0.35	0.060*	1.26 (0.99 - 1.61)*
10	rs10905718	6,154,862	*IL2RA*	G/A	0.39/0.32	0.38/0.30	2.4E-07	1.43 (1.25 - 1.64)	0.43/0.32	0.33/0.32	0.35/0.32	0.11*	1.23 (0.95 - 1.60)*
11	rs10891130	109,869,529	*FDX1*	A/G	0.02/0.01	0.03/0.01	5.0E-06	3.63 (2.09 - 6.31)	0.03/0.02	0.04/0.03	0.03/0.03	0.18	1.24 (0.91 - 1.68)
11	rs1793660	126,293,797	*KIRREL3*	G/A	0.24/0.17	0.28/0.21	3.6E-06	1.42 (1.22 - 1.64)	0.21/0.23	0.22/0.21	0.24/0.24	0.91	1.01 (0.89 - 1.14)
11	rs12808353	133,621,146	*VPS26B*	T/G	0.06/0.04	0.07/0.04	2.4E-06	1.92 (1.46 - 2.52)	0.04/0.03	0.04/0.04	0.03/0.04	0.87	0.98 (0.73 - 1.30)
12	rs11168249	46,494,635	*HDAC7*	C/T	0.55/0.45	0.51/0.45	1.0E-06	1.39 (1.22 - 1.58)	0.53/0.51	0.50/0.47	0.48/0.49	0.19	1.07 (0.96 - 1.19)
13	rs9520835	107,754,318	*TNFSF13B*	A/G	0.21/0.29	0.24/0.29	9.0E-06	0.70 (0.59 - 0.82)	0.23/0.25	0.26/0.28	0.24/0.26	0.098	0.90 (0.80 - 1.02)
15	rs4321167	95,073,908	15q26	A/C	0.05/0.03	0.06/0.04	5.9E-06	2.91 (1.83 - 4.61)	0.05/0.03	0.04/0.04	0.04/0.04	0.31	1.15 (0.88 - 1.50)
17	rs17683107	9,765,127	*GAS7*	A/G	0.19/0.11	0.18/0.14	8.5E-07	1.52 (1.29 - 1.80)	0.13/0.15	0.15/0.13	0.13/0.16	0.76	0.98 (0.84 - 1.14)
18	rs12458015	51,456,733	*TCF4*	C/T	0.29/0.34	0.26/0.35	7.3E-06	0.72 (0.63 - 0.83)	0.28/0.31	0.32/0.35	0.32/0.31	0.17	0.92 (0.82 - 1.04)

Nature Genetics: doi: 10.1038/ng.728

Complete association results for SNPs that were genotyped in the replication cohort. For the genome-wide analysis, the allele frequencies were calculated based on allele dosages and are listed separately for the German and Scandinavian discovery panels. For the replication analysis, allele frequencies are given for all three panels making up the combined replication panel. Positions refer to NCBI's build 36.

OR, Odds Ratio; CI, Confidence Interval.

[†]Odds ratios and *P*-values derived from logistic regressions of allele dosages including the six first principal components from the principal components analysis as covariates.

[‡]For SNPs with uniform effect sizes the Cochrane-Mantel-Haenszel test was used for the meta-analysis[1], while for the calculations marked with *a random effects model[2] was used (Breslow-Day *P*-value cut-off at 0.05).

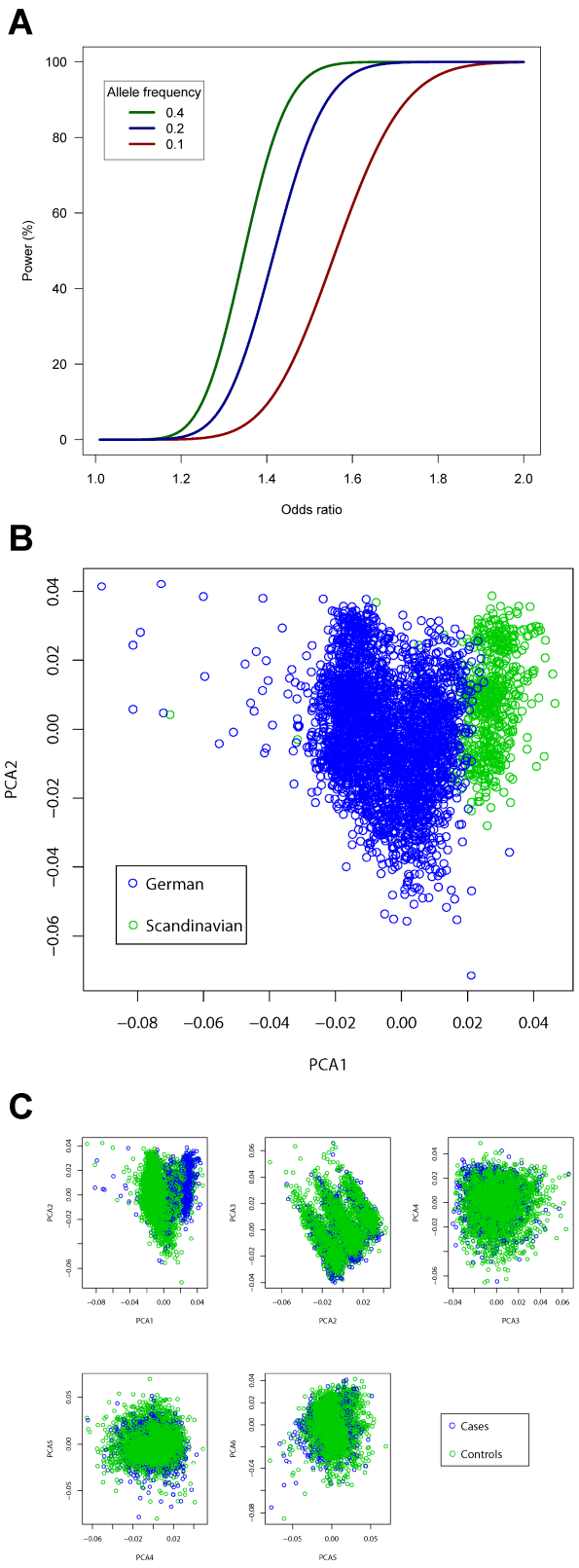
Nature Genetics: doi: 10.1038/ng.728

Supplementary Table 2. Stratified and regression analyses of non-HLA loci using rs3134792 as a marker for the HLA association.

