ARTICLE

Fine Mapping Seronegative and Seropositive Rheumatoid Arthritis to Shared and Distinct HLA Alleles by Adjusting for the Effects of Heterogeneity

Buhm Han,[1,2,3] Dorothée Diogo,[1,2,3,4] Steve Eyre,[5,6] Henrik Kallberg,[7] Alexandra Zhernakova,[8,9] John Bowes,[5,6] Leonid Padyukov,[7] Yukinori Okada,[1,2,3,4] Miguel A. González-Gay,[10] Solbritt Rantapää-Dahlqvist,[11] Javier Martin,[12] Tom W.J. Huizinga,[8] Robert M. Plenge,[13] Jane Worthington,[5,6] Peter K. Gregersen,[14] Lars Klareskog,[7] Paul I.W. de Bakker,[1,2,15] and Soumya Raychaudhuri[1,2,3,4,5,*]

Despite progress in defining human leukocyte antigen (HLA) alleles for anti-citrullinated-protein-autoantibody-positive ($ACPA^+$) rheumatoid arthritis (RA), identifying HLA alleles for ACPA-negative ($ACPA^-$) RA has been challenging because of clinical heterogeneity within clinical cohorts. We imputed 8,961 classical HLA alleles, amino acids, and SNPs from Immunochip data in a discovery set of 2,406 $ACPA^-$ RA case and 13,930 control individuals. We developed a statistical approach to identify and adjust for clinical heterogeneity within $ACPA^-$ RA and observed independent associations for serine and leucine at position 11 in HLA-DRβ1 ($p = 1.4 \times 10^{-13}$, odds ratio [OR] = 1.30) and for aspartate at position 9 in HLA-B ($p = 2.7 \times 10^{-12}$, OR = 1.39) within the peptide binding grooves. These amino acid positions induced associations at *HLA-DRB1*03* (encoding serine at 11) and *HLA-B*08* (encoding aspartate at 9). We validated these findings in an independent set of 427 $ACPA^-$ case subjects, carefully phenotyped with a highly sensitive ACPA assay, and 1,691 control subjects (HLA-DRβ1 Ser11+Leu11: $p = 5.8 \times 10^{-4}$, OR = 1.28; HLA-B Asp9: $p = 2.6 \times 10^{-3}$, OR = 1.34). Although both amino acid sites drove risk of $ACPA^+$ and $ACPA^-$ disease, the effects of individual residues at HLA-DRβ1 position 11 were distinct ($p < 2.9 \times 10^{-107}$). We also identified an association with $ACPA^+$ RA at HLA-A position 77 ($p = 2.7 \times 10^{-8}$, OR = 0.85) in 7,279 $ACPA^+$ RA case and 15,870 control subjects. These results contribute to mounting evidence that $ACPA^+$ and $ACPA^-$ RA are genetically distinct and potentially have separate autoantigens contributing to pathogenesis. We expect that our approach might have broad applications in analyzing clinical conditions with heterogeneity at both major histocompatibility complex (MHC) and non-MHC regions.

Introduction

Rheumatoid arthritis (RA [MIM 180300]) has two distinct subtypes—anti-citrullinated-protein-autoantibody-negative ($ACPA^-$ or seronegative) RA and -positive ($ACPA^+$ or seropositive) RA—with potentially different genetic risk factors, environmental risk factors, and optimal therapeutic strategies.[1,2] Despite constituting about one-third (~30%) of RA cases,[3] $ACPA^-$ RA has been relatively understudied in comparison to $ACPA^+$ RA.[4–7] We and others have demonstrated that the widely established method for identifying $ACPA^-$ RA subjects on the basis of anticyclic citrullinated peptide (anti-CCP) antibody testing is imperfect in that the absence of antibody is not sufficiently specific to $ACPA^-$ RA, whereas its presence is specific to $ACPA^+$ RA.[8–10]

The lack of a specific test for $ACPA^-$ RA can result in heterogeneity in clinical cohorts, which can confound genetic studies for $ACPA^-$ disease. For example, $ACPA^-$ RA subjects might include $ACPA^+$ RA subjects whose ACPAs have not been detected by conventional anti-CCP testing[8–11] or subjects who have other autoantibody-negative inflammatory arthritic conditions, such as ankylosing spondylitis (AS)[12] or other *HLA-B*27*-associated conditions. So, although investigators have reported associations between classical HLA alleles and $ACPA^-$ RA,[13,14] it remains unclear whether these associations are distinct from those alleles driving $ACPA^+$ disease risk, recently defined by our group.[6] Additionally, the specific amino acid sites and residues driving $ACPA^-$ RA risk have yet to be defined.

To define HLA alleles driving $ACPA^-$ RA risk, we first obtained dense SNP genotype data within the major

[1]Division of Genetics, Brigham & Women's Hospital and Harvard Medical School, Boston, MA 02115, USA; [2]Program in Medical and Population Genetics, Broad Institute of Harvard and MIT, Cambridge, MA 02142, USA; [3]Partners HealthCare Center for Personalized Genetic Medicine, Boston, MA 02115, USA; [4]Division of Rheumatology, Immunology, and Allergy, Brigham and Women's Hospital and Harvard Medical School, Boston, MA, 02115, USA; [5]Arthritis Research UK Epidemiology Unit, Musculoskeletal Research Group, University of Manchester, Manchester Academic Health Sciences Centre, Manchester M13 9PT, UK; [6]NIHR Manchester Musculoskeletal Biomedical Research Unit, Central Manchester University Hospitals NHS Foundation Trust, Manchester Academic Health Sciences Centre, Manchester M13 9PT, UK; [7]Rheumatology Unit, Department of Medicine, Karolinska Institutet and Karolinska University Hospital Solna, 171 76 Stockholm, Sweden; [8]Department of Rheumatology, Leiden University Medical Centre, 2300 RC Leiden, the Netherlands; [9]Department of Genetics, University Medical Center Groningen and University of Groningen, 9700 RB Groningen, the Netherlands; [10]Rheumatology Division, Hospital Universitario Marqués de Valdecilla, Instituto de Formación e Investigación Marqués de Valdecilla, 39008 Santander, Spain; [11]Department of Public Health and Clinical Medicine and Department of Rheumatology, Umeå University, 901 85 Umeå, Sweden; [12]Instituto de Parasitologia y Biomedicina Lopez-Neyra, Consejo Superior de Investigaciones Cientificas, 18100 Armilla, Granada, Spain; [13]Merck Research Laboratories, Merck & Co. Inc., Boston, MA 02115, USA; [14]The Feinstein Institute for Medical Research, North Shore–Long Island Jewish Health System, Manhasset, NY 11030, USA; [15]Departments of Epidemiology and Medical Genetics, University Medical Center Utrecht, 3584 CG Utrecht, the Netherlands
*Correspondence: soumya@broadinstitute.org
http://dx.doi.org/10.1016/j.ajhg.2014.02.013.

histocompatibility complex (MHC) region by applying the Immunochip custom array[3] to ACPA$^-$ case and control groups. We then used these data to impute HLA alleles, amino acids, and SNPs with a highly accurate imputation approach.[15] Recognizing that possible clinical heterogeneity within genotyped cohorts might confound associations within the MHC, we developed a statistical approach to correct for the effects of heterogeneity within cohorts; it uses genetic risk scores (GRSs) built from known risk loci for potential confounding diseases as covariates.

We observed that two amino acid positions, HLA-DRβ1 position 11 (in which serine and leucine conferred risk) and HLA-B position 9 (in which aspartate conferred risk), were driving ACPA$^-$ RA. These two positions are already known to drive ACPA$^+$ RA as well;[6] however, the specific amino acid residues conferring risk were completely distinct between the two disease subtypes. We also separately tested for associations with ACPA$^+$ disease. In addition to confirming known associations at positions 11, 71, and 74 in HLA-DRβ1, position 9 in HLA-B, and position 9 in HLA-DPβ1, we identified an additional association at amino acid position 77 within the binding groove of HLA-A. These results contribute to mounting evidence that ACPA$^+$ and ACPA$^-$ RA are distinct diseases with certain unique genetic factors.

