Fine-mapping of genetic loci driving spontaneous clearance of hepatitis C virus infection

Hailiang Huang[1,2,3], Priya Duggal[4], Chloe L. Thio[5], Rachel Latanich[5], James J. Goedert[6], Alessandra Mangia[7], Andrea L. Cox[5], Gregory D. Kirk[5], Shruti Mehta[4,5], Joel N. Blankson[5], Jasneet Aneja[3], Laurent Alric[8], Sharyne M. Donfield[9], Matthew E. Cramp[10], Salim I. Khakoo[11], Leslie H. Tobler[12], Michael Busch[12], Graeme J. Alexander[13], Hugo R. Rosen[14], Brian R. Edlin[15], Georg M. Lauer[3], David L. Thomas[5], Mark J. Daly[1,2], Raymond T. Chung[3,*], Arthur Y. Kim[3,*]

1. Analytical and Translational Genetics Unit, Massachusetts General Hospital and Harvard Medical School, Boston, MA 02114, USA
2. Stanley Center for Psychiatric Research, Broad Institute of MIT and Harvard, Cambridge, MA 02142, USA
3. Department of Medicine, Massachusetts General Hospital, Harvard Medical School, Boston, MA 02114
4. Johns Hopkins University Bloomberg School of Public Health, Baltimore, MD 21205
5. Johns Hopkins University School of Medicine, Baltimore, MD 21205
6. Division of Cancer Epidemiology and Genetics, National Cancer Institute, Rockville MD 20852
7. IRCCS Casa Sollievo della Sofferenza Hospital, San Giovanni Rotondo, Italy;
8. Department of Medicine, Purpan Hospital, University of Toulouse III, Toulouse, France
9. Rho, Chapel Hill, NC 27517
10. South West Liver Unit, Plymouth Hospitals NHS Trust, Plymouth, United Kingdom
11. Henry Welcome Laboratories, Southampton General Hospital, Southampton, UK
12. University of California, San Francisco and Blood Systems Research Institute, San Francisco, CA 94118.
13. Cambridge University Hospitals NHS Foundation Trust and Addenbrooke's Hospital, Cambridge, United Kingdom;
14. University of Colorado, Aurora, Colorado 90045
15. State University of New York Downstate College of Medicine, Brooklyn, New York

*Co-corresponding authors

Abstract

Approximately three quarters of acute HCV infections evolve to a chronic state, while one quarter are spontaneously cleared. Genetic predispositions strongly contribute to the development of chronicity. We have conducted a genome-wide association study to identify genomic variants underlying HCV spontaneous clearance using Immunochip in European and African ancestries. We confirmed two previously reported significant associations, in the *IL28B/IFNL4*[1,2] and MHC regions, with spontaneous clearance in the European population. We further fine-mapped the MHC association to a region of about 50 kilo base pairs, down from 1 mega base pairs in the previous study. Additional analyses suggested that the association in the MHC locus might be significantly stronger for virus subtype 1a than 1b, suggesting that viral subtype may have influenced the genetic mechanism underlying the clearance of HCV.

The development of chronic viral infection represents a failure to mount an adequate innate and/or adaptive response to a specific pathogen. Infection with hepatitis C virus (HCV) in humans represents a paradigm of a dichotomous outcome of infection, as approximately three quarters of acute HCV infections evolve to a chronic state, but one quarter are spontaneously cleared[3]. As such, it is likely that genetic predispositions, especially at loci that regulate the innate and/or adaptive immune response, strongly contribute to the development of chronicity. A prior genome wide association study (GWAS) conducted by our consortium demonstrated striking associations of spontaneous resolution of HCV with polymorphisms near the IFN-L3 locus (*IL28B*) and in the HLA class II locus (Duggal *et al.*[4]).

The associations identified in Duggal *et al.* span large genomic regions, specifically 55,000 base pairs for the Il28B locus and > 1 mega base pairs for the HLA class II locus. Recently advances in genomic technologies allowed a more precise characterization of genetic associations and facilitated resolving these associations to much smaller genomic regions. Firstly, Immunochip[5], a customized array platform with deeper coverage of loci enriched in autoimmune diseases, provides coverage of additional genomic variants for an opportunity to explore with greater precision the contribution of these loci to the clearance of viral infection. Secondly, additional coverage of the MHC region can be gained using an imputation algorithm that takes into account the long range linkage-disequilibrium in MHC, and a large customized reference panel with improved coverage of the MHC region[6]. Thirdly, fine-mapping algorithms[7,8], designed with the goal to resolve known genetic associations to smaller sets of variants, can be used with the high density genomic data to further improve the precision of the genetic associations.

We therefore conducted an analysis of a large pool of spontaneous resolvers and chronic patients of HCV using the Immunochip platform, the SNP2HLA algorithm with the T1DGC MHC imputation reference panel[6], and a recently-developed fine-mapping algorithm[8,9] to (1) more precisely define the susceptible variant within the known associated loci; and (2) identify additional loci associated with clearance. Similar successes have been achieved in other diseases such as inflammatory bowel disease[8,10]. Additionally, we explored the hypothesis that there are shared mechanisms that define a "brisk" immunity able to confer both susceptibility to autoimmune disease and improved control of pathogens. We also examined the influence of region (North America versus European) upon associations with HLA within the European ancestry, as previous studies have shown variability of results, especially for the class I locus[11-14].

Results

The final dataset after QC has 166,537 variants for 527 cases/828 controls of European ancestry; and 171,161 variants for 75 cases/171 controls of African ancestry (Table 1). For each ancestry, we performed logistic regression under the additive model using the first two principal components as covariates. The QQ plot (Figure 1, using common variants with >2% minor allele frequency) and the genomic control (GC) factors (0.98 for the European ancestry and 0.92 for the African ancestry using designated null variants) indicate the effective control of the population stratification.