SNP	Locus	*P*-value rs3134792 risk carriers	*P*-value rs3134792 non-risk homozygous	*P*-value logistic regression with rs3134792 as a covariate
rs6720394	*BLC2L11*	0.045	1.5×10^{-4}	2.5×10^{-5}
rs3197999	*MST1*	5.1×10^{-6}	1.3×10^{-4}	5.7×10^{-9}
rs706778	*IL2RA*	6.7×10^{-4}	0.0025	5.0×10^{-6}
rs3134883	*IL2RA*	0.0018	2.2×10^{-4}	7.3×10^{-7}
rs4147359	*IL2RA*	0.0089	5.8×10^{-4}	1.0×10^{-5}
rs10905718	*IL2RA*	0.0015	7.7×10^{-5}	3.6×10^{-7}

Nature Genetics: doi: 10.1038/ng.728

Supplementary Figure 1. Power to detect associated alleles in the discovery panel and principal components analysis.

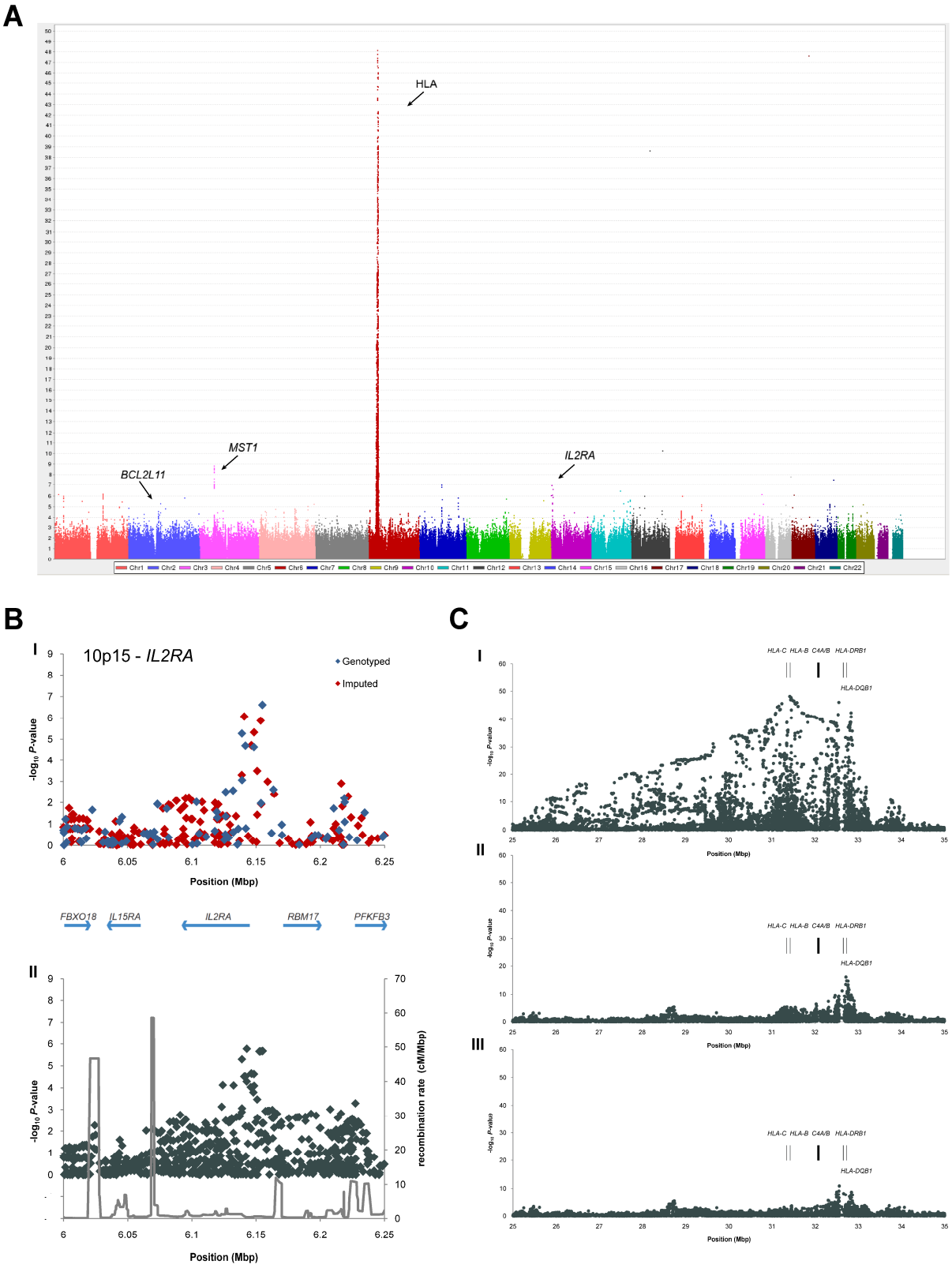


Nature Genetics: doi: 10.1038/ng.728

Panel A shows the power of the genome-wide analysis to detect association at a *P*-value threshold of 5×10^{-7} for different odds ratios using a log-additive genetic model. The power to detect association was 80% for a SNP with a frequency of 40% and an odds ratio of 1.42. **Panel B** shows the first two principal components generated from the EIGENSTRAT software[3] after removal of outliers along the first six principal components (see **Supplementary Methods** for further details). The **green** colored circles represent samples from patients and controls recruited at the Scandinavian centers while the **blue** circles represent patients and controls recruited at the German centers. **Panel C** shows the principal components that were used for removal of outliers and subsequently included as covariates in the association analysis. The colors of the circles indicate cases (**blue**) and controls (**green**). Plotted value represents the values generated after removal of outliers.