Material and Methods

Samples

Case-Control Sample Collections

We used data from six case-control collections (UK, US, Dutch, Spanish, Swedish Umeå, and Swedish Epidemiological Investigation of Rheumatoid Arthritis [EIRA], Table S1, available online).[3] All individuals provided informed consent and were recruited through protocols approved by institutional review boards. Each collection consisted of individuals who were self-described as white and of European descent, and all cases either met the 1987 American College of Rheumatology diagnostic criteria or were diagnosed by board-certified rheumatologists. We previously genotyped all samples with the Immunochip custom array, which densely covered the MHC region (7,563 SNPs), in accordance with Illumina protocols.

Classifying ACPA$^-$ RA in Discovery Samples

From these samples, we defined a total of 2,406 ACPA$^-$ RA case and 13,930 control subjects for discovery from five collections (excluding the Swedish EIRA). To do this, we followed standard clinical practice to identify ACPA$^-$ RA subjects as those who were not reactive to anti-CCP antibody by using reference cutoff levels defined at local clinical labs. In the UK cohort, we used the commercially available DiastatTM ACPA Kit (Axis-Shield Diagnostics Limited). In the US samples, we used a second-generation commercial anti-CCP enzyme immunoassay (Inova Diagnostics).[16] For Spanish samples, we used the Immunoscan ELISA test (Euro Diagnostica). For the Swedish Umeå and Dutch collections, we used the Immunoscan-RA Mark2 ELISA test (Euro Diagnostica).[17] These assays are the standard commercially available assays that are currently being widely used in clinical practice.

Clinically Homogeneous ACPA$^-$ Samples for Replication

To replicate ACPA$^-$ results, we sought to define an independent replication data set that was as clinically homogeneous as possible. To this end, we used genotype data on 987 case and 1,940 control subjects who were from the Swedish EIRA cohort and who were identified as anti-CCP antibody negative with the Immunoscan-RA Mark2 ELISA test (Euro-Diagnostica). In addition, to stringently ensure clinical homogeneity, we applied a highly sensitive ACPA typing method developed at the Karolinska Institutet[8] to test sera for reactivity to four specific citrullinated peptides (α-enolase, vimentin, fibrinogen, collagen type II). We considered samples ACPA$^-$ only if they were negative for all four of these tests. After applying this assay, we removed 106 case individuals who were reactive to the sensitive assay, as well as 381 case individuals to whom we did not apply the assay. We also excluded 73 case and 249 control subjects who were positive for *HLA-B*27*. Because *HLA-B*27* is highly sensitive for AS (>90%), excluding *HLA-B*27*-positive individuals effectively removed the effect of possible confounding from AS or related spondyloarthropathies. The resulting replication collection consisted of 427 case and 1,691 control subjects.

Sample Collections for ACPA$^+$ RA

For ACPA$^+$ RA, we used 7,279 anti-CCP-positive individuals from all six cohorts (UK, US, Swedish Umeå, Dutch, Spanish, and Swedish EIRA; Table S1). We used all 15,870 control subjects for ACPA$^+$ RA analyses.

Statistical Analyses

HLA Imputation

We imputed case and control groups together for 8,961 binary markers representing classical HLA alleles, amino acids, and SNPs by using SNP2HLA,[15] which utilizes the Beagle imputation method.[18] The binary markers included every possible grouping of amino acid residues given a multiallelic amino acid position. We used reference data collected by the Type 1 Diabetes Genetics Consortium;[19] these data consisted of genotypes for 5,863 SNPs tagging the MHC and classical alleles for *HLA-A*, *HLA-B*, *HLA-C*, *HLA-DRB1*, *HLA-DQA1*, *HLA-DQB1*, *HLA-DPA1*, and *HLA-DPB1* at four-digit resolution in 5,225 individuals of European descent.[19]

Quantifying Imputation Accuracy

To assess accuracy, we took advantage of typed *HLA-A*, *HLA-B*, *HLA-C*, *HLA-DQB1*, and *HLA-DRB1* alleles for 918 individuals in the UK cohort. We calculated imputation accuracy as the proportion of correctly imputed classical alleles:

$$\frac{\sum_i \max\left(\delta(g_{i,1}=x_{i,1})+\delta(g_{i,2}=x_{i,2}),\delta(g_{i,1}=x_{i,2})+\delta(g_{i,2}=x_{i,1})\right)}{2n},$$

where $g_{i,1}$ and $g_{i,2}$ are genotyped alleles of individual i and $x_{i,1}$ and $x_{i,2}$ are imputed alleles. For each gene, we used individuals successfully typed for four-digit alleles. The δ function is 1 if the genotyped allele is the imputed allele and 0 otherwise. The term n is the number of samples.

Statistical Framework for Association Testing

We tested associations at all 8,961 binary markers by using probabilistic genotypic dosages that take uncertainty in imputation into account. We used logistic regression under the assumption that each marker conferred a fixed log additive effect across each case-control collection. To account for population stratification, we included ten principal components (PCs) as covariates for each collection. We calculated PCs by using EIGENSOFT v.4.2[20] with HapMap Phase 2 samples as reference populations on a

subset of SNPs (minor allele frequency > 0.05) filtered for minimizing intermarker linkage disequilibrium (LD).[3] This resulted in the following logistic regression model:

$$\log(\text{odds}_i) = \theta + \beta_a g_{a,i} + \sum_{j \in \text{collections}} \delta_{i,j}\left(\gamma_j + \sum_{k=1\ldots10} \pi_{j,k}\, p_{i,k}\right), \quad \text{(Equation 1)}$$

where a indicates the marker being tested, $g_{a,i}$ is the dosage of a in individual i, and β_a is the additive effect of a. In the collection-specific term, $\delta_{i,j}$ is an indicator variable that is 1 only if individual i is in collection j. The γ_j parameter is the collection-specific effect due to the differences in case-control proportions; it is set to 0 for one arbitrarily selected reference collection. The $\pi_{j,k}$ parameter is the effect of the k^{th} PC, and $p_{i,k}$ is the k^{th} PC value for individual i.

Adjusting for Clinical Heterogeneity in ACPA$^-$ Discovery

In the discovery analysis for ACPA$^-$ disease, we adjusted for possible clinical heterogeneity within the collections. Our approach was to extend Equation 1 to include GRSs of potentially confounding diseases as covariates:

$$\log(\text{odds}_i) = \theta + \beta_a g_{a,i} + \sum_{j \in \text{Collections}} \delta_{i,j}\left(\gamma_i + \sum_{k=1\ldots10} \pi_{j,k}\, p_{i,k} + \sum_{h=1\ldots H} \alpha_{j,h}\, s_{i,h}\right), \quad \text{(Equation 2)}$$

where h indicates a confounding disease we want to adjust for and H is the total number of confounding diseases. $s_{i,h}$ is the GRS of individual i for disease h and is defined as the sum of risk-allele dosages weighted by effect sizes:

$$s_{i,h} = \sum_l \beta_{l,h} g_{l,i}, \quad \text{(Equation 3)}$$

where l iterates over known risk alleles for h, $\beta_{l,h}$ is the effect size of l for h, and $g_{l,i}$ is the dosage of l in individual i. $\alpha_{j,h}$ is the effect of $s_{i,h}$, which approximates the sample proportion of confounding disease in the collection. For a detailed description of the method, see Appendix A.