For European samples, we identified 8 genome-wide significant variants (p-value < 5E-8) in two loci (Figure 2 and Table 2). The variant on chromosome 19, rs8099917, shows the strongest association with spontaneous clearance (p-value=5E-16). Patients

carrying the minor allele, G, are on average 2.5x (odds ratio=0.39) less likely to spontaneously clear the virus compared to those with two copies of the C allele. This variant is roughly 7,000 base pairs upstream of the *IL28B* gene, and has been previously reported to be associated with HCV spontaneous clearance[4] and the response to chronic HCV therapy in Asian populations[15]. Previous studies have also shown an association between *IL28B* and interferon-based clearance of HCV[16]. Because the *IL28B/IFLN4* region was not designed as a high-density locus in Immunochip, we could not test other variants in this region for their association with HCV spontaneous clearance, and was unable to provide a better resolution in this locus.

The other genome-wide significant locus for the European samples is the major histocompatibility complex (MHC) locus. Genome-wide significant variants in this region are reported in Table 2 (before imputation). We used SNP2HLA[6] and a customized reference panel from a T1D study to impute missing variants, HLA alleles and amino acid residues for this region. We identified 12 SNPs and 5 amino acids that are genome-wide significant (Figure 3 and Table 3, boldfaced). No secondary signal in this region exceeded the suggestive significance threshold ($1x10^{-5}$) after conditioning on the primary signal. Therefore, all variants reported in Table 3 account for the same association signal. Using a fine-mapping algorithm described in another study[9], we constructed the 99% credible set, which is a set of variants that has 99% probability of having the causal variant in this locus (Table 3, full). Comparing with the previous study[4] which identified this association to a region of more than 1 mega base pairs, we mapped this association to a much smaller region of 50,562 base pairs.

Neither the MHC nor the *IL28B* locus was genome-wide significant in the African ancestry. Using the heterogeneity test (fixed-effect, implemented in the R metafor package), we found that neither the MHC locus nor the *IL28B* locus have significantly different effect size (p-values =0.47 and 0.29 respectively) across the two populations. Therefore, the difference in the significance is likely driven by the sample size and/or the allele frequency differences.

In addition to the genome-wide significant loci, we examined genes outside the HLA that have been previously associated with HCV spontaneous clearance[17]. Only genes *IFNG-AS1* (p-value=6E-4) and *STAT1* (p-value=3E-4) showed marginal evidence of association (p-value < 1E-3), and this effect was observed only in the European cohort. No genes reached the marginal p-value threshold in the samples of African ancestry. *IFNG-AS1* is a long noncoding RNA that is expressed in CD4 T cells and promotes Th1 responses[18]. STAT1 is one of the key mediators of the type I, II and III interferon responses.

Since HCV is particularly diverse, with up to a 30% difference at the amino acid level between major viral genotypes, the strain of infecting virus may influence HLA-mediated clearance[13]. Unfortunately, information regarding the virus genotype or subtype was not available in this study, especially in those persons who already cleared the virus, so a direct comparison is therefore not possible. However, an indirect comparison is possible by taking advantage of the observation that North American patients are much more likely to be infected with the 1a virus and European patients are much more likely to be infected by the 1b virus{Pawlotsky:1995eg}. We observed that the association in the class II MHC locus, after accounting for the sample size difference (**Methods**), is

136 stronger in North American samples than in European samples (Figure 4) with marginal
137 significance (p=0.05), This suggests that viral subtype may have influenced the genetic
138 mechanism underlying the clearance of HCV. Meta-analysis by cohorts confirms this
139 observation (Figure 5). We also interrogated the potentially protective effect of certain
140 SNPs associated with HLA class I alleles previously implicated in spontaneous clearance.
141 No SNP associated with class I was associated with genome-wide significance, including
142 those associated with *HLA B*27* subtypes (p-values > 0.05). The strength of association
143 with the SNP most closely linked with *HLA-B*57* and control of HIV-1 (rs2395029) was
144 not genome-wide significant but showed a marked difference by continent (North
145 America p-value=8.6E-4, Europe p-value=0.078, overall p-value=1.0E-4), suggesting
146 that any protective effect of this class I allele differs by region.
147 Autoimmune disorders have been reported to have shared genetic susceptibility
148 loci[19,20]. For each of 5 major autoimmune diseases, including inflammatory bowel
149 disease, systemic lupus erythematosus, rheumatoid arthritis, celiac disease and multiple
150 sclerosis, we listed all variants that reached p-value<0.001 (or the best variant) in this
151 analysis (Table 4). We found no shared variant after considering multiple testing. This
152 analysis was only performed in the European cohort because the African cohort has
153 limited power due to the sample size, and GWAS results in samples of African ancestry
154 for other autoimmune disorders is more limited. A full exploration of the hypothesis that
155 susceptibility to autoimmunity also confers ability to clear HCV will require a larger
156 sample size.
157 An alternate approach, taken by the International Genetics of Ankylosing
158 Spondylitis Consortium[21], is to search for the reported associations with other diseases in