Nature Genetics: doi: 10.1038/ng.728

Supplementary Figure 2. Manhattan plot demonstrating the association results for all autosomal chromosomes and detailed association results at the *IL2RA* and 6p21 loci.

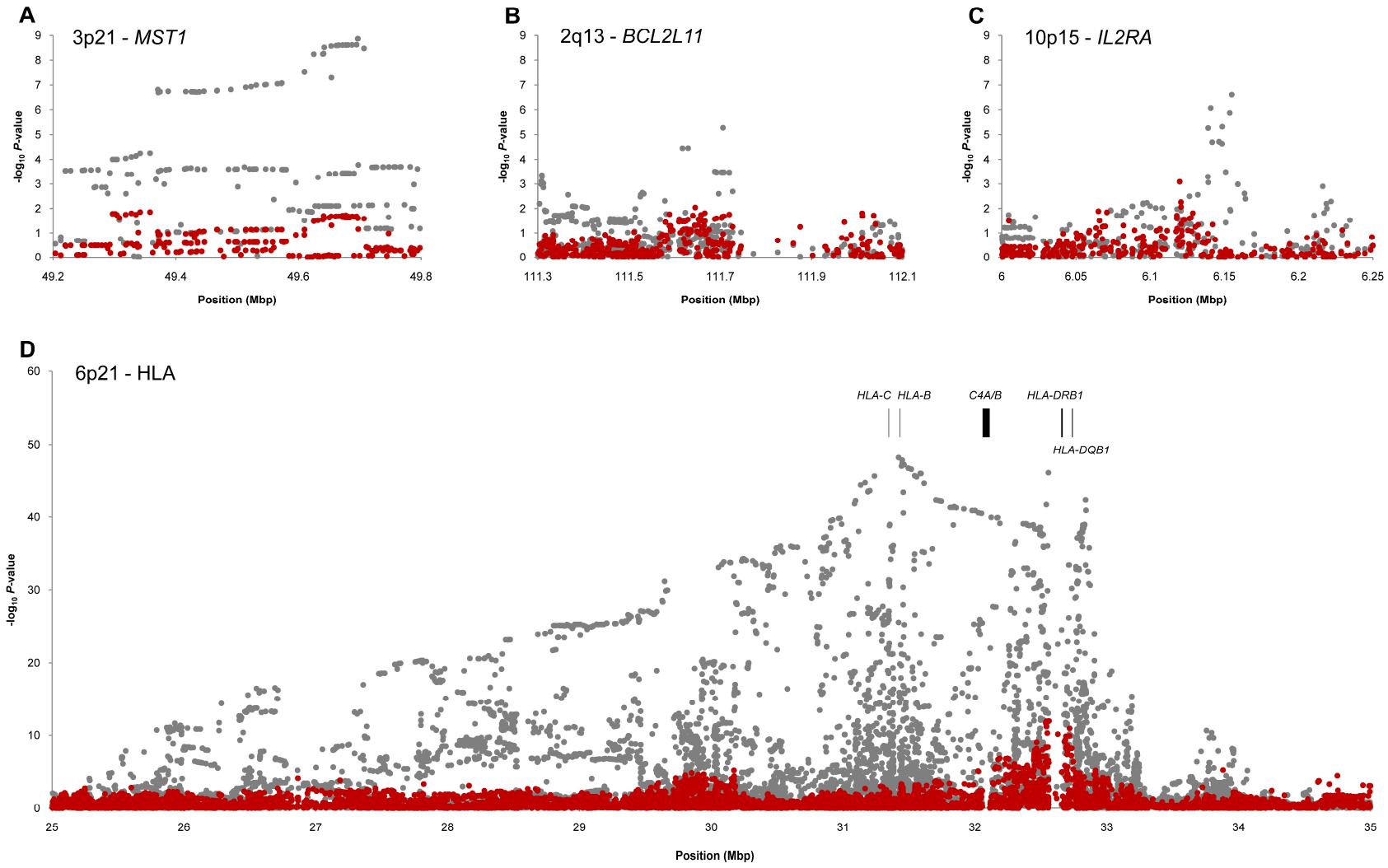


Nature Genetics: doi: 10.1038/ng.728

Panel A shows the negative decadic logarithm of the *P*-values derived from the population adjusted association statistics for all SNPs that were either directly genotyped or imputed. The SNPs are plotted against their chromosomal position on the X-axis. The plot was generated using Haploview[4]. **Panel B** shows association results at the *IL2RA* locus. In sub-panel I the association results from the genotyped and imputed markers are shown as the negative decadic logarithm of the *P*-values plotted against the genomic position (NCBI's build 36). Sub-panel II shows recombination rates (gray lines) derived from the Hapmap project along with the association results based on an imputed dataset using reference data from the 1000 Genomes project. **Panel C** shows association analysis results from 6p21 encompassing the HLA-complex as the negative decadic logarithm of the *P*-values obtained from the population adjusted association statistics with and without correction for the HLA association. Sub-panel I shows the results from the population corrected association statistics. Sub-panel II shows the population corrected association statistics corrected for rs3134792. Sub-panel III shows the population corrected association statistics corrected for rs3134792 and rs9272723. Of the patients carrying the associated allele at rs3134792 99% also carried the HLA-B*08 allele while 90% carried the HLA-DRB1*03 allele. Thus, rs3134792 is likely to represent a tag for the PSC associated HLA-B8-DR3 haplotype[5-7]. In **Panels B** and **C** all genotyped SNPs associated at a P-value$<10^{-4}$ were required to demonstrate high quality genotype clustering plots.

Nature Genetics: doi: 10.1038/ng.728

Supplementary Figure 3. Association results for the PSC associated genetic regions in ulcerative colitis.