For our analysis, we adjusted for both ACPA$^+$ RA and AS. For the ACPA$^+$ RA GRS, l iterated over 47 independent SNPs associated with ACPA$^+$ RA (Table S2),[3] all four-digit *HLA-DRB1* alleles, HLA-B Asp9, HLA-DPβ1 Phe9, and HLA-A Asn77. We estimated β_l from our ACPA$^+$ RA case-control data set presented in this paper. To estimate β_l for all four-digit *HLA-DRB1* alleles in a multivariate model, we included in the logistic regression all four-digit alleles with allele frequency > 0.1%, except for the reference allele we chose (*HLA-DRB1*15:01*). To avoid reusing the same controls both to estimate β_l and to map ACPA$^-$ RA, which could result in bias as a result of overfitting, we estimated β_l for each collection by using the other five collections. Similarly, for the AS GRS, l iterated over *HLA-B*27* and 19 AS-associated SNPs that passed our quality control (QC) (Table S2).[12] We used reported effect sizes β_l in Cortes et al.[12]

Two-Step Approach for Adjusting for Heterogeneity

Using GRSs as covariates in regression might be overly conservative and could remove true associations if the causal loci are shared between the disease of interest and the confounding disease. To account for the shared genetic structure between the two RA subtypes, we employed an alternative two-step approach: (1) we estimated the confounding proportions $\alpha_{j,h}$ in Equation 2 by using GRSs based on nonshared loci first, which gave us an unbiased estimate of $\alpha_{j,h}$, and then (2) we used this $\alpha_{j,h}$ as a fixed value in the regression framework presented above. Because we did not definitively know which loci were shared, we used a heuristic to choose nonshared loci by using 38 non-MHC SNPs not associated with ACPA$^-$ RA at a nominal significance threshold ($p > 0.01$)[3] (Table S2).

Genomic-Control Inflation Factor

We assessed the genomic-control inflation factor, λ_{GC}, by testing associations at "reading-writing-ability SNPs" included on the Immunochip platform. Out of 1,469 SNPs, we used 1,250 that passed QC in all six collections. We obtained chi-square statistics at these SNPs by using logistic regression as described above to assess λ_{GC}.

Forward Conditional Search

Once we identified an associated marker, we forward searched further associations by including the identified marker as a covariate in the logistic regression.

Exhaustive Search

To find the best pair of associations in *HLA-DRB1* and *HLA-B* for ACPA$^-$ disease, we examined every possible combination of 495 binary markers within *HLA-DRB1* and 774 binary markers within *HLA-B* (383,130 tests). We extend the single-marker model in Equation 2 to the following two-marker model:

$$\log(\text{odds}_i) = \theta + \beta_a g_{a,i} + \beta_b g_{b,i} + \sum_{j \in \text{collections}} \delta_{i,j}\left(\gamma_i + \sum_{k=1\ldots10} \pi_{j,k}\, p_{i,k} + \sum_{h=1\ldots H} \alpha_{j,h}\, s_{i,h}\right), \quad \text{(Equation 4)}$$

where a and b are the pair of binary markers being tested. We calculated the log-likelihood difference (ΔLL) in model fit due to this pair and assessed significance by comparing the deviance ($-2 \times \Delta LL$) to a chi-square distribution with 2 degrees of freedom.

Joint Analysis of Discovery and Replication Data

In order to jointly analyze five discovery collections and a replication cohort for ACPA$^-$ disease, we combined them into one logistic regression framework, including GRSs as covariates for five discovery cohorts to adjust for heterogeneity.

Forward Search outside of HLA-DRB1 for ACPA$^+$ RA

Because *HLA-DRB1* has a very strong effect in ACPA$^+$ disease, to examine the associations beyond *HLA-DRB1*, we conditioned on the *HLA-DRB1* effects by including binary variables as covariates corresponding to all four-digit *HLA-DRB1* alleles, excluding one allele as a reference (*HLA-DRB1*15:01*). If we forward searched by conditioning on an amino acid position with m residues, such as position 9 of HLA-B, we included binary variables corresponding to the $m - 1$ residues, excluding the most frequent one.

Testing for Discordant Effect Sizes

Given a multiallelic amino acid position with m residues, we wanted to test whether the effect sizes of m residues were concordant between two different conditions (e.g., ACPA$^-$ versus ACPA$^+$). To this end, we calculated multivariate odds ratios (ORs) of residues by including in the logistic regression $m - 1$ binary markers corresponding to $m - 1$ residues, excluding one residue as the reference. Let $a_1, \ldots, a_{m-1}$ and $b_1, \ldots, b_{m-1}$ be the multivariate log ORs in two different conditions. Let $v_1, \ldots, v_{m-1}$ and $u_1, \ldots, u_{m-1}$ be their variances. To test discordance of effect sizes between two conditions, we used the statistic

$$\sum_{i=1\ldots m} \frac{(a_i - b_i)^2}{v_i + u_i}, \quad \text{(Equation 5)}$$

which is chi-square distributed with $m - 1$ degrees of freedom under the null.

Assessing Accuracy of Fine Mapping with Simulations

To test the accuracy of our approach to adjust for clinical heterogeneity in fine mapping, we simulated an ACPA$^-$ RA case-control study confounded by ACPA$^+$ RA. We simulated a large study (50,000 case and 50,000 control subjects) to assess the asymptotic results. We first simulated control subjects by sampling with replacement from the UK control subjects. Then we assumed that specific amino acid positions were conferring risk to ACPA$^-$ RA with predefined ORs, and we sampled ACPA$^-$ RA subjects from the UK control subjects on the basis of the ORs. Finally, we replaced 26.3% of the case group with individuals randomly sampled from the UK ACPA$^+$ RA case group. We performed an association test with and without adjusting for heterogeneity to examine whether we could fine map the risk-conferring amino acid positions correctly. To adjust for heterogeneity, we used GRSs built from the effect sizes estimated from the other five cohorts, excluding the UK cohort.

Results

ACPA$^-$ RA Discovery Collection and HLA Imputation

To define HLA alleles driving ACPA$^-$ RA risk, we analyzed a discovery data set of 2,406 ACPA$^-$ RA case and 13,930 control subjects (from the UK, the US, Spain, Sweden, and the Netherlands, see Table S1) genotyped on the Immunochip custom array with 7,563 SNPs across the MHC region.[3] This platform represents greater SNP density than most standard genome-wide-association-study arrays and offers the potential for higher HLA imputation accuracy. Indeed, applying SNP2HLA,[15] we observed an overall imputation accuracy of 96.9% for four-digit HLA alleles in a subset of UK control subjects separately typed for HLA alleles (Table S3). We classified RA samples as ACPA$^-$ on the basis of anti-CCP antibody amounts according to standard clinical practice (see Material and Methods). After adjusting for ten PCs, we observed little evidence of population stratification ($\lambda_{GC} = 0.98$, see Material and Methods).

Correcting for Clinical Heterogeneity in ACPA$^-$ RA Collections

We considered that other syndromes clinically indistinguishable from ACPA$^-$ RA might be embedded within ACPA$^-$ RA and thus confound associations. Indeed, in an analysis unadjusted for clinical heterogeneity, we observed that as we defined ACPA$^-$ samples by increasing the level of stringency of the anti-CCP cutoff, the frequency of HLA-DRβ1 Val11 (the strongest risk factor for ACPA$^+$ disease) decreased in our ACPA$^-$ cohort ($p = 6.9 \times 10^{-5}$), suggesting confounding from ACPA$^+$ RA (Figure S1). We also noticed significant association at *HLA-B*27* ($p = 2.8 \times 10^{-9}$), a well-known risk factor for AS,[12,21,22] but not at *HLA-C*06:02* ($p > 0.001$), a risk factor for psoriatic arthritis.[23–25] However, as in most clinical settings, the phenotypic information that would be essential for identifying and excluding the specific individuals with conditions other than ACPA$^-$ RA was not available.

To correct for the effects of heterogeneous samples within our ACPA$^-$ cohort, we applied a statistical approach to adjust for confounding diseases (ACPA$^+$ RA and AS, Material and Methods). We constructed GRSs representing the log OR for an individual for the confounding disease on the basis of the known-risk-allele dosages weighted by effect sizes.[26–28] Then, adjusting association statistics in a logistic regression model for GRSs could successfully control for the effects of confounding diseases (see Appendix A).

