nature genetics

Genome-wide association study identifies susceptibility loci for IgA nephropathy

Ali G Gharavi[1], Krzysztof Kiryluk[1], Murim Choi[2], Yifu Li[1], Ping Hou[1,3], Jingyuan Xie[1,4], Simone Sanna-Cherchi[1], Clara J Men[2], Bruce A Julian[5], Robert J Wyatt[6], Jan Novak[5], John C He[7], Haiyan Wang[3], Jicheng Lv[3], Li Zhu[3], Weiming Wang[4], Zhaohui Wang[4], Kasuhito Yasuno[2], Murat Gunel[2], Shrikant Mane[2,8], Sheila Umlauf[2,8], Irina Tikhonova[2,8], Isabel Beerman[2], Silvana Savoldi[9], Riccardo Magistroni[10], Gian Marco Ghiggeri[11], Monica Bodria[11], Francesca Lugani[1,11], Pietro Ravani[12], Claudio Ponticelli[13], Landino Allegri[14], Giuliano Boscutti[15], Giovanni Frasca[16], Alessandro Amore[17], Licia Peruzzi[17], Rosanna Coppo[17], Claudia Izzi[18], Battista Fabio Viola[19], Elisabetta Prati[20], Maurizio Salvadori[21], Renzo Mignani[22], Loreto Gesualdo[23], Francesca Bertinetto[24], Paola Mesiano[24], Antonio Amoroso[24], Francesco Scolari[18], Nan Chen[4], Hong Zhang[3] & Richard P Lifton[2]

We carried out a genome-wide association study of IgA nephropathy, a major cause of kidney failure worldwide. We studied 1,194 cases and 902 controls of Chinese Han ancestry, with targeted follow up in Chinese and European cohorts comprising 1,950 cases and 1,920 controls. We identified three independent loci in the major histocompatibility complex, as well as a common deletion of *CFHR1* and *CFHR3* at chromosome 1q32 and a locus at chromosome 22q12 that each surpassed genome-wide significance (*P* values for association between 1.59×10^{-26} and 4.84×10^{-9} and minor allele odds ratios of 0.63–0.80). These five loci explain 4–7% of the disease variance and up to a tenfold variation in interindividual risk. Many of the alleles that protect against IgA nephropathy impart increased risk for other autoimmune or infectious diseases, and IgA nephropathy risk allele frequencies closely parallel the variation in disease prevalence among Asian, European and African populations, suggesting complex selective pressures.

Chronic kidney disease is a major cause of morbidity and mortality, affecting 10–20% of the world population, with glomerulonephritis accounting for a considerable proportion of cases[1–3]. IgA nephropathy is the most common form of glomerulonephritis and the most common cause of kidney failure among Asian populations[2,4]. The diagnosis of IgA nephropathy requires documentation by kidney biopsy demonstrating proliferation of the glomerular mesangium with deposition of immune complexes predominantly composed of immunoglobulin and complement C3 proteins[3,5,6]. Registry data as well as autopsy and kidney-donor biopsy series suggest substantial variation in the prevalence of IgA nephropathy among populations with different ancestries: it is most frequent among Asians, with a disease prevalence as high as 3.7% detected among Japanese kidney donors[7], but is rare among individuals of African ancestry[5] and is of intermediate prevalence among Europeans (up to 1.3%)[6].

The pathogenesis of IgA nephropathy is uncertain[8,9]. The finding of IgA1 glycosylation abnormalities among European, Asian and African-American populations suggests a shared pathogenesis among different groups[10–15]. Moreover, familial aggregation of IgA nephropathy has been reported among people of all ancestries, suggesting

[1]Department of Medicine, Columbia University College of Physicians and Surgeons, New York, New York, USA. [2]Department of Genetics, Howard Hughes Medical Institute, Yale University School of Medicine, New Haven, Connecticut, USA. [3]Renal Division, Peking University First Hospital, Peking University, Institute of Nephrology, Key Laboratory of Renal Disease, Ministry of Health of China, Beijing, China. [4]Department of Nephrology, Ruijin Hospital, Shanghai Jiaotong University, School of Medicine, Shanghai, China. [5]Departments of Microbiology and Medicine, University of Alabama at Birmingham, Birmingham, Alabama, USA. [6]Children's Foundation Research Center at the Le Bonheur Children's Hospital, and the Division of Pediatric Nephrology, University of Tennessee Health Sciences Center, Memphis, Tennessee, USA. [7]Department of Medicine, Mount Sinai School of Medicine, New York, New York, USA. [8]Yale Center for Genome Analysis, Yale University School of Medicine, New Haven, Connecticut, USA. [9]Nephrology and Dialysis Unit, Ciriè Hospital, Torino, Italy. [10]Divisione di Nefrologia Dialisi e Trapianto, Azienda Ospedaliero-Universitaria Policlinico di Modena, Modena, Italy. [11]Laboratory on Pathophysiology, Uremia Istituto Giannina Gaslini, Genova, Italy. [12]University of Calgary, Alberta, Canada. [13]Divisione di Nefrologia e Dialisi, Istituto Scientifico Humanitas, Milan, Italy. [14]Department of Clinical Medicine, Nephrology and Health Science, Section of Nephrology, University of Parma, Parma, Italy. [15]Department of Nephrology, Ospedale di Gorizia, Gorizia, Italy. [16]Nephrology and Dialysis Unit, Ospedali Riuniti, Ancona, Italy. [17]Nephrology, Dialysis and Transplantation, Regina Margherita University Hospital, Turin, Italy. [18]University of Brescia and Second Division of Nephrology, Montichiari Hospital, Montichiari, Italy. [19]Division of Nephrology, Spedali Civili, Brescia, Brescia Italy. [20]Dialysis Center, Ospedale di Desenzano, Desenzano del Garda, Italy. [21]Renal Unit Careggi University Hospital, Florence, Italy. [22]Division of Nephrology, Ospedale degli Infermi, Rimini, Italy. [23]Department of Biomedical Sciences, University of Foggia, Foggia, Italy. [24]Department of Genetics, Biology and Biochemistry, University of Torino, Torino, Italy. Correspondence should be addressed to A.G.G. (ag2239@columbia.edu) or R.P.L. (richard.lifton@yale.edu).

Received 8 October 2010; accepted 15 February 2011; published online 13 March 2011; doi:10.1038/ng.787

Table 1 Summary of study cohorts

		Genotyped		After quality control	
Cohort	Ancestry	Cases	Controls	Cases	Controls
Discovery cohort	Han Chinese	1,228	966	1,194	902
Follow-up cohort 1	Han Chinese	740	750	712	748
Follow-up cohort 2	European	1,273	1,201	1,238	1,172
All cohorts combined		3,241	2,917	3,144	2,822

a genetic component to the disease[8,16]. So far, linkage studies have identified several loci predisposing individuals to IgA nephropathy, but the underlying genes are not known[8,16–18]. A single, unreplicated genome-wide association study (GWAS) in a small European cohort (533 cases) has reported association of IgA nephropathy with the major histocompatibility complex (MHC)[19].

We report a GWAS for IgA nephropathy in a cohort of 3,144 IgA nephropathy cases of Chinese and European ancestry, leading to the identification of five loci for this disease.

RESULTS

Study design and genotyping of discovery cohort

To detect loci conferring susceptibility to IgA nephropathy, we carried out a two-stage GWAS (**Table 1**). In the discovery phase, we carried out genome-wide genotyping on the Illumina 610 Quad platform in 1,228 biopsy-proven IgA nephropathy cases and 966 healthy controls of Chinese Han ancestry recruited from Beijing (**Table 1** and **Supplementary Table 1**). We further evaluated the top signals in the discovery phase in an independent cohort of Han Chinese descent (Shanghai cohort, 740 cases and 750 controls) and a European cohort of Italian and North American origin (combined by stratified analysis, 1,273 cases and 1,201 controls). Subsequently, we analyzed the Beijing, Shanghai and European cohorts together to identify genome-wide significant loci.