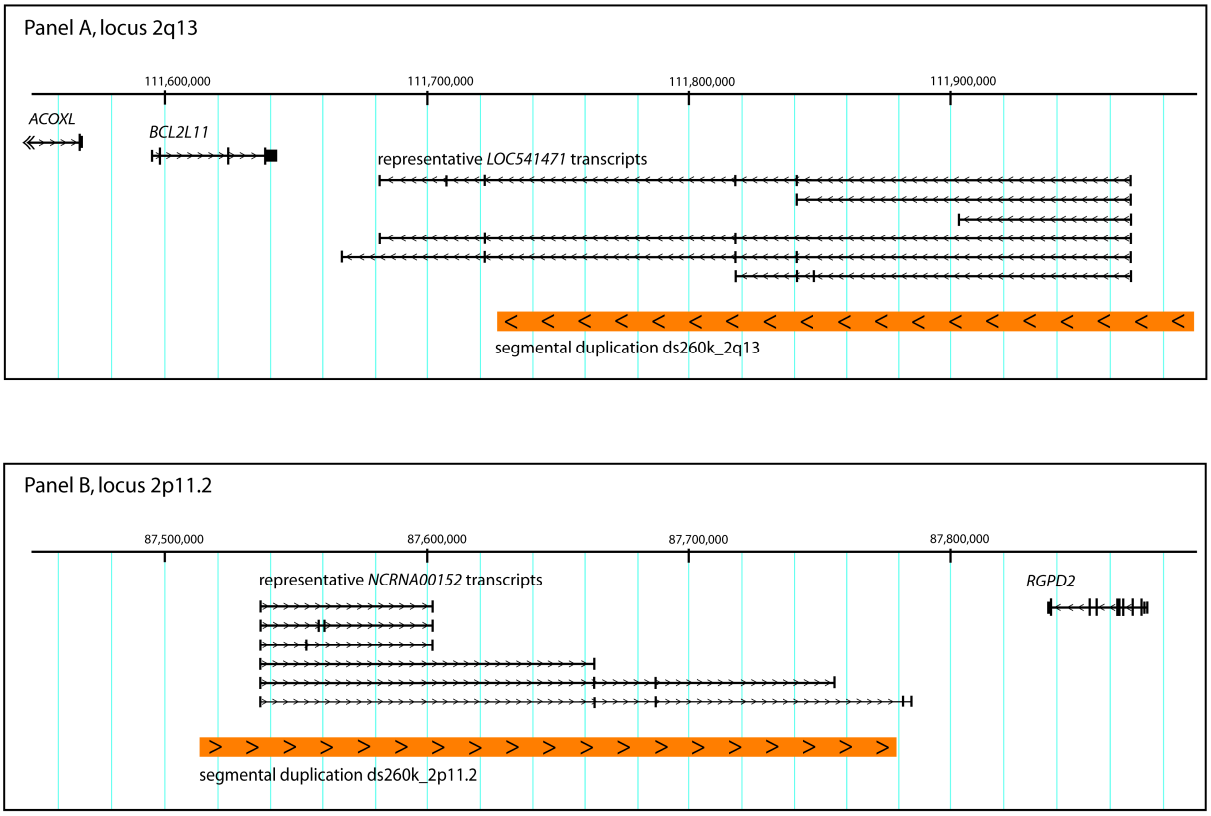


Nature Genetics: doi: 10.1038/ng.728

The **red** dots demonstrate the association statistics from the *MST1* **(Panel A)**, *BCL2L11* **(Panel B)**, *IL2RA* **(Panel C)** and HLA **(Panel D)** regions in a genome-wide association analysis in ulcerative colitis consisting of 1043 cases and 1703 controls[8]. The **gray** dots represent the association statistics for the same regions in PSC. From the ulcerative colitis study SNPs fulfilling the study specific quality criteria (for details please see ref. 8) and demonstrating high quality genotype clustering plots if associated at a *P*-value$<10^{-4}$ were included in the plot, while SNPs from the present study were required to fulfill the quality criteria of the present study. The SNPs in the ulcerative colitis study were tested for association with an allele-based Chi-squared test while population adjusted logistic regression was used in the present study.

Nature Genetics: doi: 10.1038/ng.728

Supplementary Figure 4. A duplicated segment at 2q13 and 2p11.2 is transcribed to non-protein-coding RNA of unknown function.



A segment on human chromosome 2, at the locus 2q13 (chr2: 111,726,975 - 111,993,116 in NCBI's build 36.1) (**Panel A**) is duplicated at the locus 2p11.2 (chr2: 87,512,075 - 87,778,622) (**Panel B**). The two segments span ~260 kilobases and are more than 99% identical (see ref. 9 and the Segmental Duplication Database). The duplicated segment at 2q13 (ds260k_2q13) is found in a stretch of DNA where the order of the genes is conserved (*BUB1-ACOXL-BCL2L11*-ds260k_2q13-*ANAPC1-MERTK*) in the mouse (chromosome 2) and bovine (chromosome 11) genomes. At the locus 2p11.2 the duplicated segment (ds260k_2p11.2) is found among several pseudogenes between the genes *RGPD1* and *RGPD2*, suggesting that the ancestral locus is at 2q13. The high degree of similarity between the two segments indicates that the duplication has occurred fairly recently in the human/great ape lineage[10].