ACPA$^-$ RA Is Associated with Ser11 and Leu11 in HLA-DRβ1 and Asp9 in HLA-B

After correcting for clinical heterogeneity as described above, we tested for allelic associations in ACPA$^-$ RA. Taking into account multiple hypothesis testing, we considered $p < 5.6 \times 10^{-6}$ (0.05/8,961 binary MHC-marker association tests) to be significant. After testing all amino acids and classical and SNP alleles, we observed that the strongest association was at amino acid residues at position 11 in HLA-DRβ1 (presence of Ser or Leu, OR = 1.30, $p = 1.4 \times 10^{-13}$), encoded by *HLA-DRB1* (see Figure 1A, Table 1, and Figure S2). This allele exceeded the significance of all other SNPs and classical alleles that we tested. The variation of amino acid residues at this position was attributable to a triallelic SNP (rs9269955, G/C/A) and a quadallelic SNP (rs17878703) at the first and second base positions of the codon, respectively. The association at position 11 was statistically indistinguishable ($p > 0.09$) from the association at position 13 (presence of Ser, Gly, or Phe, OR = 1.29, $p = 4.7 \times 10^{-13}$). The most strongly associated classical allele was *HLA-DRB1*03* ($p = 6.7 \times 10^{-10}$).[13,14] After conditioning on *HLA-DRB1*03*, we observed that Ser11+Leu11 remained highly significant ($p = 2.4 \times 10^{-8}$), suggesting that *HLA-DRB1*03* does not fully explain *HLA-DRB1* associations. We also observed a separate, strong association 23 kb away from *HLA-B* at SNP rs9266669 (OR = 1.38, $p = 4.0 \times 10^{-13}$; Figure 1A). This SNP was statistically indistinguishable ($p > 0.01$) from the presence of Asp9 in HLA-B (OR = 1.39, $p = 2.7 \times 10^{-12}$); these two alleles were in tight LD ($r^2 = 0.8$). HLA-B Asp9 was almost perfectly correlated with *HLA-B*08* in our data set ($r^2 = 0.997$). The *HLA-B*08* classical allele, Asp9, and SNP rs9266669 thus could not be distinguished on the basis of genetics alone. Both of these amino acid sites mapped to the binding grooves of their respective HLA receptors (Figure 2).

The *HLA-DRB1* and *HLA-B* associations were independent of each other and explained most of the MHC association with ACPA$^-$ RA. After conditioning on Ser11+Leu11 effects in HLA-DRβ1, we observed that rs9266669 in *HLA-B* (or Asp9 in HLA-B) remained the most significant association ($p = 2.0 \times 10^{-7}$, OR = 1.27; Figure 1B). Similarly, we observed that after conditioning on Asp9 in HLA-B, Ser11+Leu11 in HLA-DRβ1 remained the most significant association ($p = 1.0 \times 10^{-7}$, OR = 1.22; Figure 1C). When we conditioned on both Ser11+Leu11 in HLA-DRβ1 and Asp9 in HLA-B, no further significant association was found ($p > 0.0007$; Figure 1D).

Because the so-called 8.1 ancestral haplotype[29] harbors both HLA-DRβ1 Ser11 and HLA-B Asp9, we considered

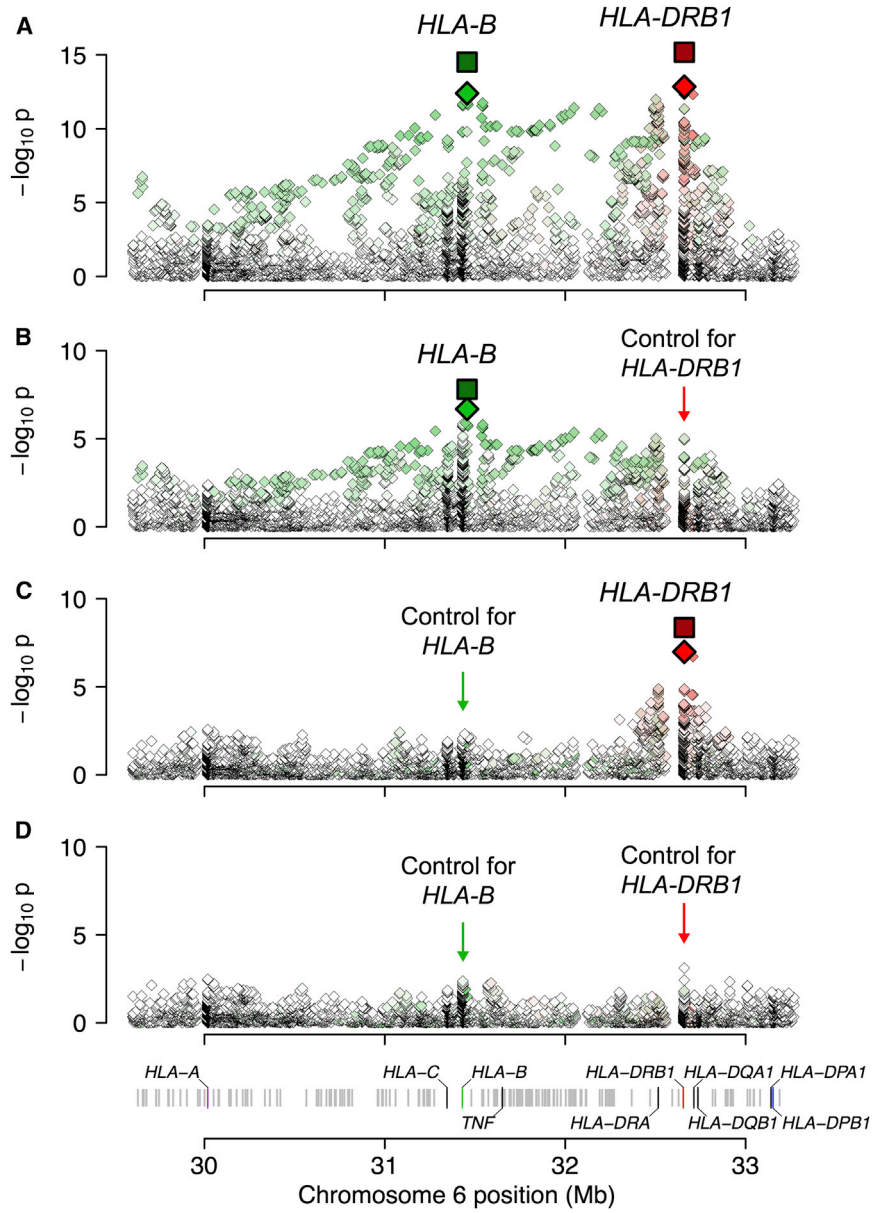


Figure 1. Association Results within the MHC to ACPA$^-$ RA

(A) We observed the most significant association at position 11 of HLA-DRβ1 (encoded by *HLA-DRB1*), where Ser and Leu conferred risk (red diamond). We also observed an independent association at SNP rs9266669, which was statistically indistinguishable from HLA-B Asp9 (green diamond). The dark-red and dark-green squares denote the statistical significance of the two positions in a joint analysis including both discovery and replication data.

(B) Conditioning on HLA-DRβ1 Ser11+Leu11, we found that the association at rs9266669 remained the most significant.

(C) Conditioning on HLA-B Asp9, we found that the association at HLA-DRβ1 Ser11+Leu11 remained the most significant.

(D) Conditioning on both HLA-DRβ1 Ser11+Leu11 and HLA-B Asp9, we did not observe any more statistically significant association within MHC ($p > 0.0007$).

the possibility that these associations were driven by that haplotype alone and not the individual amino acid sites. Given that our imputation provided phased haplotypes spanning the whole MHC region, we inferred the ancestral haplotype dosage for each individual. Then, using a trivariate logistic regression model including dosages for the 8.1 ancestral haplotype, HLA-DRβ1 Ser11+Leu11, and HLA-B Asp9, we observed that association at the ancestral haplotype was not significant ($p = 0.21$). In contrast, the other two HLA amino acid variables retained statistical significance even after adjustment for the effect of the 8.1 ancestral haplotype ($p = 1.6 \times 10^{-7}$ at HLA-DRβ1 Ser11+Leu11 and $p = 3.4 \times 10^{-3}$ at HLA-B Asp9). These results suggest that the association was driven primarily by the amino acid sites and not by the effect of the 8.1 haplotype alone.