Genome-wide association analysis

In analysis of genome-wide genotyping data, we applied quality control filters, leading to elimination of 5% of samples because of low call rate, duplication, cryptic relatedness or gender mismatch and 16.8% of markers primarily because of low minor allelic frequency (<0.01, see **Supplementary Note** and **Supplementary Table 2**). After quality control, the genotyping call rate was 0.9992. We next applied the standard 1-degree-of-freedom Cochran-Armitage trend test to analyze 498,322 SNPs in the discovery cohort of 1,194 cases (650 males and 544 females, average age 31.1 years) and 902 controls (608 males and 294 females, average age 31.5 years). The quantile-quantile plot showed no global departure from the expected distribution of *P* values and the inflation factor (λ) was 1.024, indicating negligible population stratification (**Supplementary Fig. 1** and **Fig. 1**). Accordingly, principal component analysis (PCA) indicated that cases and controls were matched along the axes of statistically significant principal components, and PCA correction did not substantially change the distribution of the association statistic or the genomic inflation factor ($\lambda = 1.022$; **Supplementary Fig. 2** and **Supplementary Table 3**). We concluded that our association results were not biased by differences in ancestry or population structure between cases and controls.

The genome-wide association analysis showed 27 SNPs exceeding genome-wide thresholds for significance ($P \leq 5 \times 10^{-8}$; **Fig. 1**). These 27 signals all resided in a 0.54-Mb interval within the MHC on chromosome 6p21, with the top signal at rs9275596 ($P = 1.9 \times 10^{-12}$). Notably, 14 MHC SNPs with suggestive *P* values ($P = 5 \times 10^{-6}$ to $P = 1 \times 10^{-4}$) showed little or no linkage disequilibrium (LD) with rs9275596 (**Fig. 2a**).

Follow-up of top signals from discovery stage

After we removed MHC SNPs, additional loci remained that showed departure from the expected *P*-value distribution. We ranked signals on the basis of the false discovery rate and chose to follow up loci with $P \leq 1.3 \times 10^{-5}$, corresponding to a *Q* value ≤ 0.10 (**Supplementary Fig. 3**)[20]. Power calculations indicated that this strategy would provide 80% power to detect loci with allelic frequencies >0.10 and relative risk >1.5 with genome-wide significance ($P < 5 \times 10^{-8}$) in the combined cohort (**Supplementary Table 4**). In total, 65 SNPs from ten distinct loci met these criteria (including three potentially independent loci in MHC and two in the chromosome 22q12.2 interval). We genotyped the top-scoring SNPs and one additional SNP from each of these intervals in follow-up cohorts (20 SNPs total in 3,870 individuals after quality control; **Table 1**). We carried out tests of association within each cohort, followed by a combined analysis with the discovery cohort using Mantel's extension of Cochran-Armitage trend test (**Table 2** and **Supplementary Table 5**).

Five of the ten loci selected for follow up surpassed the threshold for significant genome-wide association: three loci within 6p21, one locus at 1q32 and one locus at 22q12.2 (**Table 2** and **Supplementary Tables 5,6**). Each signal had significant association with consistent effect size for the same risk allele in each individual cohort, with little evidence for heterogeneity.

The strongest association in the combined cohort was located within a ~170-kb interval that includes *HLA-DRB1*, *HLA-DQA1* and *HLA-DQB1* (rs9275596, odds ratio (OR) = 0.63, $P = 1.6 \times 10^{-26}$). This SNP has genome-wide significance with a consistent effect size in each cohort (**Table 2** and **Fig. 2b**) and has strong supporting association from a nearby SNP in strong LD (rs2856717).

This locus, however, did not explain all of the signal at 6p21. Conditioning for the effect of rs9275596 eliminated evidence for association for the majority of SNPs in close proximity, however two distinct loci maintained genome-wide significance. The second independent locus is defined by rs9357155 (which has an $r^2 = 0.01$ for correlation with rs9275596 in the combined cohort) and has an OR of 0.74 and a *P* value of 6.9×10^{-9} for association with IgA nephropathy after conditional analysis (**Table 3** and **Fig. 2c**). This SNP lies in a ~100-kb segment of LD and lies 128 kb centromeric to rs9275596. This LD segment contains the genes *TAP2*, *TAP1*, *PSMB8* and *PSMB9*, and the supporting SNP in this region (rs2071543) is a missense variant in *PSMB8* (p.Gln49Lys) that is at a position completely conserved

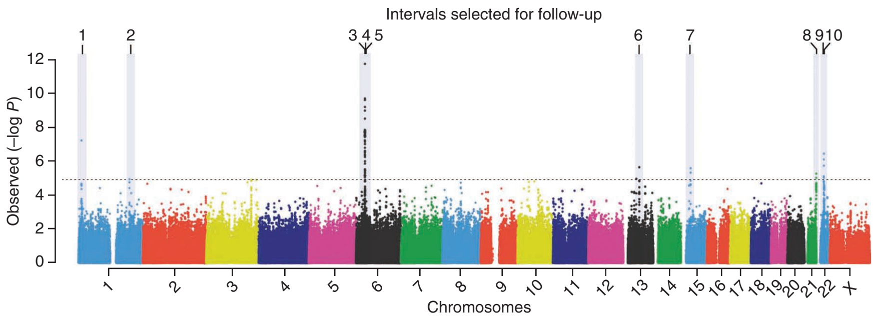


Figure 1 Manhattan plot of *P* values for SNP associations to IgA nephropathy. Observed *P* values versus chromosomal location; highlighted are the ten independent loci followed up in additional cohorts. Dashed line, follow-up threshold.

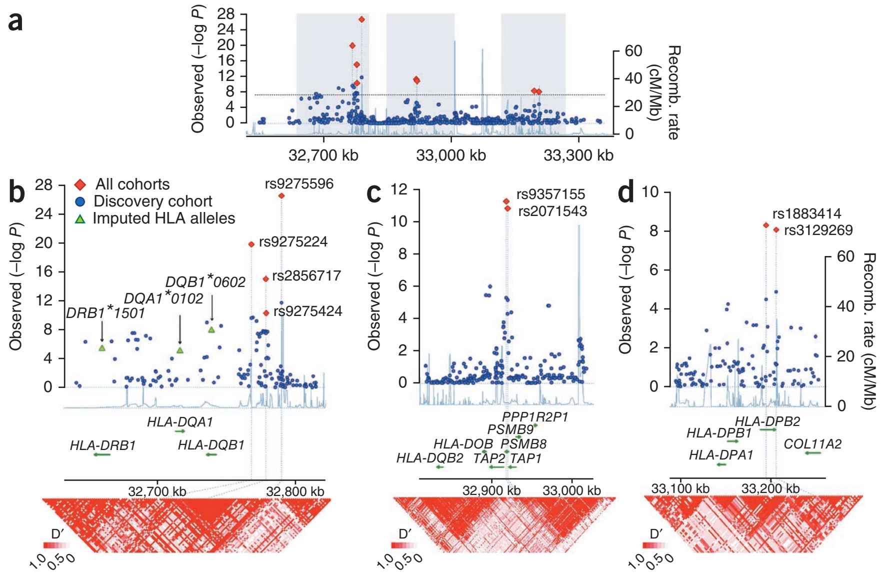


Figure 2 High-resolution view of MHC locus. *x* axes, physical distance (kb); left *y* axes, –log *P* for association statistics. The –log *P* values in the discovery and combined cohorts are blue circles and red diamonds, respectively. Right *y* axes, average recombination rates based on the phased HapMap haplotypes. Recombination rates, light blue lines. (**a**) The three intervals associated with IgA nephropathy reside within a 0.54 Mb segment on chromosome 6. Shaded areas correspond to regional plots in **b**–**d**. (**b**) Regional plot for interval containing *HLA-DQB1*, *HLA-DQA1* and *HLA-DRB1*. The classical HLA alleles imputed in the discovery cohort (green triangles) formed a protective haplotype, *DQB1*0602-DQA1*0102-DRB1*1501*. (**c**) Regional plot for the second MHC interval: SNPs typed in the combined cohorts reside within *PSMB8*. (**d**) Regional plot for the *HLA-DPB2*, *HLA-DPB1* and *HLA-DPA1* interval. Bottom of **b**–**d**, LD heat maps (*D*′) calculated based on the genotype data of the Beijing cohort.

among all orthologs (most distantly related ortholog is in platypus; **Tables 2,3**, **Fig. 2c** and **Supplementary Tables 7,8**).