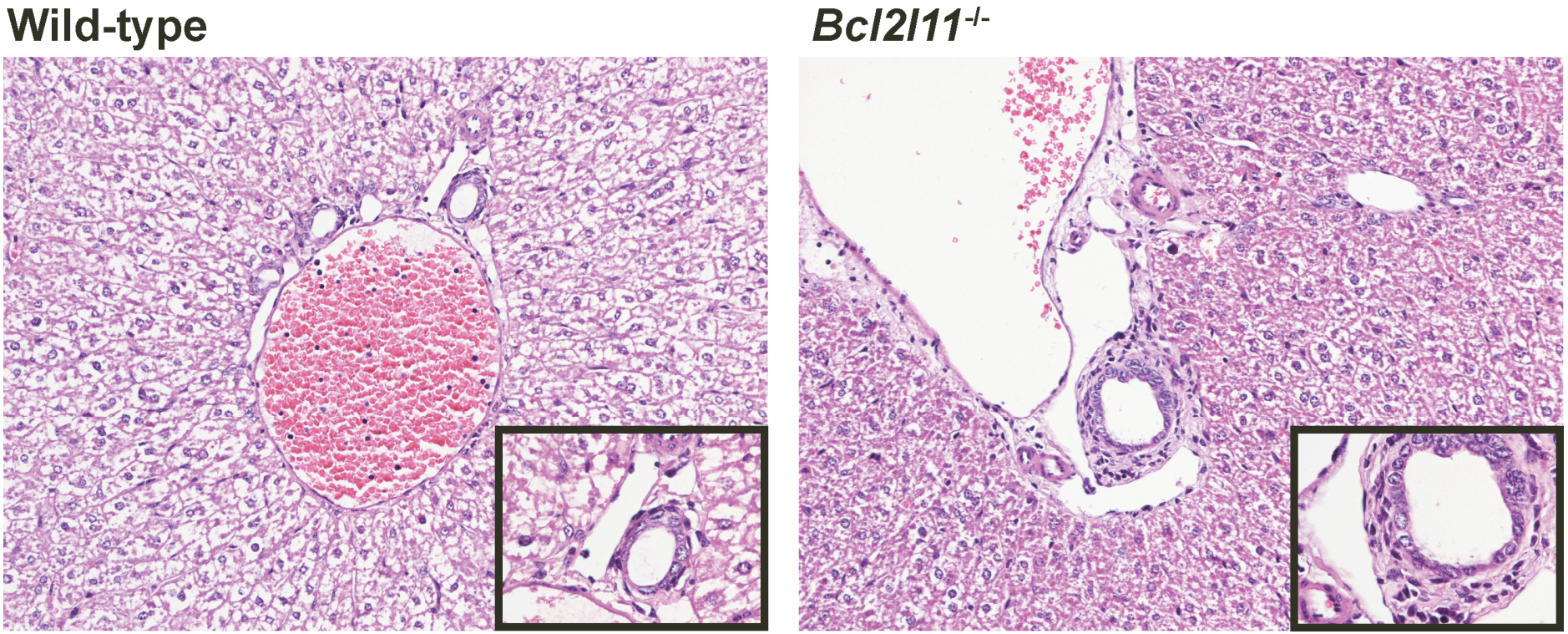
Nature Genetics: doi: 10.1038/ng.728

The putative genes *LOC541471* and *NCRNA00152* are located with their 5' exons within ds260k_2q13 and ds260k_2p11.2, respectively. Approximately 50-100 spliced transcripts, a large fraction containing polyadenylation tails, from each of these loci are available in public databases, demonstrating that a significant amount of transcription is taking place. There is also transcription at the mouse ortholog of *LOC541471*, *Gm14005*. *LOC541471* and *NCRNA00152* are processed into a large number of splice variants with the only common theme being that exon 1 is retained in the major fraction of splices transcripts. This is also the only exon showing any detectable homology with mouse *Gm14005* (~70% sequence identity). No homologous transcripts from any other species appear to be known.

The *LOC541471/NCRNA00152* and *Gm14005* transcripts only contain very short open reading frames and these have no detectable homology with any known protein, indicating that they are non-protein-coding genes. The spliced transcripts also contain a significant fraction of long interspersed repetitive elements (LINEs) and other retrotransposon genetic elements. Whether the *LOC541471/NCRNA0*0152 transcripts serve any biological function as non-protein coding RNA is not known. The only hint of a possible function for *LOC541471/NCRNA00152* appears to be the recent work by Kikuchi *et al.* [11] that shows one *LOC541471* transcript (*BX647931/AGD2*) to be down-regulated upon differentiation of mesenchymal stem cells to adipocytes.

Nature Genetics: doi: 10.1038/ng.728

Supplementary Figure 5. Presence of mononuclear cells in the portal tracts of *Bcl2l11*$^{-/-}$ mice.



8 week old gender- and age-matched mice were assessed for subtle histopathologic abnormalities in the livers and portal tracts in particular. 4/4 *Bcl2l11*$^{-/-}$ compared to 0/4 wild-type mice exhibited mononuclear cells in the vicinity of bile ducts, and representative sections are shown at 200x and 1000x (insert) magnification.

Nature Genetics: doi: 10.1038/ng.728

Supplementary Methods

Study panels

The diagnosis of primary sclerosing cholangitis (PSC) was based on standard clinical, biochemical, histological and cholangiographic criteria[12], including the exclusion of secondary causes of sclerosing cholangitis[13].

Supplementary Table 1 includes listing of the case/control panels included in the genome-wide analysis and subsequent replication analysis. The PSC patients in the Scandinavian discovery and replication panels were recruited on admission to Oslo University Hospital, Rikshospitalet, Oslo, Norway, Huddinge University Hospital, Stockholm, Sweden and Sahlgrenska University Hospital, Gothenburg, Sweden. The German PSC patients in the discovery and replication panels were recruited via the Grosshadern University Clinic, Munich, and the University Hospital of Heidelberg, Heidelberg, or through the Northern German biobank popgen for patients recruited at the University Medical Center Hamburg-Eppendorf, Hamburg, the Hannover Medical School, Hannover, the University Hospital of Mainz, Mainz, the Christian-Albrechts-University Hospital Kiel, Kiel, the University Hospital Freiburg, Freiburg and the Charité University Hospital Berlin, Berlin. The PSC patients from Belgium and the Netherlands in the Central European replication panel were recruited at the University Hospital Leuven, Belgium, the Academic Medical Center, Amsterdam, the Netherlands and the University Medical Center Groningen, Groningen, the Netherlands. The PSC patients and controls in the US replication panel are participants in the PSC Resource Of Genetic Risk, Environment, and Synergy Studies (PROGRESS). DNA samples from healthy controls of Scandinavian descent were randomly selected from the Norwegian Bone Marrow Donor Registry for the Scandinavian panels. DNA from the German controls for the German

Nature Genetics: doi: 10.1038/ng.728

discovery panel and Central European replication panel was received from blood donors through the Northern German biobank popgen and the Southern German population-based study KORA F4[14]. DNA from the Belgian controls in the Central European replication panel consisted of volunteers recruited via the University Hospital in Leuven, Belgium. Written informed consent was obtained from all study participants. The patient recruitment was approved by the ethics committees at each of the recruitment centers and the study was approved by the Medical Faculty of the Christian-Albrechts-University, Kiel, Germany and the South-Eastern Norwegian Regional Ethics Committee (S-93178 and S-08872b).