We further considered that our approach to correcting for heterogeneity might be conservative and might remove shared loci between two subtypes of RA. To address this concern, we developed a two-step alternative approach that estimates the confounding proportion (proportion of misdiagnosed ACPA$^+$ RA samples within ACPA$^-$ RA cohorts) by using a GRS calculated on the basis of an approximated set of nonshared loci (i.e., known loci associated with ACPA$^+$ RA but with $p > 0.01$ association in ACPA$^-$ RA) and then regresses out only this amount from the model (see Material and Methods). The confounding proportion estimates by this approach were comparable to the estimates by the previous approach with the full GRS (mean proportion across cohorts was 26.3% with the full GRS and 28.3% with the nonshared-loci GRS; see Figure S3). Consistent with the previous approach, this two-step approach produced the most significant associations at rs9266669 ($p = 1.8 \times 10^{-13}$, OR = 1.38 at HLA-B Asp9) and HLA-DRβ1 Ser11+Leu11 ($p = 2.3 \times 10^{-13}$, OR = 1.27). Again, these two associations were independent ($p = 5.4 \times 10^{-8}$).

Replicating HLA Associations in a Clinically Homogeneous ACPA$^-$ Collection

We wanted to validate these findings in an independent cohort without significant clinical heterogeneity. To this end, we assessed association in an independent data set of 427 phenotypically homogeneous ACPA$^-$ individuals and 1,691 control subjects (Swedish EIRA). According to a state-of-the-art commercially unavailable assay,[8] these

Table 1. Effect Estimates for Amino Acids Associated with Risk of ACPA$^-$ and ACPA$^+$ RA

RA Subtypes	HLA Protein	Amino Acid Position	Amino Acid Residue	OR after Adjustment for Known Associated Positions (95% CI)			Frequency in Control Group	Frequency in Case Group	Classical Alleles
				Discovery	Replication	Joint			
ACPA$^-$	HLA-DRβ1	11	Ser+Leu	1.22 (1.14–1.32)	1.22 (1.04–1.43)	1.22 (1.14–1.31)	0.514	0.548	*HLA-*01*, *HLA-*03*, *HLA-*08*, *HLA-*11*, *HLA-*12*, *HLA-*13*, *HLA-*14*
	HLA-B	9	Asp	1.27 (1.15–1.40)	1.23 (0.99–1.52)	1.26 (1.15–1.38)	0.131	0.161	*HLA-*08*
ACPA$^+$	HLA-A	77	Asn	0.85 (0.81–0.90)			0.343	0.279	*HLA-*01*, *HLA-*23*, *HLA-*24*, *HLA-*26*, *HLA-*29*, *HLA-*30*, *HLA-*36*, *HLA-*80*

For each amino acid identified in this study, we show the OR and 95% confidence interval (95% CI), unadjusted frequencies in the case and control groups, and corresponding classical HLA alleles. All ORs were conditioned on known associated positions; for ACPA$^-$ RA, we estimated ORs of HLA-DRβ1 Ser11+Leu11 and HLA-B Asp9 by conditioning on each other. For ACPA$^+$ RA, we estimated the OR of HLA-A Asn77 by conditioning on all alleles at *HLA-DRB1*, amino acids at HLA-B position 9, and amino acids at HLA-DPβ1 position 9. See Table S7 for the complete table, including previously identified positions.

ACPA$^-$ individuals were negative for not only anti-CCP antibody but also antibodies for four specific citrullinated peptide antigens. We also excluded *HLA-B*27*-positive individuals (>90% sensitive for AS) from case and control groups. We tested for association without any adjustment for heterogeneity. We confirmed associations both at HLA-DRβ1 Ser11+Leu11 ($p = 5.8 \times 10^{-4}$, OR = 1.28) and at HLA-B Asp9 ($p = 2.6 \times 10^{-3}$, OR = 1.34) with comparable effect sizes (Table 1). These associations were again independent of each other. Conditioning on HLA-DRβ1 Ser11+Leu11, we observed an independent effect at HLA-B Asp9 (p = 0.03, OR = 1.23). Conversely, conditioning on HLA-B Asp9, we observed an independent effect at HLA-DRβ1 Ser11+Leu11 (p = 0.007, OR = 1.22).

In a joint analysis of the discovery and replication cohorts, we observed increased significance at both HLA-DRβ1 and HLA-B positions ($p = 6.7 \times 10^{-16}$ and OR = 1.30 for HLA-DRβ1 Ser11+Leu11; $p = 5.3 \times 10^{-14}$ and OR = 1.38 for HLA-B Asp9; Figure 1A and Table S4) and that their effects were independent ($p < 2 \times 10^{-8}$; Figures 1B and 1C and Table S4). Conditioning on both of these effects, we observed no other independent association throughout the MHC (p > 0.0002).

Exhaustive Search Confirms Associations with Ser11 and Leu11 in HLA-DRβ1 and Asp9 in HLA-B

Because the conditional forward search might miss the best explanations, we exhaustively tested every possible pair of binary markers in *HLA-DRB1* and *HLA-B* in a joint analysis. Out of 383,130 pairs we tested, HLA-DRβ1 Ser11+Leu11 and HLA-B Asp9 in HLA-B (or equivalently *HLA-B*08* and *HLA-B*0801*) constituted the most significant pair ($p = 1.1 \times 10^{-20}$; Table S5), confirming that our model provides the most parsimonious explanation of the data.

Associations Are Independent of Rheumatoid Factor Status

We examined whether the associations we identified were independent of rheumatoid factor (RF) status. We obtained RF data for 1,016 affected individuals in the UK cohort; 470 individuals (46%) were RF$^+$, and 546 individuals (54%) were RF$^-$. We stratified the samples into two groups on the basis of RF status. The associations were consistent between the two groups in that they showed the same direction of effects at both HLA-DRβ1 Ser11+Leu11 and HLA-B Asp9 (Table S6). We observed that effect sizes tended to be greater in the RF$^+$ subjects than in the RF$^-$ subjects at both loci (p = 0.02). A thorough investigation of this phenomenon will require larger sample sizes.

Asn77 at HLA-A Is Associated with ACPA$^+$ RA

We also mapped associations within the MHC to ACPA$^+$ RA in 7,279 ACPA$^+$ RA subjects and 15,870 control subjects (see Table S1 and Material and Methods). We observed little evidence of stratification after adjusting for ten PCs ($\lambda_{GC} = 1.07$). We confirmed previously published associations in HLA-DRβ1 at amino acid positions 11 ($p < 10^{-692}$), 71 ($p < 10^{-37}$), and 74 ($p < 10^{-23}$) (Table S7). Conditioning on *HLA-DRB1* alleles, we confirmed associations at Asp9 in HLA-B ($p < 10^{-36}$, OR = 1.93) and Phe9 in HLA-DPβ1 ($p < 10^{-19}$, OR = 1.31)[6] (Figure S4). Conditioning on all of these previously known associated positions (the *HLA-DRB1* alleles, position 9 in HLA-B, and position 9 in HLA-DPβ1), we observed an independent association with ACPA$^+$ RA with the presence of Asn77 in HLA-A ($p = 2.7 \times 10^{-8}$, OR = 0.85; Figure S4D and Table 1). Similar to the other amino acid sites associated with RA,[6] position 77 in HLA-A was also located in the binding groove (Figure 2 and Figure S5). We previously observed that Ser77 in HLA-A confers protection in HIV controllers.[31] After conditioning on this sixth position, we observed no convincing associations ($p > 4 \times 10^{-6}$).