After we conditioned for the effects of both rs9275596 and rs9357155, we found that a third locus within MHC, defined by rs1883414, which lies 400 kb centromeric to rs9275596 (and which showed r^2 = 0.005 and r^2 = 0.002 with rs9275596 and rs9357155, respectively), showed a conditioned OR of 0.77 and *P* value of 3.1 × 10^{-8} for association (**Table 3**). This signal, in the region of *HLA-DPA1*, *HLA-DPB1* and *HLA-DPB2*, is supported by a second SNP (rs3129269) and showed consistent effect size across cohorts (**Tables 2,3**, **Fig. 2d**, and **Supplementary Tables 7,8**).

To better delineate the risk associated with the MHC region and detect potential functional variants, we imputed classical HLA alleles in the discovery cohort[21] (**Supplementary Table 9**). This showed a genome-wide significant association with a protein-altering variant of known functional importance, the *DQB1*0602* allele (OR = 0.47, *P* = 6.6 × 10^{-9}). *DQB1*602* is in strong LD with another functional allele, *DRB1*1501*, but conditional analysis suggested that *DQB1*602* best explains this association signal (**Supplementary Table 10**). The strength of the *DQB1*602* association is probably underestimated because of the limitations of current imputation algorithms (sensitivity of 56.6% for detection of the *DQB1*602* allele; **Supplementary Table 11**).

A major signal outside the MHC locus resided in a 100-kb segment on chromosome 1q31-q32.1 containing *CFH* (encoding complement factor H) and the related *CFHR3*, *CFHR1*, *CFHR4*, *CHFR2* and *CFHR5* genes (rs6677604, OR = 0.68, *P* = 3.0 × 10^{-10} in the combined cohort). This locus was also the top signal in our genome-wide copy number polymorphism (CNP) analysis (**Supplementary Fig. 4** and **Supplementary Table 12**). The top SNP, rs6677604, is located in intron 12 of *CFH* and is supported by multiple highly correlated SNPs (**Fig. 3a** and **Table 2**). After controlling for rs6677604, there were no other independent signals in the entire *CFH* region. The association results at rs6677604 were far less significant under a recessive model (*P* = 5.6 × 10^{-5}), supporting an additive risk. The rs6677604 A allele is protective in all three cohorts but has a much higher allele frequency in Europeans (0.23 in European controls versus 0.07 in Chinese controls; **Table 2**). This allele perfectly tags a common deletion spanning *CFHR1* and *CFHR3* (*CFHR1,3Δ*)[22,23]. We confirmed the association of the rs6677604 A allele with *CFHR1,3Δ* in our cohort: PCR of multiple amplicons within *CFHR1* and *CFHR3* failed, and we did not

Table 2 Association results for ten SNPs representing five independent regions that reach genome-wide significance in combined analyses

			Beijing discovery cohort[a] *n* = 2,096 (1,194 cases, 902 controls)			Shanghai replication cohort[a] *n* = 1,460 (712 cases, 748 controls)			European replication cohort[b] *n* = 2,410 (1,238 cases, 1,172 controls)			All cohorts combined[b] *n* = 5,966 (3,144 cases, 2,822 controls)				
												OR				
Chr.	Location (kb)	SNP (minor allele)	MAF (cases/ controls)	OR	*P*	MAF (cases/ controls)	OR	*P* value	MAF (cases/ controls)	OR	*P*	Per allele	Het.	Hom.	*P*	*Q*
1	194,918	**rs3766404 (C)**	0.052/0.086	0.59	1.84 × 10^{-5}	0.078/0.080	0.98	8.18 × 10^{-1}	0.12/0.14	0.82	1.46 × 10^{-2}	0.77	0.79	0.45	4.24 × 10^{-5}	0.01*
1	194,953	**rs6677604 (A)**	0.041/0.073	0.55	1.20 × 10^{-5}	0.052/0.070	0.73	3.22 × 10^{-2}	0.17/0.23	0.71	1.19 × 10^{-5}	0.68	0.69	0.41	2.96 × 10^{-10}	0.17
6	32,778	**rs2856717 (T)**	0.19/0.26	0.66	3.31 × 10^{-8}	0.14/0.20	0.69	1.51 × 10^{-4}	0.28/0.33	0.77	3.32 × 10^{-6}	0.73	0.69	0.59	8.44 × 10^{-16}	0.44
6	32,789	**rs9275596 (C)**	0.14/0.22	0.56	1.91 × 10^{-12}	0.09/0.16	0.54	6.29 × 10^{-8}	0.20/0.27	0.70	7.40 × 10^{-10}	0.63	0.62	0.43	1.59 × 10^{-26}	0.31
6	32,917	**rs9357155 (A)**	0.15/0.20	0.69	5.19 × 10^{-6}	0.12/0.18	0.64	1.79 × 10^{-5}	0.11/0.13	0.77	8.26 × 10^{-4}	0.71	0.66	0.62	2.11 × 10^{-12}	0.35
6	32,919	**rs2071543 (A)**	0.16/0.22	0.70	7.19 × 10^{-6}	0.14/0.20	0.65	1.59 × 10^{-5}	0.12/0.14	0.81	1.66 × 10^{-3}	0.73	0.67	0.64	5.77 × 10^{-12}	0.27
6	33,194	**rs1883414 (T)**	0.19/0.24	0.73	3.26 × 10^{-5}	0.17/0.20	0.82	3.55 × 10^{-2}	0.29/0.33	0.82	2.17 × 10^{-4}	0.78	0.77	0.61	4.84 × 10^{-9}	0.55
6	33,205	**rs3129269 (T)**	0.21/0.27	0.73	1.32 × 10^{-5}	0.20/0.23	0.83	3.48 × 10^{-2}	0.33/0.38	0.83	6.67 × 10^{-4}	0.79	0.79	0.61	8.54 × 10^{-9}	0.42
22	28,824	**rs2412971 (A)**	0.31/0.39	0.72	8.21 × 10^{-7}	0.24/0.28	0.83	2.79 × 10^{-2}	0.46/0.51	0.82	1.61 × 10^{-3}	0.80	0.75	0.66	1.86 × 10^{-9}	0.29
22	28,859	**rs2412973 (A)**	0.32/0.39	0.73	1.91 × 10^{-6}	0.26/0.30	0.83	2.68 × 10^{-2}	0.46/0.51	0.83	2.09 × 10^{-3}	0.80	0.76	0.66	4.46 × 10^{-9}	0.28

The per-allele, heterozygote (het.) and homozygote (hom.) ORs are indicated for the combined cohort. MAF, minor allele frequency; Het., heterozygous; Hom., homozygous.
[a]Cochran-Armitage trend test. [b]Stratified analysis using Mantel's extension of Cochran-Armitage trend test. Chr, chromosome. *Q*, *P* value for the Cochran's *Q* statistic. *Significant heterogeneity (*P* < 0.05).