159 loci having suggestive evidence (p-value < 1E-5), *i.e.*, the MHC and the *IL28B* loci in
160 this study. We only performed the search in *IL28B* because MHC has been already
161 implicated in many autoimmune disorders. We searched within 0.5Mb around the lead
162 SNP (rs8099917) in *IL28B* for associations with other diseases that have been reported in
163 the NHGRI GWAS catalog (search performed through UCSC Table Browser on August
164 24, 2014). Two SNPs were found to be in partial linkage disequilibrium (R^2>0.4) with
165 our lead SNP in *IL28B*, including rs12980275 (R^2=0.41) associated with lipid levels in
166 hepatitis C treatment[22], and rs12979860 (R^2=0.42) associated with chronic hepatitis C
167 infection[4]/response to hepatitis C treatment[16], which has been discussed in the previous
168 sections.
169
170 **Discussion**
171 We have conducted a genome-wide association study to identify genomic variants
172 underlying the HCV spontaneous clearance using immunochip. Consistent with previous
173 reports[4], two loci were found to be significantly associated with the HCV spontaneous
174 clearance in the European cohort. The immunochip design, the imputation pipeline
175 specifically designed for the MHC region and the novel fine-mapping algorithm
176 facilitated the accurate characterization of classical HLA types and allowed us to achieve
177 a higher resolution in the MHC region. Twelve SNPs and 5 amino acids in the MHC
178 region were found to be significantly associated and no secondary signal remains after
179 conditioning on the best SNP. Fine-mapping mapped this association to a region of about
180 50 kilo base pairs, down from 1 mega base pairs in the previous study. We found no
181 associated variants in the African cohort, probably due to different genetic background

182 (in the case of the *IL28B* locus) and limited sample size (in the case of the MHC locus).
183 Additional analyses suggested that the association in the class II MHC locus might be
184 significantly stronger for virus subtype 1a than 1b, suggesting that viral subtype may
185 have influenced the genetic mechanism underlying the clearance of HCV[13,23]. Additional
186 evidence, such as virus typing, is needed to confirm this finding.
187 Limitations of this study include inability to dissect SNPs near the *IL28B/IFLN4*
188 region, as this loci had not been previously implicated in autoimmune GWAS studies.
189 While the Immunochip did include rs8099917 as a surrogate for this region, additional
190 information regarding associations with rs12979860 and ss469415590 is not
191 available{Kumar:2016ea}. Also, this study was a fine-mapping exercise that narrowed
192 the MHC significantly but was not fully independent due to considerable overlap with the
193 previous GWAS.
194 Previous studies of GWAS data revealed that there are SNPs and loci with evidence
195 of association across multiple immune-mediated diseases[19]. We found several variants
196 that have suggestive and plausible evidence of associations with both HCV spontaneous
197 clearance and another autoimmune disorders. Despite the observation that none of these
198 variants are significant after the strict Bonferroni correction, they jointly confirm the
199 concept that shared genetic mechanisms underlie autoimmune disorders and suggest the
200 hypothesis that susceptibility to autoimmunity may also confer ability to clear HCV.
201 Fuller exploration of this hypothesis will require further analyses with larger sample
202 sizes.
203

204 **Methods**

Overview of samples

1,944 samples from 13 cohorts (ALIVE, BBAASH, HGDS, MHCS, Rosen and colleagues, REVELL, BAHSTION, SWAN, Toulouse, Cramp and colleagues, Hencore, Mangia and colleagues, UK Drug Use Cohort) were genotyped in this study, as previously described[4]. Self-clearance of HCV was coded as cases (718 samples) and persistence of HCV was coded as controls (1,180 samples). Samples with unidentified clearance status were not used (46 samples). All samples were genotyped using Illumina's Immunochip, a custom Infinium chip with 196,524 SNPs and small in/dels. A large number of these variants are in 187 high-density regions known to be associated with twelve autoimmune disorders and inflammatory diseases. Variants in these high-density regions include 289 established associations, variants from 1000 genome project low coverage pilot 1 study[24], and variants discovered in re-sequencing[25]. In addition, roughly 25,000 variants were included as replication of unrelated diseases as part of the WTCCC2 project, with the purpose of serving as null SNPs in analyses.

Sample ethnicities

To identify the sample ethnicities, we first constructed the principal component axes using Hapmap samples. 988 founders from Hapmap phase 3 (draft release 2)[26], including samples from ethnicities ASW, CEU, CHB, CHD, GIH, JPT, LWK, MEX, MKK, TSI and YRI were used. To calculate the principal components, only common variants that are also present in the Immunochip were used, and AT/GC SNPs were excluded to avoid ambiguous strand alignment. We performed LD pruning of the variants, resulting in a total of 15,525 variants used to create the principal components. The study samples were then projected to the principal component axes and assigned the ethnicities based on their

228 distance to the Hapmap samples. Out of 1,898 samples, 1,416 samples were mapped to
229 European ancestry, 225 samples were mapped to African ancestry and 227 samples were
230 admixtures and were not used in this study.

231 **Quality control**

232 QC was performed separately on samples of European and African ancestries separately.
233 Variants that failed the Hardy–Weinberg equilibrium test in controls (p-value ≤ 1E-5) or
234 had low call rate (≤95%) were identified, and 24,820 variants were removed in European
235 samples and 20,196 variants were removed in African samples. The remaining variants
236 were used to perform QC in samples. Samples were cleaned for having low call rate
237 (≤95%) or having high heterozygosity rate (> 3 standard deviations from the mean).

238 We then created a LD pruned dataset for calculating the identity by state (IBS)
239 matrix and the principal components. We pruned the variants using a sliding window of
240 50 variants, step size of 5 variants and variance inflation factor threshold of 1.25. There
241 were 20,782 variants in European samples, and 21,778 variants in African samples after
242 the pruning. The IBS matrix was calculated using this LD pruned dataset and checked for
243 sample relatedness. 28 duplicated samples in European cohorts and 9 duplicated samples
244 in cohorts of African ancestry have been identified and removed (pi_hat>0.9). The final
245 dataset has 527 cases and 828 controls for European cohorts, and 75 cases and 171
246 controls for African cohorts.

247 To correct for within European and within African population stratification, we
248 calculated the principal components for samples of European ancestry and African
249 ancestry, respectively. The first two principal components sufficiently control the

250 population stratification in both ancestries (results not shown) and were use in the
251 association analysis as covariates.