Power calculations

The power of the discovery panel to detect association for different allele frequencies and odds ratios was estimated for a log-additive model in the software package QUANTO 1.2.4 (ref. 15). Power graphs based on these calculations were plotted using the R statistical package.

Genome-wide genotyping and genotype calling

Genome-wide SNP genotyping was performed using the Affymetrix® Genome-Wide Human SNP Array 6.0 (Affymetrix, Santa Clara, CA, USA) according to manufactures protocols. The genotyping was performed at three different sites; Affymetrix® service facility (South San Francisco, CA, USA), Helmholtz Zentrum München, Genome Analysis Centre (GAC) (Munich, Germany) and ATLAS biolabs (Berlin, Germany) (total n=4121). The subsequent genotype calling was performed in batches defined on the basis of these different genotyping sites/recruitment centers. In brief, 5 µl of genomic DNA at 50 ng/µl was aliquoted to the

Nature Genetics: doi: 10.1038/ng.728

corresponding wells of two 96-well plates. One plate was digested with *Nsp*I and the other plate was digested with *Sty*I. The reaction was incubated at 37 °C for 2 hours and at 65 °C for 20 minutes to deactivate the enzymes. The digested DNAs were subsequently ligated to their respective *Nsp*I and *Sty*I adaptors and the ligated product amplified by PCR using a common primer. The *Nsp*I and *Sty*I PCR products were combined and purified by ethanol precipitation in combination with a membrane filter plate. Purified PCR product was further fragmented with DNase I then labeled with biotin. Labeled DNA was combined with hybridization mix and injected into the arrays for hybridization for 18 to 22 hours at 50°C. DNA samples were recovered from the arrays and washed and stained using Affymetrix® FS450 fluidic stations, and the stained arrays ultimately scanned using Affymetrix® GeneChip Scanner 3000 7G. Raw image files were converted into CEL-files using the Affymetrix® genotyping console.

Genotype calling was performed using the Birdseed v2 algorithm implemented in Affymetrix Power Tools version 1.10.2. Prior to genotype calling, samples with a low contrast-QC value (<0.40) were excluded (n=54) to avoid bias in the genotype calling of other samples. After genotype calling, the genotypes and their corresponding intensity values were converted into a custom database format by means of perl scripts. This database managed the generation of PLINK[16] input files and generated genotyping cluster plots for the disease associated SNPs considered for replication genotyping.

Quality control of genotype data

Following genotyping, all samples underwent stringent quality control procedures[17]. Samples with a call-rate <95% from the genotype calling process (n=110) were excluded from subsequent analysis. For four additional samples, there was a mismatch between the recorded

Nature Genetics: doi: 10.1038/ng.728

gender and the gender inferred from the genotype data and these samples were therefore also excluded. Duplicate samples were identified using identity-by-state (IBS) calculations implemented in the PLINK software version 1.06 (ref. 16) (3 cases and 17 controls). For each duplicate pair the sample with the lowest genotyping call-rate was excluded. Subsequently, the sample heterozygosity values in terms of the F measure for all the samples were plotted in a histogram and outliers identified based on this plot were excluded (n=179). Following removal of duplicate samples and heterozygosity outliers, a new global IBS estimation was performed. This procedure identified 28 first-degree relatives (PI_HAT=0.5), eight samples with a PI_HAT>0.1875 and three samples with an average pair-wise PI_HAT>0.05. For each pair of related samples the sample with the highest SNP call-rate was retained in the dataset.

The first step of the SNP-based quality control was exclusion of SNPs with a batch-wise call-rate <95%. Secondly, the samples were resorted according to panel allocation (**Supplementary Table 1**), and SNPs with minor allele frequency <1% and markers deviating from Hardy-Weinberg-Equilibrium (HWE) in the controls (P-value<10^{-4}) were excluded. The SNP based quality control was performed iteratively following the removal of samples for any cause (see above) to reduce potential bias on the allele frequency distribution from excluded samples. Ultimately, removal of population outliers was performed based on results from the principal components analysis of population heterogeneity (see separate paragraph). Removal of outliers along the six first principal components led to the exclusion of 38 samples, leaving a total of 715 PSC patients and 2962 controls for the association analysis.

Nature Genetics: doi: 10.1038/ng.728

Principal component analysis

The EIGENSTRAT software version 3.0 was used to investigate possible population structure in the dataset[3]. The EIGENSTRAT analysis was performed using SNPs passing the quality control measures described above in both the Scandinavian and German discovery panels. Mitochondrial SNPs and SNPs located on the X and Y chromosomes were not included in the EIGENSTRAT analysis. Also, SNPs in the region encompassing the human leukocyte antigen (HLA) complex (positions 25 Mbp to 35 Mbp) at chromosome 6p21 that were expected to bias the principal components analysis were excluded. The segregation of the German and Scandinavian discovery panels along the two first principal components are shown in **Supplementary Figure 1**.

We included principal components from the EIGENSTRAT analysis as covariates in the logistic regression analysis to correct for potential bias introduced by population structure. The number of principal components to include in model was decided by comparing the deviances when each principal component was sequentially added to the model. The first six principal components were significantly correlated with case/control status and outliers were therefore removed along these axes (see **Supplementary Figure 1**). After removal of outliers, the first six principal components were still significantly correlated with case/control status, and were subsequently used to correct for population structure in the association analysis.