Table 3 Stepwise conditional analysis of association among the signals in the HLA region

		Beijing discovery cohort n = 2,096 (1,194 cases, 902 controls)		Shanghai follow-up cohort n = 1,460 (712 cases, 748 controls)		European follow-up cohort n = 2,410 (1,238 cases, 1,172 controls)		All cohorts combined n = 5,966 (3,144 cases, 2,822 controls)	
Test SNP	Conditioning SNP(s)	Unconditioned P	Conditioned P	Unconditioned P	Conditioned P	Unconditioned P	Conditioned P	Unconditioned P	Conditioned P
rs2856717	rs9275596	3.30×10^{-8}	0.280	1.51×10^{-4}	0.271	3.32×10^{-6}	0.354	8.44×10^{-16}	0.114
rs9275596		1.91×10^{-12}	NA	6.29×10^{-8}	NA	7.40×10^{-10}	NA	1.59×10^{-26}	NA
rs9357155		5.19×10^{-6}	2.29×10^{-3}	1.79×10^{-5}	3.12×10^{-4}	8.26×10^{-4}	8.83×10^{-4}	2.11×10^{-12}	6.87×10^{-9}
rs1883414		1.32×10^{-5}	2.16×10^{-4}	0.0348	0.164	6.67×10^{-4}	3.64×10^{-4}	8.54×10^{-9}	9.94×10^{-8}
rs2856717	rs9275596, rs9357155	3.30×10^{-8}	0.236	1.51×10^{-4}	0.225	3.32×10^{-6}	0.303	8.44×10^{-16}	0.0754
rs9275596		1.91×10^{-12}	NA	6.29×10^{-8}	NA	7.40×10^{-10}	NA	1.59×10^{-26}	NA
rs9357155		5.19×10^{-6}	NA	1.79×10^{-5}	NA	8.26×10^{-4}	NA	2.11×10^{-12}	NA
rs1883414		1.32×10^{-5}	7.04×10^{-5}	0.0348	0.059	6.67×10^{-4}	7.18×10^{-4}	8.54×10^{-9}	3.13×10^{-8}
rs2856717	rs9275596, rs9357155, rs1883414	3.30×10^{-8}	0.278	1.51×10^{-4}	0.241	3.32×10^{-6}	0.272	8.44×10^{-16}	0.0760
rs9275596		1.91×10^{-12}	NA	6.29×10^{-8}	NA	7.40×10^{-10}	NA	1.59×10^{-26}	NA
rs9357155		5.19×10^{-6}	NA	1.79×10^{-5}	NA	8.26×10^{-4}	NA	2.11×10^{-12}	NA
rs1883414		1.32×10^{-5}	NA	0.0348	NA	6.67×10^{-4}	NA	8.54×10^{-9}	NA

rs9275596 and rs2856717 represent the major HLA signal near DQB1. rs9357155 and rs1883414 represent the other two independent signals in the HLA region. NA, not applicable.

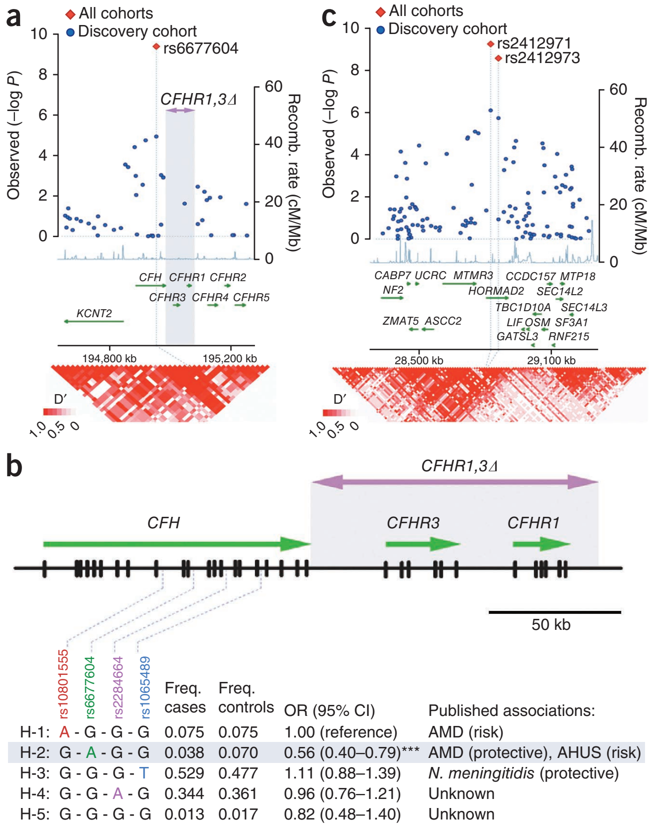
detect the CFHR1 protein in serum from all AA homozygotes tested (**Supplementary Fig. 5**). We evaluated evidence for association of IgA nephropathy with alleles in *CFH* that confer risk of age-related macular degeneration (AMD) and found no contribution to risk (for example, the p.Tyr402His variant, tagged by rs10801555, showed OR = 1.0 and P = 0.99 in discovery cohort; **Fig. 3b**). Haplotype-based analysis in the Beijing discovery cohort indicated protection by the haplotype containing the rs6677604 A allele (OR = 0.56, $P = 1 \times 10^{-6}$ versus all other haplotypes in the discovery cohort; **Fig. 3b** and **Supplementary Fig. 6**) but no significant effect of other haplotypes.

The fifth signal in the GWAS resided in an intronic SNP in *HORMAD2* on chromosome 22.q12.2 (rs2412971, OR = 0.80, $P = 1.9 \times 10^{-9}$) and was supported by a second SNP within 35 kb of this signal (rs2412973, OR = 0.80, $P = 4.5 \times 10^{-9}$). After controlling for rs2412971, there were no other independent signals in this region. The association extends across a large LD segment that encompasses genes including *HORMAD2, MTMR3, LIF* and *OSM* (**Fig. 3c**).

Cumulative effects on disease risk

To determine the cumulative risk conferred by these loci, we computed a genetic risk score, calculated as the weighted sum of the number of protective alleles multiplied by the log of the OR for each of the individual loci (**Table 4** and **Supplementary Tables 13,14**). The disease risk varied up to tenfold between individuals with no protective alleles and those with five or more. The risk score model was similar in all cohorts and collectively explained 5–7% of the variation in disease risk in the Chinese cohorts and ~4% of the risk in the European cohort (**Table 4**). The risk score did not reproducibly correlate with any of the parameters of disease severity, such as estimated glomerular filtration rate, degree of proteinuria or histologic severity grade.

Most notably, consistent with the higher prevalence of IgA nephropathy in Asians, the frequency of protective alleles was significantly lower in the Chinese cohort compared with the European group. The differences in distribution of protective alleles were highly significant between the Asian and European cohorts (**Fig. 4a**, $P = 4.8 \times 10^{-72}$ and $P = 6.4 \times 10^{-60}$ for differences within cases and controls, respectively). To confirm this finding in independent populations, we examined three HapMap groups and similarly found that frequencies of risk alleles correlate with disease frequency among these populations: risk allele frequencies were highest in Asians, intermediate in Europeans and lowest in Africans (**Fig. 4b** and **Supplementary Fig. 7**). For



		Freq. cases	Freq. controls	OR (95% CI)	Published associations:
H-1:	A - G - G - G	0.075	0.075	1.00 (reference)	AMD (risk)
H-2:	G - A - G - G	0.038	0.070	0.56 (0.40–0.79)***	AMD (protective), AHUS (risk)
H-3:	G - G - G - T	0.529	0.477	1.11 (0.88–1.39)	*N. meningitidis* (protective)
H-4:	G - G - A - G	0.344	0.361	0.96 (0.76–1.21)	Unknown
H-5:	G - G - G - G	0.013	0.017	0.82 (0.48–1.40)	Unknown

Figure 3 Analysis of the chromosome 1 and chromosome 22 loci. (**a**) Regional association plot of the chromosome 1q32 locus; although the most strongly associated SNP resides within *CFH*, it is a perfect proxy for *CFHR1,3Δ*. Bottom, LD heat map (*D′*) calculated based on the genotype data of the Beijing cohort. (**b**) Haplotype analysis indicated five common haplotypes (H-1 to H-5) in the Beijing discovery cohort (frequency (freq.) > 0.01). The haplotype frequencies, corresponding tag SNPs and reported disease associations are shown[22–24,36,37,41,43]. The H2 haplotype perfectly tags *CFHR1,3Δ*. The ORs and 95% CIs are calculated in reference to H-1, which has an identical frequency among cases and controls. ***$P = 7.7 \times 10^{-6}$ for comparison of H-2 versus all other haplotypes. (**c**) Regional association plot of the chromosome 22 locus: the strongest association stems from the SNPs residing within *HORMAD2*, but the area of association spans a region over ~0.7 Mb containing multiple genes.