252 **Imputation**

253 Imputation of the MHC region was performed on QC cleaned data using SNP2HLA[6].
254 This software package takes advantage of the long-range linkage disequilibrium between
255 HLA loci and SNP markers across the MHC region and can perform accurate imputation
256 of classical HLA types starting from SNP genotype data. The reference panel was created
257 using the Type 1 Diabetes Genetics Consortium's high quality HLA reference panel
258 (roughly 5,000 European samples), which includes classical HLA alleles and amino acids
259 at class I (HLA-A, -B, -C) and class II (-DPA1, -DPB1, -DQA1, -DQB1, and -DRB1)
260 loci.

261 **Association test**

262 All association tests were performed in PLINK 1.07 [27] using the logistic regression. We
263 assumed additive models and used the first two principal components as covariates in the
264 regression. HCV spontaneous clearance was coded as case so an odds ratio > 1 indicates
265 the tested allele increases the probability of spontaneous clearance.

266 **Test whether an association has the same effect size across two cohorts**

267 For a cohort of sample size N, assume the association of interest has p-value of p, and the
268 chi-square statistic, calculated from p using the chi-square distribution, is χ^2. If this
269 association has the same effect size in the other cohort, the expected chi-square statistic
270 in the other cohort with population M is approximately $\chi^2 \bullet M/N$. The significance of
271 whether this association has the same effect size can be evaluated using the non-central

272 chi-square distribution with one degree of freedom and non-centrality parameter of

273 $\chi^{2\bullet}M/N$.

Acknowledgements: This project was funded in whole or in part by the National Institute of Allergy and Infectious Diseases (U19AI088791 and AI082630), the Office of AIDS Research through the Center for Inherited Diseases at Johns Hopkins University, the National Institute on Drug Abuse (R01DA033541, R01DA013324, R01DA12568, and R01DA04334). The MACS is funded by the National Institute of Allergy and Infectious Diseases, with additional supplemental funding from the National Cancer Institute (UO1-AI-35042, UL1-RR025005, UO1-AI-35043, UO1-AI-35039, UO1-AI-35040, and UO1-AI-35041). The REVELL cohort was funded by R01HL076902. Swan cohort was funded by R01-DA16159, R01-DA21550, and UL1-RR024996. The HGDS is funded by the National Institutes of Health, National Institute of Child Health and Human Development, R01-HD-41224.

Author contributions: Study design: AYK, RTC., DLT, PD, HH; Collecting samples and clinical information: AYK, CLT, RL, JJG, AM, ALC, GDK, SM, JNB, JA, LA, SMD, MEC, SIK, LHT, MB, GJA, HRR, BRE, GML; Performed Quality Control: JA, HH; Statistical analysis: HH, PD; Writing of the manuscript: AYK, RTC, HH.

Data availability: The data that support the findings of this study are available from the corresponding authors but restrictions apply to the availability of these data, which were used under license for the current study, and so are not publicly available. Data are however available from the authors upon reasonable request.

Author Information: The study protocols were approved by the institutional review board (IRB) at each center involved with recruitment. Informed consent and permission to share the data were obtained from all subjects, in compliance with the guidelines

297 specified by the recruiting center's IRB. The authors declare no competing financial

298 interests. Correspondence and requests for materials should be addressed to

299 AKIM1@mgh.harvard.edu or rtchung@partners.org

300 **Table 1. Variants and samples in this study**

	European ancestry	African ancestry
Variants		
Initial	191,357	191,357
HWE (p-value <1E–6)	-394	-300
Missingness (5%)	-24,426	-19,896
After QC	166,537	171,161
Samples		
Initial	1,416	227
Missingness (5%)	-24	-2
Heterozygosity	-9	-1
Duplicated	-28	-6
After QC	527 cases/828 controls	75 cases/171 controls

301 Negative numbers indicate numbers of variants or samples removed.
302

Table 2. Genome-wide significant associations

				European ancestry		**African ancestry**	
CHR	SNP	POSITION	Tested Allele	OR	p-value	OR	p-value
19	rs8099917	39743165	G	0.385	5.24E-16	0.555	0.2309
6	rs6457620	32663999	G	0.605	1.47E-09	0.758	0.1571
6	rs6457617	32663851	C	0.614	3.43E-09	0.758	0.1573
6	rs9275224	32659878	A	0.615	4.76E-09	0.685	0.0550
6	rs6932517	32678182	C	0.600	1.10E-08	0.557	0.0064
6	rs9357152	32664960	G	1.664	1.20E-08	1.484	0.1075
6	rs9378125	32679732	G	1.657	1.57E-08	1.437	0.1397
6	rs2858324	32660375	A	0.604	2.89E-08	0.590	0.0178

List of variants that have genome-wide significant association with HCV spontaneous clearance (before imputation). The genomic position is in HG18.