Imputation

After the application of stringent quality control procedures in the German and Scandinavian discovery panels, imputation was performed in parallel for these two panels using the MACH

Nature Genetics: doi: 10.1038/ng.728

version 1.0.16 software[18]. The HapMap dataset release 22 from the Centre d'Etude du Polymorphisme Humain (CEPH) population was used as the reference for imputation in both populations. All the genotyped SNPs were appropriately adjusted to have their alleles specified along the forward strand in concordance with the HapMap reference dataset. To reduce the computing time, model parameter calculation for the imputation procedure was performed in a random subset of 350 individuals in each panel with the flags --greedy and --rounds 50. The random subsets included both cases and controls. Second, the parameters generated were used for imputation of the entire dataset with maximum likelihood imputation and the flags --mle, --mldetails and --greedy. Both the estimation of model parameters and the imputation itself were done separately for each chromosome.

For downstream analyses, we included imputed SNPs with a minor allele frequency >1%, HWE *P*-value>10^{-4} in the controls and good imputation quality (defined as a r^2>0.3) in both the German and Scandinavian discovery panels. The calculations of the HWE *P*-values were based on the best-guess genotypes provided by the imputation software and were calculated using the exact test implemented in PLINK version 1.06 (ref. 16). For all other analyses of the imputed SNPs, we used the allele dosages from the imputation procedure.

Association testing of the genome-wide data

A logistic regression procedure was used to test both the genotyped and the imputed SNPs for association. The allele dosages from the imputation procedure were used to account for uncertainty in the imputation procedure. The association analysis was performed for the entire discovery panel for subsets of 20,000 SNPs in parallel. The R statistical package version 2.9.1 was used for the association testing with custom scripts (freely available from the authors on

Nature Genetics: doi: 10.1038/ng.728

request). To adjust for the population structure the first six principal components (for the selection of these components, see separate section on principal components analysis) were included in a logistic regression procedure.

Imputation and analysis using 1000 Genomes data

For substantiation of the association signals at chromosomes 2, 3 and 10 additional imputation using reference data from prerelease of the 1000 Genomes project (August 2009 release of phased data) was performed. This reference data was downloaded from the MACH website. This dataset represents 112 haplotypes from the CEPH population with the exclusion of singleton SNPs only appearing in one individual. This imputation procedure was conducted in the same manner as described in the imputation section except for 1) Use of the --autoFlip flag to ensure concordance of allele alignments with the 1000 Genomes data, 2) SNPs flagged by MACH to have largely diverging allele frequencies in the target regions were removed prior to imputation and 3) To avoid high memory consumption the --compact switch was used. As suggested by the Abecasis group in the 1000 Genomes Imputation Cookbook the quality criteria for inclusion of imputed SNPs in down-stream analysis was increased to r^2>0.5 for the 1000 Genomes data.

Selection of markers for follow-up

All associated SNPs demonstrating a *P*-value<10^{-4} located outside of the HLA complex at 6p21 (defined as positions 25 Mbp to 35 Mbp on chromosome 6) were considered for replication genotyping. The genotyped SNPs were required to demonstrate high-quality genotype clustering plots (see section on genome-wide genotyping and genotype calling). The

Nature Genetics: doi: 10.1038/ng.728

imputed markers were required to have good imputation quality (see section on imputation). The number of SNPs considered for follow-up genotyping was thereafter reduced by taking account of linkage disequilibrium (LD) between closely situated markers with a clumping procedure in PLINK version 1.06 (ref. 16). The LD estimates used for the clumping procedure were obtained directly from the HapMap CEU reference dataset to avoid potential bias introduced by the imputation procedure. The thresholds used for clumping SNPs were an r^2>0.8 and a distance <50 kbp. For SNPs not genotyped in HapMap, the LD measures were manually assessed based on a combination of genotyped and imputed markers, and SNPs were excluded if a high correlation (r^2>0.8) could be consistently demonstrated. In particular, the 3p21 region was characterized by strong LD and we observed several highly significantly associated SNPs over a 0.34 Mbp region (see **Figure 1** of main manuscript and ref. 19). Since associations in this region had been previously reported, additional SNPs were manually excluded if they demonstrated strong LD outside the 50 kbp boundary. Two SNPs at the *IL2RA* locus and one SNP at the 1q24 locus were included for redundancy to ensure upfront genotyping success; these SNPs were also included in the association analysis.

Replication genotyping and association testing

Replication genotyping was performed using the Sequenom® system using matrix-assisted laser desorption/ionization time-of-flight (MALDI-TOF) mass spectrometry at the CIGENE genotyping platform (Norwegian University of Life Sciences, Ås, Norway). This method is based on allele assignment by mass-spectrometry[20]. An advantage of the method is that allele assignment is performed for each sample individually, making sample-sample variation in DNA concentration and quality less important.

Nature Genetics: doi: 10.1038/ng.728

Association testing of the replication data was performed in the PLINK package version 1.06 (ref. 16) with a Cochrane-Mantel-Haenzel (CMH) test[1]. Along with the CMH test a Breslow-Day test for heterogeneity of odds ratios was performed, and for SNPs demonstrating non-uniformity of the effect sizes, a random-effects model[2] implemented in the meta-library of the R statistical package was used. Combined analyses of the discovery and replication panels were performed with weighted Z-scores as described previously[21].

Analysis of the HLA region

The most strongly associated SNP in the HLA complex was rs3134792 in *HLA-B*. In a subset of the genotyped individuals (287 patients and 262 controls) data on HLA-types were available from previous studies[22,23] and used to examine relationship to reported associations. In order to correct for the effect of rs3134792, rs3134792 SNP data was included as a covariate in logistic regressions evaluating SNPs located in the HLA complex. This analysis demonstrated independent effects with the association signal peaking at rs9272723, and this SNP was therefore also included as a covariate in the third regression analysis of SNPs in the HLA complex.

***Bcl2l11*$^{-/-}$ mouse model and histological examination**

Bcl2l11$^{-/-}$ mice[24] on a C57Bl/6 background were maintained under specific pathogen-free conditions, as were C57Bl/6 mice. Age- and gender-matched *Bcl2l11*$^{-/-}$ and *Bcl2l11*$^{+/+}$ mice were harvested at an age of 8 weeks, livers collected in formalin, and mounted in paraffin.