Table 4 Cumulative effect of replicated loci stratified by the number of protective alleles

	Beijing discovery cohort (n = 2,074)* 1,176 cases / 898 controls			Asian replication cohort (n = 1,397)* 685 cases / 712 controls			European replication cohort (n = 2,160)* 1,098 cases / 1,062 controls		
No. protective alleles	Frequency (cases/controls)	Risk score (mean ± s.d.)	OR (95% CI)	Frequency (cases/controls)	Risk score (mean ± s.d.)	OR (95% CI)	Frequency (cases/controls)	Risk score (mean ± s.d.)	OR (95% CI)
0 (highest risk)	0.17/0.07	0.00	1.00 (reference)	0.24/0.13	0.00	1.00 (reference)	0.07/0.03	0.00	1.00 (reference)
1	0.31/0.26	−0.37 ± 0.09	0.50 (0.36–0.69)	0.38/0.32	−0.30 ± 0.15	0.66 (0.48–0.90)	0.19/0.12	−0.11 ± 0.04	0.59 (0.36–0.97)
2	0.29/0.29	−0.77 ± 0.14	0.40 (0.29–0.56)	0.24/0.31	−0.65 ± 0.23	0.43 (0.31–0.60)	0.26/0.24	−0.23 ± 0.05	0.39 (0.25–0.63)
3	0.16/0.20	−1.17 ± 0.15	0.31 (0.22–0.44)	0.10/0.14	−1.06 ± 0.26	0.40 (0.27–0.60)	0.26/0.30	−0.35 ± 0.06	0.30 (0.19–0.48)
4	0.06/0.12	−1.61 ± 0.17	0.20 (0.13–0.31)	0.04/0.08	−1.44 ± 0.28	0.28 (0.16–0.47)	0.15/0.19	−0.47 ± 0.06	0.28 (0.17–0.45)
≥5 (lowest risk)	0.01/0.06	−2.11 ± 0.25	0.09 (0.05–0.16)	0.004/0.03	−1.86 ± 0.36	0.10 (0.03–0.33)	0.08/0.13	−0.65 ± 0.10	0.21 (0.12–0.35)
OR change[a]			11.1			10.0			4.8
***P* value**[b]			6.76×10^{-27}			3.13×10^{-14}			6.24×10^{-17}
C-statistic (95% CI)[c]			0.63 (0.60–0.65)			0.61 (0.58–0.64)			0.60 (0.58–0.62)
Nagelkerke[d] r^2			0.072			0.054			0.042

*Risk scores were calculated on the basis of the ORs and allele frequencies for each specific cohort. Only individuals with nonmissing genotypes for all ten alleles were included in this analysis. [a]Fold-change in OR between highest- and lowest-risk group. [b]*P* value for the risk score prediction model. [c]C-statistic indicates the area under the receiver operating characteristic (ROC) curve for the risk score prediction model. [d]Nagelkerke's pseudo r^2 indicates fraction of the variance in risk explained by the risk score model.

example, the protective allele at the chromosome 1 locus has a frequency of 0.08 in Asians, 0.24 in Europeans and 0.49 in Africans.

DISCUSSION

In this GWAS, we identified five loci imparting significant and consistent effects on the risk of IgA nephropathy across three independent cohorts. These five loci explained up to a tenfold variation in interindividual risk and cumulatively accounted for 4–7% of the disease variance. The effect sizes at these loci are relatively large and consistent across the European and Chinese cohorts, with four having inverse OR ≥ 1.4, which is comparable to those detected in earlier studies of autoimmune or inflammatory diseases[21,24–30]. The risk allele frequencies also strongly paralleled the prevalence of IgA nephropathy among different populations.

We detected a major signal in the MHC region, which was identified but not localized in a recent GWAS with 533 affected subjects[19]. Our study of the markedly larger cohorts reported here showed that this signal originated from three distinct loci within *HLA*; we also identified two non-*HLA* loci. Evidence supporting the presence of three independent risk loci on chromosome 6p21 includes their position within distinct LD segments, as well as genome-wide significance after conditioning for the other two loci, with consistent effects within each cohort.

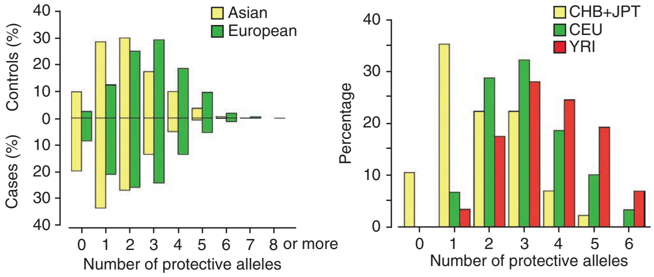


Figure 4 Differences in the distributions of protective alleles by subject ancestry. (**a**) Distributions of protective alleles by subject ancestry and case-control status. Numbers of protective alleles were scored for the combined Asian (n = 3,556) and European (n = 2,410) cohorts. Europeans harbor much greater numbers of protective alleles. The differences in the distribution of protective alleles between Asians and Europeans are highly significant within both case and control groups (χ^2 $P = 4.9 \times 10^{-72}$ and $P = 6.4 \times 10^{-60}$ for cases and controls, respectively). (**b**) Distributions of protective alleles among the three HapMap populations: there were highly significant differences between Asian (CHB+JPT) and European (CEU, $P = 1.3 \times 10^{-3}$) and Asian and Yoruban (YRI, $P = 7.1 \times 10^{-6}$) populations.

The strongest HLA signal was in the region of *HLA-DRB1* and *HLA-DQB1*. Imputation of classical alleles suggested that this signal is fully or partially conveyed by a strong protective effect of the *DRB1*1501-DQB1*0602* haplotype; the strength of this association was probably underestimated by limitations of imputation. This haplotype is relatively common in the European and Asian populations (frequency, ~0.1–0.2) and, in contrast to its protective effect for IgA nephropathy, has been associated with increased risk of systemic lupus erythematosus[25], multiple sclerosis[31], narcolepsy[32] and hepatotoxicity from COX2 inhibitors[30] but is also highly protective for type I diabetes mellitus[26]. This haplotype is also protective in selective IgA deficiency[27], yet we found no association with IgA levels at this locus among cases (**Supplementary Table 15**). This region has a complex LD structure, and our conditional analysis suggests the possibility of an independent signal within this region (at rs9275424; **Supplementary Tables 7**,**8**). High-resolution mapping and direct genotyping of classical alleles will be required to further dissect this interval and identify the functional variant(s).

The second independent interval at 6p21 contained *TAP2, TAP1, PSMB8* and *PSMB9*, interferon-regulated genes implicated in antigen generation and processing for presentation by MHC I molecules; they also have an important role in modulation of cytokine production and cytotoxic T-cell response[33,34]. *PSMB8* expression is increased in peripheral blood mononuclear cells from individuals with IgA nephropathy, motivating further investigation[35]. To our knowledge, this locus has not been identified in any earlier GWAS.

The third signal at 6p21 comprised the *HLA-DPA1*, *HLA-DPB1* and *HLA-DPB2* genes. This locus is associated with risk of chronic hepatitis B infection[29] (a major clinical problem in China) and systemic sclerosis stratified for antibody to DNA topoisomerase I or autoantibodies to centromeres[31], but the risk alleles associated with these phenotypes are not in LD with any of the IgA nephropathy risk alleles.