307 **Table 3. Associations in the 99% credible set in the MHC region**

CHR	SNP	POSITION	Tested Allele	OR	P	Probability
6	**rs6457617**	32771829	C	0.6172	8.06E-09	0.1197
6	**rs6457620**	32771977	G	0.6172	8.06E-09	0.1197
6	**rs9378125**	32787710	G	1.664	1.34E-08	0.0732
6	**rs5000632**	32771666	G	1.665	1.35E-08	0.0727
6	**rs9357152**	32772938	G	1.665	1.35E-08	0.0727
6	**rs9394113**	32773145	G	1.665	1.35E-08	0.0727
6	**rs9275224**	32767856	A	0.6229	1.60E-08	0.0616
6	**AA_DQB1_167_32737787_R**	32737787	A	1.717	2.05E-08	0.0483
6	**AA_DQB1_13_32740798_G**	32740798	A	1.708	3.03E-08	0.0331
6	**rs9275516**	32782621	A	0.6079	3.60E-08	0.0280
6	**SNP_DQB1_32737787**	32737787	T	1.7	3.79E-08	0.0266
6	**SNP_DQB1_32740798**	32740798	G	1.7	3.79E-08	0.0266
6	**SNP_DQB1_32740759_T**	32740759	P	1.7	3.79E-08	0.0266
6	**SNP_DQB1_32740760_A**	32740760	P	1.7	3.79E-08	0.0266
6	**AA_DQB1_13_32740798_A**	32740798	P	1.7	3.79E-08	0.0266
6	**AA_DQB1_26_32740759_Y**	32740759	P	1.7	3.79E-08	0.0266
6	**AA_DQB1_167_32737787_H**	32737787	P	1.7	3.79E-08	0.0266
6	rs6932517	32786160	C	0.6123	5.38E-08	0.0190
6	SNP_DQB1_32737837	32737837	A	0.627	8.09E-08	0.0128
6	SNP_DQB1_32737148	32737148	G	1.68	9.42E-08	0.0110
6	AA_DQB1_45_32740702	32740702	E	1.68	9.42E-08	0.0110
6	SNP_DQB1_32740702	32740702	T	1.68	9.42E-08	0.0110
6	SNP_DQB1_32742309_A	32742309	P	1.68	9.42E-08	0.0110
6	HLA_DQB1_0301	32739039	P	1.675	1.14E-07	0.0092
6	rs9469220	32766288	G	0.6393	2.24E-07	0.0048
6	rs2856717	32778286	T	0.6255	2.38E-07	0.0045
6	AA_DQB1_26_32740759_L	32740759	A	1.534	5.32E-07	0.0021
6	SNP_DQB1_32740759_A	32740759	A	1.534	5.32E-07	0.0021
6	SNP_DQB1_32740760_G	32740760	A	1.534	5.32E-07	0.0021
6	rs2858324	32768353	T	0.6397	7.11E-07	0.0016
6	rs2647012	32772436	A	0.6397	7.11E-07	0.0016

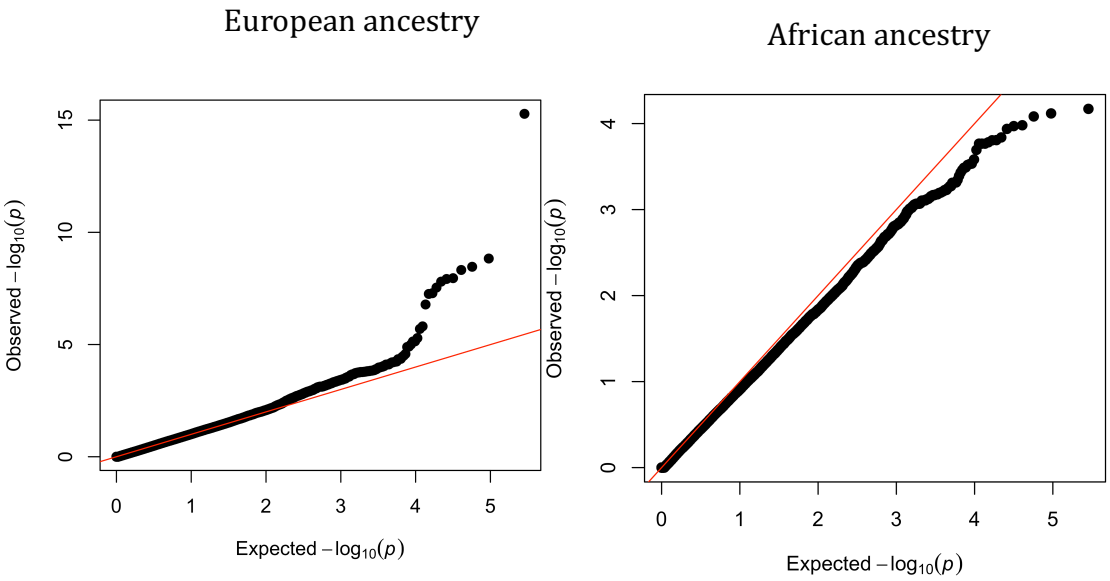
308 List of variants and HLA alleles in the MHC locus that are in the 99% credible set. Genome-wide
309 significant variants, HLA alleles and amino acid residues are boldfaced.

Table 4. Overlap with other autoimmune disorders

Disease	# variant [a]	Source	Variant	p-value	Direction[c]
IBD[28]	125/163	Table 1	rs2413583	0.0064	+
SLE[29]	5/6	Table 2	rs729302	0.2714	-
RA[30]	31/36	Table 2 and 3	rs11586238	0.0156	+
Celiac[31]	20/27	Table 2 [b]	rs2187668	0.0080	-
MS[32]	90/97	Table 1 and 2	rs9967792	0.0159	+