Discussion

In this study, we observed that associations with ACPA$^-$ RA within the MHC were driven by *HLA-DRB1* and *HLA-B*. In

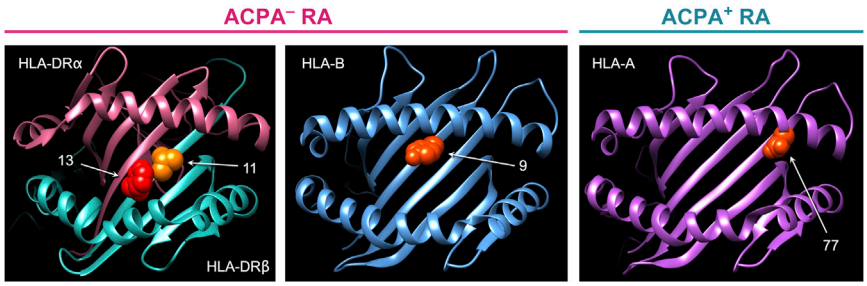


Figure 2. 3D Models of Amino Acid Positions Identified in This Study
Key amino acid positions are highlighted as spheres. We used Protein Data Bank entries 3pdo (HLA-DR), 2bvp (HLA-B), and 1x7q (HLA-A) with UCSF Chimera to prepare the figure.[30] See Figure S5 for all known associated positions.

addition, we identified the specific residues and specific amino acid sites that parsimoniously explained these associations. These positions mapped to the peptide binding grooves of these receptors, pointing to an important role for antigen recognition. The success of this study was contingent on our ability to distinguish the effects from other conditions contributing to heterogeneity within the case individuals.

Intriguingly, the positions that drove ACPA$^-$ risk were the same positions that drove most risk for ACPA$^+$ RA as well (Table S8). The risk of Asp9 in HLA-B in ACPA$^-$ RA was shared with ACPA$^+$ disease but had a more modest effect size (OR = 1.38 in ACPA$^-$ versus OR = 1.93 in ACPA$^+$). This allele, also associated with myasthenia gravis,[32] might affect nonspecific immune reactivity.

In contrast, at position 11 of HLA-DRβ1, different residues drove risk of the two diseases (discordance $p < 2.9 \times 10^{-107}$; Figure 3). For example, Ser11 conferred risk of ACPA$^-$ disease (OR = 1.31) but was protective against ACPA$^+$ disease (OR = 0.39). On the other hand, Gly11 and Pro11 showed protective effects for both subsets. We speculate that citrullinated antigens that drive ACPA$^+$ RA risk might be biochemically distinct from the antigens driving ACPA$^-$ RA risk, for example, carbamylated antigens.[33] The different set of risk and protective residues for the two disease subsets might be related to differential binding affinity and reactivity to these autoantigens.

In a multicohort study where allele frequencies can differ between cohorts, it is crucial to account for population stratification. For example, the frequency of ancestral 8.1 haplotype differed from 5% to 17% depending on cohorts (Table S9). As described in the Material and Methods, we took two approaches to account for population structure: (1) we stratified the data by country of origin, and (2) we used ten PCs to aggressively adjust for any residual population effects. The effectiveness of this standard approach is reflected in the relatively modest inflation factors for the study ($\lambda_{1,000} = 1.00$ for ACPA$^-$ RA and $\lambda_{1,000} = 1.01$ for ACPA$^+$ RA).

In this study, we addressed the issue of heterogeneity within cohorts. Like for population stratification, if the heterogeneity is present and we fail to adequately adjust for it, spurious associations can occur. For example, without adjusting for heterogeneity, the top ACPA$^-$ RA association appeared to be at Leu67 in HLA-DRβ1 ($p = 2.9 \times 10^{-28}$). Despite its remarkable significance in our heterogeneous discovery sample, Leu67 failed to replicate when we examined it in our homogenous replication data set ($p = 0.26$). In contrast, after adjusting for heterogeneity in our discovery data set, we observed the strongest effect at position 11 of HLA-DRβ1 (Table 1); not only did this effect replicate in our homogenous replication data set, but the effect sizes of each amino acid residue at that site were also highly concordant between discovery and replication sets (discordance $p > 0.4$ after adjustment; Figure S6).

To further demonstrate the potential for accounting for heterogeneity in fine mapping, we performed simulations. We simulated a study under the assumption that HLA-DRβ1 Ser11+Leu11 (OR = 1.30) and HLA-B Asp9 (OR = 1.39) confer risk, which is the model that we found in this study, and included ACPA$^+$ RA subjects in 26.3% of affected individuals (Material and Methods). Without adjustment for heterogeneity, the top association was deceivingly at HLA-DRβ1 Leu67 ($p < 10^{-331}$), which was exactly what we observed in discovery cohorts without adjusting for heterogeneity. Using our statistical approach to adjust for heterogeneity, we were able to map the correct positions we simulated; the top associations were HLA-DRβ1 Ser11+Leu11 ($p = 1.3 \times 10^{-189}$), and conditioned on this, rs2853986 ($p = 7.2 \times 10^{-59}$), which was statistically indistinguishable ($p > 0.05$) from HLA-B Asp9. We also showed that adjusting for heterogeneity not only removed spurious associations but also provided accurate estimation of the proportion of confounding samples under the null model (Figure S7).

We note that we adjusted for possible confounding from AS by correcting for AS GRSs in discovery cohorts and removing *HLA-B*27*-positive individuals in the replication cohort. This approach effectively adjusted for putative *HLA-B*27* associations with ACPA$^-$ RA if there were any. Currently, it is difficult to distinguish true *HLA-B*27* associations from confounding from AS. We expect that we will be able to accurately distinguish these two situations as we identify a greater number of non-MHC AS risk loci in the future.

The concern of clinical heterogeneity extends beyond RA to a wide range of diseases where clinical classification might be uncertain because of imperfect diagnostic tests, for example, (1) subclassification of inflammatory bowel disease (MIM 266600) into Crohn disease or ulcerative colitis or (2) distinguishing early bipolar disease (MIM

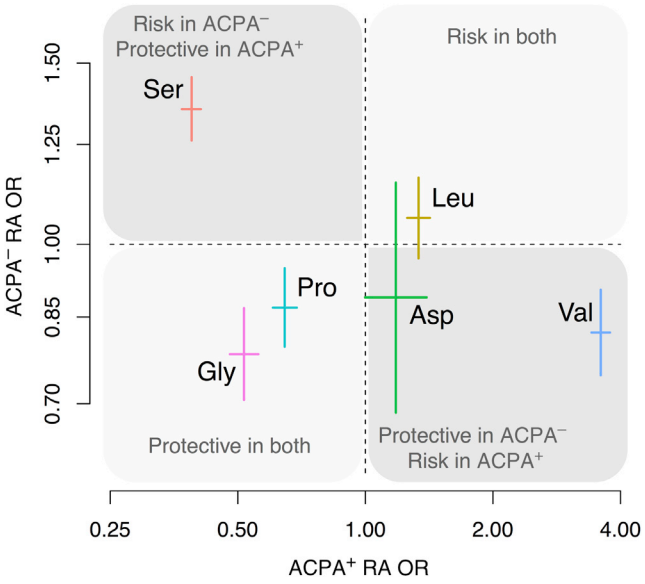


Figure 3. Distinct Effect Sizes of Amino Acid Residues at HLA-DRβ1 Position 11 for ACPA⁻ and ACPA⁺ RA
For each residue, we show the univariate OR (OR with respect to the other residues as a reference) and the 95% confidence interval. Effect sizes were distinct between the two disease subsets ($p < 2.9 \times 10^{-107}$).

125480) from major depressive disorder (MIM 608516). We expect that our statistical approach might have application to genetic studies of these conditions as well. The applicability of our approach is contingent on adequate power to detect confounding genetic effects; such power is only possible when sufficient numbers of genetic loci for confounding diseases are known. We also expect that our approach might have utility in better characterizing non-HLA loci of the conditions with clinical heterogeneity.