CFH has a critical role in dampening the alternative complement cascade through inhibition of the C3 and C5 convertases[36]. The functions of the CFH-related proteins are less well understood[36,37]. Loss-of-function mutations in *CFH* produce uncontrolled C3 activation, leading to membranoproliferative glomerulonephritis type II, which is pathologically distinct from IgA nephropathy[36]. Other rare *CFH* mutations can produce hemolytic uremic syndrome, a thrombotic disorder[36], whereas distinct common haplotypes predispose individuals to AMD and susceptibility to meningococcal infection[22–24]. Notably, the *CFH* haplotype bearing the *CFHR1,3Δ* variant may be protective in AMD, but detection of an independent effect has been complicated because of the presence of additional haplotypes imparting both high and low risk[22,23]. Here we found an unambiguous protective effect of the *CFHR1,3Δ*–containing haplotype in IgA nephropathy, strongly suggesting that *CFHR1,3Δ* is the functional variant. Nevertheless, it is unclear how loss of CFHR1 and/or CFHR3 may confer protection for IgA nephropathy. The protective effects may be due to the competing roles of CFH and CFHR1 proteins[37], such that loss of CFHR1 enhances CFH effects, reducing inflammation at tissue surfaces.

The chromosome 22q12.2 locus spans a large interval that contains *OSM* and *LIF*, encoding cytokines implicated in mucosal immunity and inflammation. Notably, inactivation of *Osm* leads to autoimmune glomerulonephritis in mouse[38]. The functions of other genes such as *HORMAD2* and *MTMR3* have not been as well characterized[39]. In addition, the rs2412973 A allele, which is protective for IgA nephropathy, has also been associated with increased risk of early-onset inflammatory bowel disease and altered expression of *MTMR3* expression in individuals with ulcerative colitis[28]. This finding is notable given the known clinical association between inflammatory bowel disease and secondary forms of IgA nephropathy, but the underlying signal within this locus remains to be clarified. Lastly, the protective allele at this locus is also associated with lower serum IgA levels among cases ($P = 3.9 \times 10^{-3}$; **Supplementary Table 15** and **Supplementary Fig. 8**).

Many of the protective alleles for IgA nephropathy have been implicated as risk factors for other immune-mediated and infectious disorders, suggesting that complex selection pressures (potentially balancing selection) may influence the frequencies of these alleles among world populations. Statistical proof of balancing selection on allele frequencies or genotypes may be particularly challenging if alleles have been maintained in the population over very long evolutionary periods. Notably, a recent genome-wide survey detected a signal of selection in the vicinity of the *CFH* gene cluster[40], and there is a large difference in the frequency of the rs6677604 A allele among world populations (**Supplementary Table 16**).

The loci identified in this study clarify the genetic architecture of sporadic IgA nephropathy, identifying new pathogenic pathways and connections to other immune-mediated disorders. On the basis of our power calculations, we identified virtually all loci imparting an OR ≥ 1.5 in the Chinese discovery cohort, but additional loci with large effects may be present among Europeans. Considering the effectiveness of GWAS for studies of immunologic disorders[22,27–31,41] and the increased power imparted by larger sample size[42], genome-wide examination of larger cohorts will probably define additional genetic components of IgA nephropathy.

URLs. HLA genotype data of the HapMap Chinese samples, https://www.sanger.ac.uk/HGP/Chr6/ng2006-data/; IBD software, http://www1.cs.columbia.edu/~itsik/hla_ibd/; HapMap, http://hapmap.ncbi.nlm.nih.gov/.

METHODS

Methods and any associated references are available in the online version of the paper at http://www.nature.com/naturegenetics/.

Note: Supplementary information is available on the Nature Genetics website.

ACKNOWLEDGMENTS
We are grateful to all study participants for their contribution to this work. We also thank the staff of the Yale West Campus Center for Genome Analysis for their excellent support. We also appreciate the assistance of C.V. Barker and S.Y. Woodford with sample collection. This study was supported by RC1DK087445 (A.G.G., R.P.L.), R01DK082753 (A.G.G., J.N., B.A.J. and R.J.W.) and KL2 RR24157 (K.K.), the Center for Glomerular Diseases at Columbia University, the Yale Center Translational Science Award and the Yale Center for Human Genetics and Genomics. R.P.L. is an investigator of the Howard Hughes Medical Institute.

AUTHOR CONTRIBUTIONS
Subject clinical characterization, recruitment and contribution of samples: P.H., J.X., S.S.C., B.A.J., R.J.W., J.N., J.C.H., H.W., J.L., L.Z., W.W., Z.W., S.S., R. Magistroni, G.M.G., M.B., P.R., C.P., L.A., G.B., G.F., A. Amore, L.P., R.C., C.I., B.F.V., E.P., M.S., R. Mignani, L.G., F.B., P.M., A. Amoroso, F.S., N.C. and H.Z.

DNA preparation: Y.L., P.H., J.X., F.L., I.B., K.K., C.J.M. and M.C.

Genotyping and wet lab experiments: S.M., S.U., I.T., C.J.M., M.C., P.H., J.X. and Y.L.

Data management: K.K., Y.L., S.S.C. and M.C.

Data analysis: K.K., M.C., A.G.G. and R.P.L.

Analytical support and discussion: K.Y. and M.G.

Manuscript preparation: A.G.G., K.K., M.C. and R.P.L.

Conception and overall supervision of project: A.G.G. and R.P.L.

COMPETING FINANCIAL INTERESTS
The authors declare no competing financial interests.

Published online at http://www.nature.com/naturegenetics/.
Reprints and permissions information is available online at http://npg.nature.com/reprintsandpermissions/.

1. Coresh, J. *et al.* Prevalence of chronic kidney disease in the United States. *J. Am. Med. Assoc.* **298**, 2038–2047 (2007).
2. Tsukamoto, Y. *et al.* Report of the Asian Forum of Chronic Kidney Disease Initiative (AFCKDI) 2007. "Current status and perspective of CKD in Asia": diversity and specificity among Asian countries. *Clin. Exp. Nephrol.* **13**, 249–256 (2009).
3. Gesualdo, L., Di Palma, A.M., Morrone, L.F., Strippoli, G.F. & Schena, F.P. The Italian experience of the national registry of renal biopsies. *Kidney Int.* **66**, 890–894 (2004).
4. D'Amico, G. The commonest glomerulonephritis in the world: IgA nephropathy. *Q. J. Med.* **64**, 709–727 (1987).
5. Nair, R. & Walker, P.D. Is IgA nephropathy the commonest primary glomerulopathy among young adults in the USA? *Kidney Int.* **69**, 1455–1458 (2006).
6. Varis, J. *et al.* Immunoglobulin and complement deposition in glomeruli of 756 subjects who had committed suicide or met with a violent death. *J. Clin. Pathol.* **46**, 607–610 (1993).
7. Suzuki, K. *et al.* Incidence of latent mesangial IgA deposition in renal allograft donors in Japan. *Kidney Int.* **63**, 2286–2294 (2003).
8. Kiryluk, K. *et al.* Genetic studies of IgA nephropathy: past, present, and future. *Pediatr. Nephrol.* **25**, 2257–2268 (2010).
9. Barratt, J. & Feehally, J. IgA nephropathy. *J. Am. Soc. Nephrol.* **16**, 2088–2097 (2005).
10. Hastings, M.C. *et al.* Galactose-deficient IgA1 in African Americans with IgA nephropathy: serum levels and heritability. *Clin. J. Am. Soc. Nephrol.* **5**, 2069–2074 (2010).
11. Gharavi, A.G. *et al.* Aberrant IgA1 glycosylation is inherited in familial and sporadic IgA nephropathy. *J. Am. Soc. Nephrol.* **19**, 1008–1014 (2008).
12. Lin, X. *et al.* Aberrant galactosylation of IgA1 is involved in the genetic susceptibility of Chinese patients with IgA nephropathy. *Nephrol. Dial. Transplant.* **24**, 3372–3375 (2009).
13. Moldoveanu, Z. *et al.* Patients with IgA nephropathy have increased serum galactose-deficient IgA1 levels. *Kidney Int.* **71**, 1148–1154 (2007).
14. Mestecky, J. *et al.* Defective galactosylation and clearance of IgA1 molecules as a possible etiopathogenic factor in IgA nephropathy. *Contrib. Nephrol.* **104**, 172–182 (1993).
15. Tomana, M. *et al.* Circulating immune complexes in IgA nephropathy consist of IgA1 with galactose-deficient hinge region and antiglycan antibodies. *J. Clin. Invest.* **104**, 73–81 (1999).
16. Gharavi, A.G. *et al.* IgA nephropathy, the most common cause of glomerulonephritis, is linked to 6q22–23. *Nat. Genet.* **26**, 354–357 (2000).
17. Bisceglia, L. *et al.* Genetic heterogeneity in Italian families with IgA nephropathy: suggestive linkage for two novel IgA nephropathy loci. *Am. J. Hum. Genet.* **79**, 1130–1134 (2006).
18. Paterson, A.D. *et al.* Genome-wide linkage scan of a large family with IgA nephropathy localizes a novel susceptibility locus to chromosome 2q36. *J. Am. Soc. Nephrol.* **18**, 2408–2415 (2007).
19. Feehally, J. *et al.* HLA has strongest association with IgA nephropathy in genome-wide analysis. *J. Am. Soc. Nephrol.* **21**, 1791–1797 (2010).
20. Storey, J.D. & Tibshirani, R. Statistical significance for genomewide studies. *Proc. Natl. Acad. Sci. USA* **100**, 9440–9445 (2003).
21. de Bakker, P.I. *et al.* A high-resolution HLA and SNP haplotype map for disease association studies in the extended human MHC. *Nat. Genet.* **38**, 1166–1172 (2006).