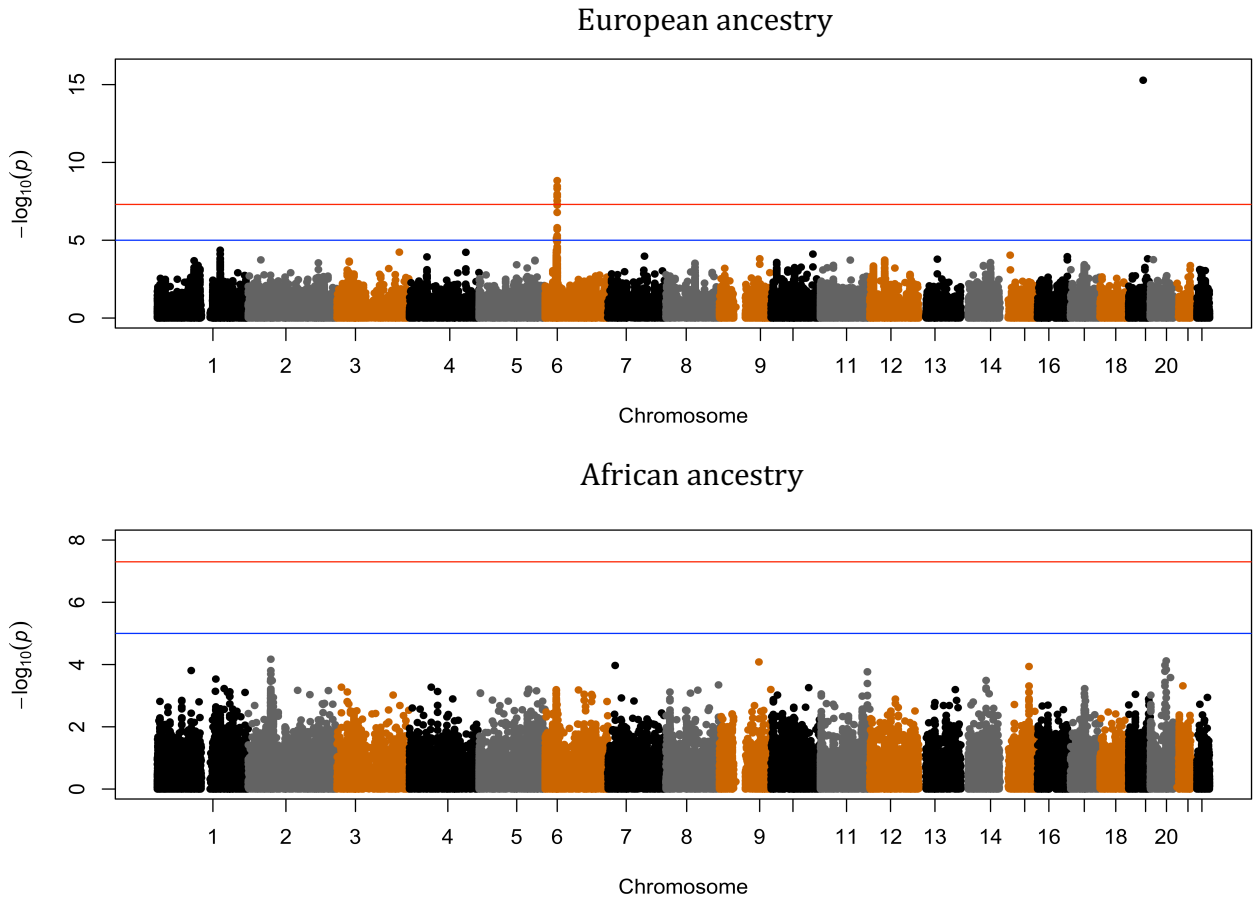
IBD: inflammatory bowel diseases, including Crohn's disease and ulcerative colitis; SLE: systemic lupus; RA: rheumatoid arthritis; Celiac: celiac disease; MS: multiple sclerosis. [a]the Number of variants in this study and the number of variants reported in the publication. Not all variants are available because of QC and chip design. [b] Only "Previously reported risk variants" and "New loci with genome-wide significant evidence" were included. [c] the '+' direction means the tested allele has the same direction of effect in both diseases, and the '-' direction means the opposite.

318 **Figure 1. QQ plot for cohorts of European and African ancestries.** The red line
319 indicates the null distribution. Only variants with minor allele frequency >2% were used
320 in this figure.



321

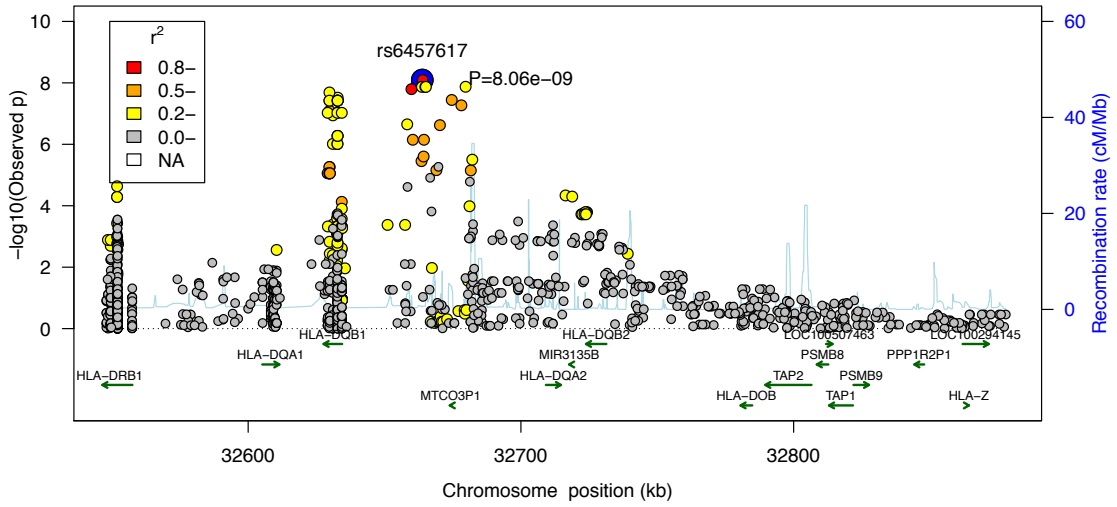
322 **Figure 2. Manhattan plot for cohorts of European and African ancestries.** The blue
323 horizontal line indicates the suggestive significance threshold, and the red horizontal line
324 indicates the genome-wide significance threshold.



325

326 **Figure 3. Regional association plot for the MHC class II region.** Color indicates the

327 linkage equilibrium with the top associated variant (rs6457620).