Our results have important implications for the clinical practice of ACPA⁻ RA. Investigators have long speculated that individuals diagnosed with ACPA⁻ RA might have other inflammatory arthritic conditions, such as AS, that mimic RA and have atypical clinical presentations. Our analysis supports this; we estimated here that each ACPA⁻ RA cohort contained 4%–11% of the affected individuals who most likely had AS and 15%–37% of affected individuals who most likely had ACPA⁺ RA (Table S10 and Figure S3). We note the possibility that other conditions that we did not account for, such as Sjögren syndrome (MIM 270150),[34] might have been included within the ACPA⁻ RA samples. These subjects were identified through research protocols, and in clinical practice, these diagnostic uncertainties can be even more pronounced. Clinical misclassifications can be particularly concerning in this setting given that optimal pharmacological treatment and long-term prognosis for these different arthritic conditions vary. Our data not only underscore the need for more accurate clinical tests than the conventional anti-CCP antibody testing but also illuminate the potential role of genetic data in helping categorize individuals with ACPA⁻ inflammatory arthritis.

Appendix A

Asymptotic Mean of Effect-Size Estimate in the Presence of Confounding

We first consider linear regression for quantitative traits. We assume a single locus, which we will extend to multiple loci later. Suppose that two groups of samples are mixed in a cohort. Let x_1 and x_2 be the genotype vectors of the two groups at the locus and y_1 and y_2 be the phenotype vectors. Let β_1 and β_2 be the effect sizes, such that the true model is $y_1 = x_1\beta_1 + \varepsilon_1$ and $y_2 = x_2\beta_2 + \varepsilon_2$, where ε_1 and ε_2 are error terms. Without loss of generality, assume that x_1, x_2, y_1, and y_2 have zero mean. Because of sample mixture, what we observe are $x = (x_1^T|x_2^T)^T$ and $y = (y_1^T|y_2^T)^T$. The standard linear regression formula gives us the least-squares estimate of effect size:

$$\begin{aligned}\widehat{\beta} &= (x^T x)^{-1} x^T y \\ &= \left(x_1^T x_1 + x_2^T x_2\right)^{-1}\left(x_1^T \,|\, x_2^T\right)\left(\left(x_1\beta_1 + \varepsilon_1\right)^T \,|\, \left(x_2\beta_2 + \varepsilon_2\right)^T\right)^T \\ &= \left(x_1^T x_1 + x_2^T x_2\right)^{-1}\left(\left(x_1^T x_1\beta_1 + x_1^T\varepsilon_1\right) + \left(x_2^T x_2\beta_2 + x_2^T\varepsilon_2\right)\right) \\ &= \left(x_1^T x_1 + x_2^T x_2\right)^{-1}\left(\left(x_1^T x_1\right)\left(\beta_1 + \left(x_1^T x_1\right)^{-1} x_1^T\varepsilon_1\right)\right. \\ &\qquad \left. + \left(x_2^T x_2\right)\left(\beta_2 + \left(x_2^T x_2\right)^{-1} x_2^T\varepsilon_2\right)\right)\end{aligned}$$

Given that $E[(x_1^T x_1)^{-1} x_1^T \varepsilon_1] = 0$ and $E[(x_2^T x_2)^{-1} x_2^T \varepsilon_2] = 0$,

$$E\left[\widehat{\beta}\right] = \left(x_1^T x_1 + x_2^T x_2\right)^{-1}\left(x_1^T x_1\beta_1 + x_2^T x_2\beta_2\right)$$

If we assume that the minor allele frequency of the variant is the same for the two groups and the genotypes follow Hardy-Weinberg equilibrium, $(x_1^T x_1)/(x_2^T x_2) \approx N_1/N_2$, where N_1 and N_2 are the sample sizes of the two groups. Thus, the effect-size estimate asymptotically converges to an average effect size weighted by the sample sizes of two groups.

This result has the following implication. Suppose that β_1 is the true effect size of interest and β_2 is the effect size for confounding samples. Consider the null model ($\beta_1 = 0$). What we observe will be $E[\widehat{\beta}] = \alpha\beta_2$, where α is the confounding proportion. Thus, we will have spurious association ($E[\widehat{\beta}] \neq 0$). Suppose that we build GRSs with respect to confounding disease as $s = x\beta_2$. If we regress out s as a covariate, it will remove spurious association. Moreover, the regression coefficient of s will be an unbiased estimator of α.

Under the alternative model ($\beta_1 \neq 0$), using risk score as a covariate might be conservative and remove true association. If we know α a priori, one approach is fixing the coefficient of s to the constant α. That is, we subtract $s\alpha = x\beta_2\alpha$ from y. This approach will retain true association. The effect-size estimate can still be conservative, given that what we would want to subtract is actually $x(\beta_2 - \beta_1)\alpha$, which is unknown.

Logistic Regression

Similar results extend to logistic regression. For simplicity, we assume the null model (true OR is 1). Suppose that α% of the case group is confounded by a disease whose OR is $\gamma \neq 1$. Let p be the control minor allele frequency. Then, the asymptotic mean of the observed log OR $\widehat{\beta}$ will be

$$E\left[\widehat{\beta}\right] = \pi = \log \frac{(\alpha p_A + (1-\alpha)p)(1-p)}{(\alpha(1-p_A) + (1-\alpha)(1-p))p},$$

where $p_A = \gamma p/((\gamma - 1)p + 1)$ is the case minor allele frequency of the confounding disease. Thus, we will have spurious association ($E[\widehat{\beta}] \neq 0$).

If γ is small, we can establish an approximate relationship, $\pi \approx \alpha \log(\gamma)$, which we show by simulations (Figure S8). Thus, using risk score $s = \log(\gamma)x$ as a covariate, we can not only remove spurious association but also approximate α from the regression coefficient of s.

Generalization to Multiple Loci

We can generalize our approach to multiple loci. Suppose that we know m independent loci associated with the confounding disease. Let $\beta_1, \ldots, \beta_m$ be their effect sizes. We build GRSs for each individual locus,

$$s_i = x_i \beta_i \quad i \in \{1, \ldots, m\},$$

where x_i is the genotype vector at locus i. In order to estimate the confounding proportion α, we look at all loci together by including all s_i in the regression:

$$y = \alpha s_1 + \alpha s_2 + \ldots + \alpha s_m + \varepsilon.$$

Application to logistic regression is also straightforward. Because α is invariant across loci, this is equivalent to the model using a combined GRS, $y = \alpha S + \varepsilon$, where $S = \sum s_i = \sum x_i \beta_i$, which results in the approach presented in the Material and Methods. The advantage of a combined GRS over multiple loci is that it can be less conservative under the alternative model. For example, if we test locus i and include s_i as a covariate, it will remove true association. However, if we include S as a covariate, the information from other loci $(s_1, s_2, \ldots, s_{i-1}, s_{i+1}, \ldots, s_m)$ will help in finding correct α and preventing overly regressing out s_i. Another possible way to more strictly prevent overly regressing out GRS can be estimating α with nonoverlapping loci first, as presented in the Material and Methods.

Supplemental Data

Supplemental Data include eight figures and ten tables and can be found with this article online at http://www.cell.com/ajhg.

Acknowledgments

This work was supported by funds from the National Institutes of Health (K08AR055688, 1R01AR062886-01, 1R01AR063759-01A1, and 5U01GM092691-04), the Arthritis Foundation, and the Doris Duke Foundation and in part through the Be the Cure For Rheumatoid Arthritis grant funded by the Innovative Medicine Initiative program from the European Union. This research used data provided by the Type 1 Diabetes Genetics Consortium (a collaborative clinical study sponsored by the National Institute of Diabetes and Digestive and Kidney Diseases, National Institute of Allergy and Infectious Diseases, National Human Genome Research Institute, National Institute of Child Health and Human Development, and Juvenile Diabetes Research Foundation International). A.Z. was supported by a grant from the Dutch Reumafonds (11-1-101) and the Rosalind Franklin Fellowship from the University of Groningen (the Netherlands). These data also included data generously provided by the Rheumatoid Arthritis International Consortium. P.I.W.d.B. is the recipient of a Vidi award from the Netherlands Organization for Scientific Research (project 016.126.354). This work was partially supported by the Red de Investigación en Inflamación y Enfermedades Reumáticas (RD12/0009) of the Redes Temáticas de Investigación Cooperativa en Salud from the Instituto de Salud Carlos III Health Ministry (Spain).