22. Hughes, A.E. *et al.* A common CFH haplotype, with deletion of CFHR1 and CFHR3, is associated with lower risk of age-related macular degeneration. *Nat. Genet.* **38**, 1173–1177 (2006).
23. Raychaudhuri, S. *et al.* Associations of CFHR1-CFHR3 deletion and a CFH SNP to age-related macular degeneration are not independent. *Nat. Genet.* **42**, 553–555, author reply 555–556 (2010).
24. Davila, S. *et al.* Genome-wide association study identifies variants in the CFH region associated with host susceptibility to meningococcal disease. *Nat. Genet.* **42**, 772–776 (2010).
25. Barcellos, L.F. *et al.* High-density SNP screening of the major histocompatibility complex in systemic lupus erythematosus demonstrates strong evidence for independent susceptibility regions. *PLoS Genet.* **5**, e1000696 (2009).
26. Erlich, H. *et al.* HLA DR-DQ haplotypes and genotypes and type 1 diabetes risk: analysis of the type 1 diabetes genetics consortium families. *Diabetes* **57**, 1084–1092 (2008).
27. Ferreira, R.C. *et al.* Association of IFIH1 and other autoimmunity risk alleles with selective IgA deficiency. *Nat. Genet.* **42**, 777–780 (2010).
28. Imielinski, M. *et al.* Common variants at five new loci associated with early-onset inflammatory bowel disease. *Nat. Genet.* **41**, 1335–1340 (2009).
29. Kamatani, Y. *et al.* A genome-wide association study identifies variants in the HLA-DP locus associated with chronic hepatitis B in Asians. *Nat. Genet.* **41**, 591–595 (2009).
30. Singer, J.B. *et al.* A genome-wide study identifies HLA alleles associated with lumiracoxib-related liver injury. *Nat. Genet.* **42**, 711–714 (2010).
31. Zhou, X. *et al.* HLA-DPB1 and DPB2 are genetic loci for systemic sclerosis: a genome-wide association study in Koreans with replication in North Americans. *Arthritis Rheum.* **60**, 3807–3814 (2009).
32. Mignot, E. *et al.* Complex HLA-DR and -DQ interactions confer risk of narcolepsy-cataplexy in three ethnic groups. *Am. J. Hum. Genet.* **68**, 686–699 (2001).
33. Begley, G.S., Horvath, A.R., Taylor, J.C. & Higgins, C.F. Cytoplasmic domains of the transporter associated with antigen processing and P-glycoprotein interact with subunits of the proteasome. *Mol. Immunol.* **42**, 137–141 (2005).
34. Muchamuel, T. *et al.* A selective inhibitor of the immunoproteasome subunit LMP7 blocks cytokine production and attenuates progression of experimental arthritis. *Nat. Med.* **15**, 781–787 (2009).
35. Coppo, R. *et al.* Upregulation of the immunoproteasome in peripheral blood mononuclear cells of patients with IgA nephropathy. *Kidney Int.* **75**, 536–541 (2009).
36. Atkinson, J.P. & Goodship, T.H. Complement factor H and the hemolytic uremic syndrome. *J. Exp. Med.* **204**, 1245–1248 (2007).
37. Heinen, S. *et al.* Factor H-related protein 1 (CFHR-1) inhibits complement C5 convertase activity and terminal complex formation. *Blood* **114**, 2439–2447 (2009).
38. Esashi, E. *et al.* Oncostatin M deficiency leads to thymic hypoplasia, accumulation of apoptotic thymocytes and glomerulonephritis. *Eur. J. Immunol.* **39**, 1664–1670 (2009).
39. Wojtasz, L. *et al.* Mouse HORMAD1 and HORMAD2, two conserved meiotic chromosomal proteins, are depleted from synapsed chromosome axes with the help of TRIP13 AAA-ATPase. *PLoS Genet.* **5**, e1000702 (2009).
40. Grossman, S.R. *et al.* A composite of multiple signals distinguishes causal variants in regions of positive selection. *Science* **327**, 883–886 (2010).
41. Maller, J. *et al.* Common variation in three genes, including a noncoding variant in CFH, strongly influences risk of age-related macular degeneration. *Nat. Genet.* **38**, 1055–1059 (2006).
42. Wellcome Trust Case Control Consortium. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature* **447**, 661–678 (2007).
43. Zipfel, P.F. *et al.* Deletion of complement factor H-related genes CFHR1 and CFHR3 is associated with atypical hemolytic uremic syndrome. *PLoS Genet.* **3**, e41 (2007).

ONLINE METHODS

Genome-wide genotyping, genotype quality control and primary association analyses. Study populations, genome-wide genotyping and genotype quality control are described in the **Supplementary Note** and **Supplementary Tables 1** and **2**. After quality control analysis, the discovery cohort consisted of 1,194 cases and 902 controls genotyped with the Illumina Human 610-Quad BeadChip. The primary genome-wide association analyses were carried out using PLINK v1.07 (ref. 44). We selected a standard 1-degree-of-freedom Cochran-Armitage trend test as the primary association test, as it demonstrates greater robustness to deviations from Hardy-Weinberg equilibrium compared with its alternatives. We estimated the per-allele ORs and 95% confidence intervals for all tested SNPs. The genome-wide distributions of *P* values were examined using quantile-quantile plots, before and after exclusion of the HLA region (**Supplementary Fig. 1**). The assessment of population stratification in the discovery cohort is described in **Supplementary Note**, **Supplementary Table 3** and **Supplementary Figure 2**.

False discovery rate and power analysis. The calculation of positive false discovery rate (pFDR) was carried out using the *Q*-value package (R). The proportion of SNPs that were truly null (Πo) was estimated at 0.991 using the empirical distribution of genome-wide *P* values[20]. The *Q* value of 0.10 (pFDR of 10%) corresponded to the *P* value of 1.3×10^{-5} (**Supplementary Fig. 3**). This *Q*-value threshold defined 65 top SNPs that were subsequently analyzed for replication. The power analysis was carried out using methods described[45]. The calculations were carried out under the following assumptions: a disease prevalence of 1%; an additive risk model for stage I (discovery) with sample size of 1,000 cases and 1,000 controls and stage II (follow-up) with sample size of 2,000 cases and 2,000 controls; a follow-up significance threshold of 1.3×10^{-5}; and joint (stage I and II) significance level of 5×10^{-8}. The joint power of our study design (**Supplementary Table 4**) was calculated for a range of disease allele frequencies (0.10–0.50) and effect sizes (genotypic risk ratio 1.10–1.80). The effect sizes detectable at $\alpha = 5 \times 10^{-8}$ and a power of 0.80 in the joint analysis were estimated using CaTS software[45].