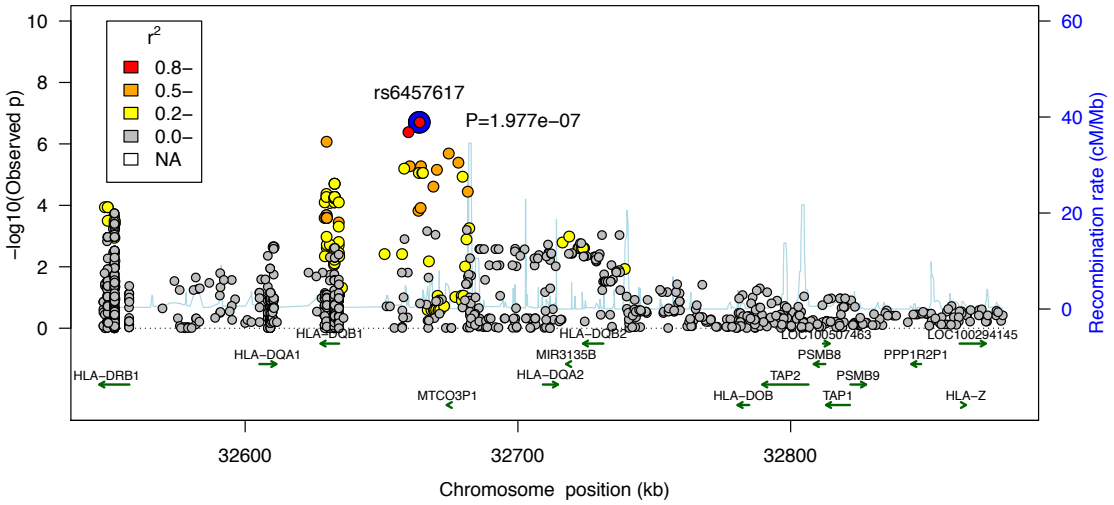


328

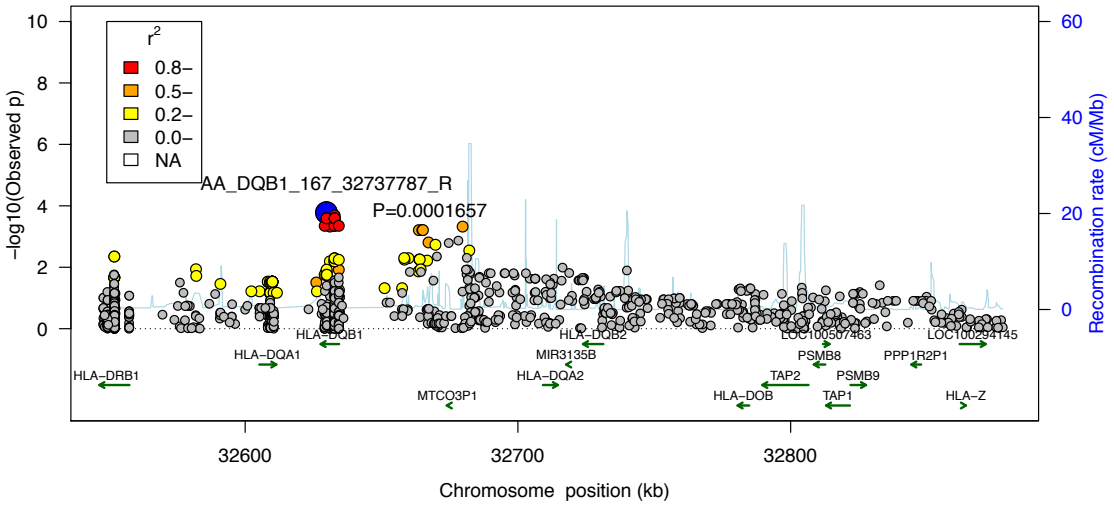
329 **Figure 4. Regional association plot for the MHC class II region in North American**

330 **samples (top) and European samples (bottom).** Color indicates the linkage equilibrium

331 with their respective top associated variant.

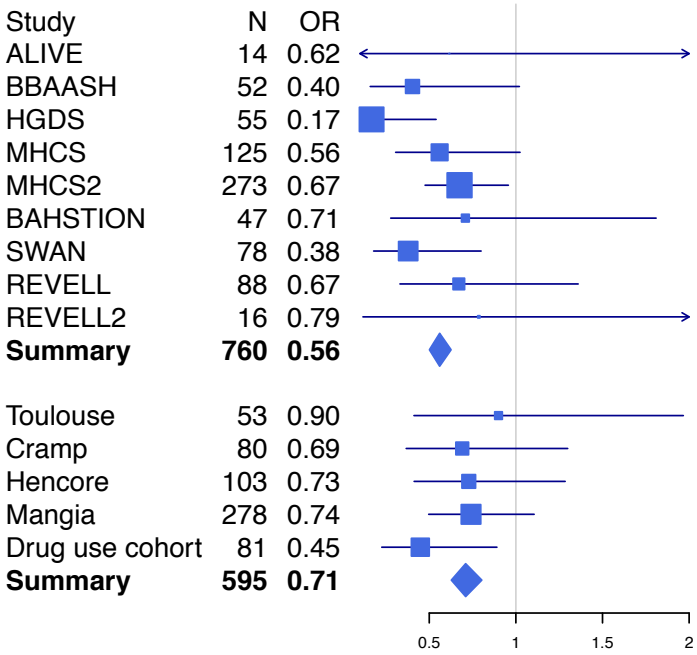


332



333

334 **Figure 5. Forest plot for the top MHC class II association (rs6457620).** Cohorts have
335 been grouped by the geographical locations of where they were collected: the top panel
336 includes cohorts collected in North America, and the bottom panel includes cohorts
337 collected in Europe.