Nature Genetics: doi: 10.1038/ng.728

5µm sections were stained with hematoxylin & eosin by standard methods. Sections were blindly assessed by an expert pathologist (F.A.O.) unaware of the genotypes. The Fisher exact test was used to compare observations in *Bcl2l11*$^{-/-}$ and wild-type mice.

URLs

Segmental Duplication Database, http://humanparalogy.gs.washington.edu

popgen, http://www.popgen.de

PSC Resource Of Genetic Risk, Environment, and Synergy Studies (PROGRESS), http://mayoresearch.mayo.edu/lazaridis_lab/genomics_of_psc.cfm

Birdseed algorithm, http://www.broadinstitute.org/mpg/birdsuite/birdseed.html

Affymetrix Power Tools, http://www.affymetrix.com/partners_programs/programs/developer/tools/powertools.affx

HapMap data, http://www.hapmap.org

R statistical package, http://www.r-project.org

1000 Genomes project, http://www.1000genomes.org

MACH website with 1000 Genomes data, http://www.sph.umich.edu/csg/abecasis/MACH/download/1000G-Sanger-0908.html

1000 Genomes Imputation Cookbook, http://genome.sph.umich.edu/wiki/MaCH:_1000_Genomes_Imputation_Cookbook

Nature Genetics: doi: 10.1038/ng.728

Reference List

1. Mantel, N. & Haenszel, W. Statistical aspects of the analysis of data from retrospective studies of disease. *J. Natl. Cancer Inst.* **22**, 719-748 (1959).
2. DerSimonian, R. & Laird, N. Meta-analysis in clinical trials. *Control Clin. Trials* **7**, 177-188 (1986).
3. Price, A.L. *et al.* Principal components analysis corrects for stratification in genome-wide association studies. *Nat. Genet.* **38**, 904-909 (2006).
4. Barrett, J.C., Fry, B., Maller, J. & Daly, M.J. Haploview: analysis and visualization of LD and haplotype maps. *Bioinformatics* **21**, 263-265 (2005).
5. Spurkland, A. *et al.* HLA class II haplotypes in primary sclerosing cholangitis patients from five European populations. *Tissue Antigens* **53**, 459-469 (1999).
6. Schrumpf, E. *et al.* HLA Antigens and Immunoregulatory T-Cells in Ulcerative-Colitis Associated with Hepatobiliary Disease. *Scand. J. Gastroenterol.* **17**, 187-191 (1982).
7. Donaldson, P.T. & Norris, S. Immunogenetics in PSC. *Best Pract. Res. Clin. Gastroenterol.* **15**, 611-627 (2001).
8. Franke, A. *et al.* Genome-wide association study for ulcerative colitis identifies risk loci at 7q22 and 22q13 (IL17REL). *Nat. Genet.* **42**, 292-4 (2010).
9. Bailey, J.A. *et al.* Recent segmental duplications in the human genome. *Science* **297**, 1003-1007 (2002).
10. Marques-Bonet, T. *et al.* A burst of segmental duplications in the genome of the African great ape ancestor. *Nature* **457**, 877-881 (2009).
11. Kikuchi, K. *et al.* Transcripts of unknown function in multiple-signaling pathways involved in human stem cell differentiation. *Nucleic Acids Res.* **37**, 4987-5000 (2009).
12. Chapman, R.W. *et al.* Primary sclerosing cholangitis: a review of its clinical features, cholangiography, and hepatic histology. *Gut* **21**, 870-877 (1980).
13. Abdalian, R. & Heathcote, E.J. Sclerosing cholangitis: a focus on secondary causes. *Hepatology* **44**, 1063-1074 (2006).
14. Wichmann, H.E., Gieger, C. & Illig, T. KORA-gen--resource for population genetics, controls and a broad spectrum of disease phenotypes. *Gesundheitswesen* **67 Suppl 1**, S26-30 (2005).
15. Gauderman, W.J. & Morrison, J.M. QUANTO 1.1: A computer program for power and sample size calculations for genetic-epidemiology studies (http://hydra.usc.edu/gxe). (2006).
16. Purcell, S. *et al.* PLINK: A tool set for whole-genome association and population-based linkage analyses. *Am. J. Hum. Genet.* **81**, 559-575 (2007).
17. Karlsen, T.H., Melum, E. & Franke, A. The utility of genome-wide association studies in hepatology. *Hepatology* **51**, 1833-42 (2010).
18. Li, Y. & Abecasis, G.R. Mach 1.0: rapid haplotype reconstruction and missing genotype inference. *Am. J. Hum. Genet.* **S79**, 2290 (2006).
19. Goyette, P. *et al.* Gene-centric association mapping of chromosome 3p implicates MST1 in IBD pathogenesis. *Mucosal Immunol.* **1**, 131-138 (2008).
20. Storm, N., Darnhofer-Patel, B., van den Boom, D. & Rodi, C.P. MALDI-TOF mass spectrometry-based SNP genotyping. *Methods Mol. Biol.* **212**, 241-262 (2003).

Nature Genetics: doi: 10.1038/ng.728

21. Skol, A.D., Scott, L.J., Abecasis, G.R. & Boehnke, M. Joint analysis is more efficient than replication-based analysis for two-stage genome-wide association studies. *Nat. Genet.* **38**, 209-213 (2006).
22. Karlsen, T.H. *et al.* Genome-Wide Association Analysis in Primary Sclerosing Cholangitis. *Gastroenterology* **138**, 1102-11 (2010).
23. Karlsen, T.H. *et al.* Particular genetic variants of ligands for natural killer cell receptors may contribute to the HLA associated risk of primary sclerosing cholangitis. *J. Hepatol.* **46**, 899-906 (2007).
24. Bouillet, P. *et al.* Proapoptotic Bcl-2 relative Bim required for certain apoptotic responses, leukocyte homeostasis, and to preclude autoimmunity. *Science* **286**, 1735-8 (1999).

Nature Genetics: doi: 10.1038/ng.728