Selection of SNPs for follow-up. The 65 SNPs that reached our *Q*-value threshold were first clustered into ten distinct loci on the basis of their physical location and regional patterns of LD. The correctness of genotype calls was verified for each SNP individually by visual inspection of the Illumina cluster plots. Conditional logistic regression analysis was carried out to confirm correct SNP grouping and detect independence signals. These analyses suggested three distinct loci on chromosome 6p21 and two distinct loci on chromosome 22q12.2. The SNPs with the lowest *P* value within each locus were selected for follow up. The selection of the second SNP for back-up genotyping was based mainly on its strength of association, high LD with the top-scoring SNP in European and Chinese HapMap populations, robustness of Illumina clustering plots and high genotyping rate. In total, we selected 20 representative SNPs for genotyping in 2,013 cases and 1,951 controls recruited for stage 2 of the study. Genotyping and genotype quality control in the follow-up cohorts is described in **Supplementary Note**.

Association analyses across multiple cohorts. Results across multiple cohorts were combined using a stratified trend test with Mentel's extension of the Cochran-Armitage test (snpMatrix package, R)[46]. We tested for heterogeneity across cohorts with the heterogeneity index (I^2), and by carrying out Cochran's *Q* heterogeneity test. To ensure findings were robust to methodology, we also combined the per-allele effect estimates using Cochran-Mantel-Haenszel stratified analysis, as well as an inverse variance-weighted method under a fixed-effects model. The results were concordant regardless of the meta-analytic method used.

Conditional analyses. We carried out stepwise logistic regression after controlling for the genotypes of the conditioning SNPs using PLINK (v1.07). The adjusted (conditioned) effect estimates were then combined across cohorts by adding cohort information as an additional covariate in the stratified analysis (**Table 3** and **Supplementary Table 8**). A similar approach was used for the conditional analysis of classical HLA alleles (**Supplementary Table 10**).

Haplotype-based association at *CFH* locus. These analyses were carried out in PLINK v1.07. Haplotypes were phased across the *CFH* locus in the Beijing cohort (**Fig. 3b** and **Supplementary Fig. 6**) and haplotype frequencies were estimated in the cases and controls separately, as well as jointly in the entire cohort. Only the haplotypes with overall frequency >1% were included in association analyses. The *P* values were derived for tests of association of one haplotype versus all others. The ORs and the corresponding 95% confidence intervals were estimated in reference to the AMD risk haplotype (H-1; **Fig. 3b**), which has an identical frequency between cases and controls.

Imputation and association analysis of classical *HLA* alleles. The *HLA* classical alleles at *DQB1*, *DQA1* and *DRB1* loci were imputed on the basis of the genotype data from the Beijing cohort (**Supplementary Tables 9,10**). In short, the genotype data were first phased using BEAGLE[47] and pairwise inflammatory bowel disease status was determined using GERMLINE software[48]. The HLA classical allele status and genotype data of the HapMap Han Chinese individuals were used as a reference panel (see URLs)[21]. The imputation was carried out using the HLA-via-IBD software (see URLs). The accuracy of the imputation procedure was tested by direct sequencing of the informative coding segments of *HLA-DQB1* gene in a random subset of 420 samples. This demonstrated that imputation had 57% sensitivity and 96% specificity for identifying the *HLA-DQB1*602* alleles (**Supplementary Table 11**).

Risk score discovery and validation. Among the five independent regions of association, alleles with lower frequency conveyed a protective effect. Therefore, the risk score model was based on protective genotypes for the top five independent and most strongly associated SNPs (rs9275596, rs9357155, rs1883414, rs2412971 and rs6677604). The risk score was calculated as a weighted sum of the number of protective alleles at each locus multiplied by the log of the OR for each of the individual loci for a specific cohort. Only individuals with nonmissing genotypes for all ten alleles were included in this analysis (**Table 4** and **Supplementary Table 13**). The predictive risk score models were built using association results for each of the three model-building cohorts and were validated by testing their predictive properties against all other cohorts (target cohorts, **Supplementary Table 14** and **Supplementary Fig. 7**). The percentage of the total variance in disease state explained by the risk score was estimated by Nagelkerke's pseudo r^2 from the logistic regression model with the risk score as a quantitative predictor and disease state as an outcome. The C-statistic was estimated as an area under the receiver operating characteristic curve provided by the above logistic model. These analyses were carried out with SPSS Statistics version 17.0.

Distributions of protective alleles. Each individual study participant was scored for the number of protective alleles and the distributions of protective alleles were compared between groups of various ancestries (**Fig. 4**). Only individuals with complete genotype information were included. Because relatively few individuals had five or more protective alleles, they were binned into a single category for the purpose of statistical testing and a χ^2 goodness-of-fit test was used to derive *P* values. Analysis of the HapMap release 23 data set included 30 unrelated individuals from Yoruba in Ibadan, Nigeria (YRI), 30 unrelated Utah residents with ancestry from northern and western Europe (CEU), and a combined group of 45 unrelated Japanese individuals from Tokyo (JPT) and 45 Han Chinese from Beijing (CHB). The genotype data were downloaded directly from the HapMap Project website (see URLs). Our exploratory association analyses of protective alleles with clinical subphenotypes are described in **Supplementary Note** and **Supplementary Table 15**.

Common copy number polymorphisms analysis. For the purpose of this analysis, we used publicly available CNP discovery data obtained with 2.1 million NimbleGen CGH arrays[49,50]. We identified 1,051 SNPs present on the Illumina HumanHap 610K chip that tag known common (>1%) copy number variations at $r^2 > 0.8$. The genotypes for these SNPs were extracted from the data set and analyzed separately for association with the disease state

doi:10.1038/ng.787

(**Supplementary Fig. 4** and **Supplementary Table 12**). These SNPs underwent all quality control steps as outlined above before association analysis. A simple 1-degree-of-freedom χ^2 allelic test was used to screen for association (PLINK) and the results were ranked and visualized using a quantile-quantile plot (R). The top associated CNPs were validated using quantitative real-time PCR.

Quantitative real-time PCR. Quantitative real-time PCR was carried out on genomic DNA using the iQ5 Real-Time PCR Detection System (Bio-Rad) and amplification was achieved using SYBR Green Supermix (Bio-Rad) with a standard two-step amplification protocol. All samples were analyzed in triplicate. Three amplicons spanning *CFHR1* and *CFHR3* were tested and the signal was normalized to an amplicon in B-actin (**Supplementary Table 17**). Pooled DNA from ten individuals homozygous for G alleles at rs6677604 was used as reference.

Protein blotting. Diluted plasma samples were separated on 4–15% Ready Gel (Bio-Rad), transferred to PVDF membranes (Millipore), and protein blotted with primary antibodies to CFH (AbD Serotec) and CFHR1 (R&D Systems) using standard protocols.

44. 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).
45. 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).
46. Clayton, D. & Leung, H.T. An R package for analysis of whole-genome association studies. *Hum. Hered.* **64**, 45–51 (2007).
47. Browning, S.R. & Browning, B.L. Rapid and accurate haplotype phasing and missing-data inference for whole-genome association studies by use of localized haplotype clustering. *Am. J. Hum. Genet.* **81**, 1084–1097 (2007).
48. Gusev, A. *et al.* Whole population, genome-wide mapping of hidden relatedness. *Genome Res.* **19**, 318–326 (2009).
49. Conrad, D.F. *et al.* Origins and functional impact of copy number variation in the human genome. *Nature* **464**, 704–712 (2010).
50. Craddock, N. *et al.* Genome-wide association study of CNVs in 16,000 cases of eight common diseases and 3,000 shared controls. *Nature* **464**, 713–720 (2010).

doi:10.1038/ng.787