338

339

Reference

1. Prokunina-Olsson, L. *et al.* A variant upstream of IFNL3 (IL28B) creating a new interferon gene IFNL4 is associated with impaired clearance of hepatitis C virus. *Nat Genet* **45,** 164–171 (2013).
2. O'Brien, T. R. *et al.* Comparison of functional variants in IFNL4 and IFNL3 for association with HCV clearance. *Journal of Hepatology* **63,** 1103–1110 (2015).
3. Shin, E.-C., Sung, P. S. & Park, S.-H. Immune responses and immunopathology in acute and chronic viral hepatitis. *Nat Rev Immunol* **16,** 509–523 (2016).
4. Duggal, P. *et al.* Genome-wide association study of spontaneous resolution of hepatitis C virus infection: data from multiple cohorts. *Ann. Intern. Med.* **158,** 235–245 (2013).
5. Cortes, A. & Brown, M. A. Promise and pitfalls of the Immunochip. *Arthritis Res Ther* **13,** 101 (2010).
6. Jia, X. *et al.* Imputing amino acid polymorphisms in human leukocyte antigens. *PLoS ONE* **8,** e64683 (2013).
7. Maller, J. B. *et al.* Bayesian refinement of association signals for 14 loci in 3 common diseases. *Nat Genet* **44,** 1294–1301 (2012).
8. Huang, H. *et al. Association mapping of inflammatory bowel disease loci to single variant resolution*. *bioRxiv* (2015). doi:10.1101/028688
9. Gormley, P. *et al.* Meta-analysis of 375,000 individuals identifies 38 susceptibility loci for migraine. *Nat Genet* **48,** 856–866 (2016).
10. Liu, J. Z. *et al.* Association analyses identify 38 susceptibility loci for inflammatory bowel disease and highlight shared genetic risk across populations. *Nat Genet* (2015). doi:10.1038/ng.3359
11. Thio, C. L. *et al.* HLA-Cw*04 and hepatitis C virus persistence. *J. Virol.* **76,** 4792–4797 (2002).
12. McKiernan, S. M. *et al.* Distinct MHC class I and II alleles are associated with hepatitis C viral clearance, originating from a single source. *Hepatology* **40,** 108–114 (2004).
13. Kim, A. Y. *et al.* Spontaneous Control of HCV Is Associated With Expression of HLA-B*57 and Preservation of Targeted Epitopes. *Gastroenterology* **140,** 686–696.e1 (2011).
14. Kuniholm, M. H. *et al.* Relation of HLA class I and II supertypes with spontaneous clearance of hepatitis C virus. *Genes and Immunity* **14,** 330–335 (2013).
15. Ochi, H. *et al.* IL-28B predicts response to chronic hepatitis C therapy - fine-mapping and replication study in Asian populations. *Journal of General Virology* **92,** 1071–1081 (2011).
16. Ge, D. *et al.* Genetic variation in IL28B predicts hepatitis C treatment-induced viral clearance. *Nature* **461,** 399–401 (2009).
17. Mosbruger, T. L. *et al.* Large-Scale Candidate Gene Analysis of Spontaneous Clearance of Hepatitis C Virus. *J INFECT DIS* **201,** 1371–1380 (2010).
18. Peng, H. *et al.* The Long Noncoding RNA IFNG-AS1 Promotes T Helper Type 1 Cells Response in Patients with Hashimoto's Thyroiditis. *Sci. Rep.* **5,** 17702 (2015).
19. Cotsapas, C. *et al.* Pervasive sharing of genetic effects in autoimmune disease.

PLoS Genet **7,** e1002254 (2011).

20. Richard-Miceli, C. & Criswell, L. A. Emerging patterns of genetic overlap across autoimmune disorders. *Genome Med* **4,** 6 (2012).
21. Consortium, I. G. O. A. S. Identification of multiple risk variants for ankylosing spondylitis through high-density genotyping of immune-related loci. *Nat Genet* **45,** 730–738 (2013).
22. Clark, P. J. *et al.* Interleukin 28B polymorphisms are the only common genetic variants associated with low-density lipoprotein cholesterol (LDL-C) in genotype-1 chronic hepatitis C and determine the association between LDL-C and treatment response. *J. Viral Hepat.* **19,** 332–340 (2012).
23. Nitschke, K. *et al.* HLA-B*27 subtype specificity determines targeting and viral evolution of a hepatitis C virus-specific CD8+ T cell epitope. *Journal of Hepatology* **60,** 22–29 (2014).
24. 1000 Genomes Project Consortium. A map of human genome variation from population-scale sequencing. *Nature* **467,** 1061–1073 (2010).
25. Trynka, G. *et al.* Dense genotyping identifies and localizes multiple common and rare variant association signals in celiac disease. *Nat Genet* **43,** 1193–1201 (2011).
26. International HapMap Consortium. A haplotype map of the human genome. *Nature* **437,** 1299–1320 (2005).
27. Purcell, S. *et al.* PLINK: A Tool Set for Whole-Genome Association and Population-Based Linkage Analyses. *The American Journal of Human Genetics* **81,** 559–575 (2007).
28. Jostins, L. *et al.* Host–microbe interactions have shaped the genetic architecture of inflammatory bowel disease. *Nature* **491,** 119–124 (2012).
29. Lee, Y. H., Bae, S.-C., Choi, S. J., Ji, J. D. & Song, G. G. Genome-wide pathway analysis of genome-wide association studies on systemic lupus erythematosus and rheumatoid arthritis. *Mol Biol Rep* **39,** 10627–10635 (2012).
30. Stahl, E. A. *et al.* Genome-wide association study meta-analysis identifies seven new rheumatoid arthritis risk loci. *Nat Genet* **42,** 508–514 (2010).
31. Dubois, P. C. A. *et al.* Multiple common variants for celiac disease influencing immune gene expression. *Nat Genet* **42,** 295–302 (2010).
32. International Multiple Sclerosis Genetics Consortium (IMSGC) *et al.* Analysis of immune-related loci identifies 48 new susceptibility variants for multiple sclerosis. *Nat Genet* **45,** 1353–1360 (2013).