Received: December 16, 2013
Accepted: February 24, 2014
Published: March 20, 2014

Web Resources

The URLs for data presented herein are as follows:

Online Mendelian Inheritance in Man (OMIM), http://www.omim.org/
Protein Data Bank (PDB), http://www.rcsb.org/pdb/home/home.do

References

1. Daha, N.A., and Toes, R.E.M. (2011). Rheumatoid arthritis: Are ACPA-positive and ACPA-negative RA the same disease? Nat. Rev. Rheumatol. *7*, 202–203.
2. van der Helm-van Mil, A.H., and Huizinga, T.W. (2008). Advances in the genetics of rheumatoid arthritis point to subclassification into distinct disease subsets. Arthritis Res. Ther. *10*, 205.
3. Eyre, S., Bowes, J., Diogo, D., Lee, A., Barton, A., Martin, P., Zhernakova, A., Stahl, E., Viatte, S., McAllister, K., et al.; Biologics in Rheumatoid Arthritis Genetics and Genomics Study Syndicate; Wellcome Trust Case Control Consortium (2012). High-density genetic mapping identifies new susceptibility loci for rheumatoid arthritis. Nat. Genet. *44*, 1336–1340.
4. Ding, B., Padyukov, L., Lundström, E., Seielstad, M., Plenge, R.M., Oksenberg, J.R., Gregersen, P.K., Alfredsson, L., and Klareskog, L. (2009). Different patterns of associations with anti-citrullinated protein antibody-positive and anti-citrullinated protein antibody-negative rheumatoid arthritis in the extended major histocompatibility complex region. Arthritis Rheum. *60*, 30–38.
5. Stahl, E.A., Raychaudhuri, S., Remmers, E.F., Xie, G., Eyre, S., Thomson, B.P., Li, Y., Kurreeman, F.A.S., Zhernakova, A., Hinks, A., et al.; BIRAC Consortium; YEAR Consortium (2010). Genome-wide association study meta-analysis identifies seven new rheumatoid arthritis risk loci. Nat. Genet. *42*, 508–514.

6. Raychaudhuri, S., Sandor, C., Stahl, E.A., Freudenberg, J., Lee, H.-S., Jia, X., Alfredsson, L., Padyukov, L., Klareskog, L., Worthington, J., et al. (2012). Five amino acids in three HLA proteins explain most of the association between MHC and seropositive rheumatoid arthritis. Nat. Genet. *44*, 291–296.
7. Raychaudhuri, S., Remmers, E.F., Lee, A.T., Hackett, R., Guiducci, C., Burtt, N.P., Gianniny, L., Korman, B.D., Padyukov, L., Kurreeman, F.A.S., et al. (2008). Common variants at CD40 and other loci confer risk of rheumatoid arthritis. Nat. Genet. *40*, 1216–1223.
8. Lundberg, K., Bengtsson, C., Kharlamova, N., Reed, E., Jiang, X., Källberg, H., Pollak-Dorocic, I., Israelsson, L., Kessel, C., Padyukov, L., et al. (2013). Genetic and environmental determinants for disease risk in subsets of rheumatoid arthritis defined by the anticitrullinated protein/peptide antibody fine specificity profile. Ann. Rheum. Dis. *72*, 652–658.
9. Wiik, A.S., van Venrooij, W.J., and Pruijn, G.J.M. (2010). All you wanted to know about anti-CCP but were afraid to ask. Autoimmun. Rev. *10*, 90–93.
10. van der Linden, M.P.M., van der Woude, D., Ioan-Facsinay, A., Levarht, E.W.N., Stoeken-Rijsbergen, G., Huizinga, T.W.J., Toes, R.E.M., and van der Helm-van Mil, A.H.M. (2009). Value of anti-modified citrullinated vimentin and third-generation anti-cyclic citrullinated peptide compared with second-generation anti-cyclic citrullinated peptide and rheumatoid factor in predicting disease outcome in undifferentiated arthritis and rheumatoid arthritis. Arthritis Rheum. *60*, 2232–2241.
11. Viatte, S., Plant, D., and Raychaudhuri, S. (2013). Genetics and epigenetics of rheumatoid arthritis. Nat. Rev. Rheumatol. *9*, 141–153.
12. Cortes, A., Hadler, J., Pointon, J.P., Robinson, P.C., Karaderi, T., Leo, P., Cremin, K., Pryce, K., Harris, J., Lee, S., et al.; International Genetics of Ankylosing Spondylitis Consortium (IGAS); Australo-Anglo-American Spondyloarthritis Consortium (TASC); Groupe Française d'Etude Génétique des Spondylarthrites (GFEGS); Nord-Trøndelag Health Study (HUNT); Spondyloarthritis Research Consortium of Canada (SPARCC); Wellcome Trust Case Control Consortium 2 (WTCCC2) (2013). Identification of multiple risk variants for ankylosing spondylitis through high-density genotyping of immune-related loci. Nat. Genet. *45*, 730–738.
13. Verpoort, K.N., van Gaalen, F.A., van der Helm-van Mil, A.H.M., Schreuder, G.M.T., Breedveld, F.C., Huizinga, T.W.J., de Vries, R.R.P., and Toes, R.E.M. (2005). Association of HLA-DR3 with anti-cyclic citrullinated peptide antibody-negative rheumatoid arthritis. Arthritis Rheum. *52*, 3058–3062.
14. Irigoyen, P., Lee, A.T., Wener, M.H., Li, W., Kern, M., Batliwalla, F., Lum, R.F., Massarotti, E., Weisman, M., Bombardier, C., et al. (2005). Regulation of anti-cyclic citrullinated peptide antibodies in rheumatoid arthritis: contrasting effects of HLA-DR3 and the shared epitope alleles. Arthritis Rheum. *52*, 3813–3818.
15. Jia, X., Han, B., Onengut-Gumuscu, S., Chen, W.-M., Concannon, P.J., Rich, S.S., Raychaudhuri, S., and de Bakker, P.I.W. (2013). Imputing amino acid polymorphisms in human leukocyte antigens. PLoS ONE *8*, e64683.
16. Lee, H.-S., Irigoyen, P., Kern, M., Lee, A., Batliwalla, F., Khalili, H., Wolfe, F., Lum, R.F., Massarotti, E., Weisman, M., et al. (2007). Interaction between smoking, the shared epitope, and anti-cyclic citrullinated peptide: a mixed picture in three large North American rheumatoid arthritis cohorts. Arthritis Rheum. *56*, 1745–1753.
17. Klareskog, L., Stolt, P., Lundberg, K., Källberg, H., Bengtsson, C., Grunewald, J., Rönnelid, J., Harris, H.E., Ulfgren, A.-K., Rantapää-Dahlqvist, S., et al. (2006). A new model for an etiology of rheumatoid arthritis: smoking may trigger HLA-DR (shared epitope)-restricted immune reactions to autoantigens modified by citrullination. Arthritis Rheum. *54*, 38–46.
18. Browning, B.L., and Browning, S.R. (2009). A unified approach to genotype imputation and haplotype-phase inference for large data sets of trios and unrelated individuals. Am. J. Hum. Genet. *84*, 210–223.
19. Brown, W.M., Pierce, J., Hilner, J.E., Perdue, L.H., Lohman, K., Li, L., Venkatesh, R.B., Hunt, S., Mychaleckyj, J.C., and Deloukas, P.; Type 1 Diabetes Genetics Consortium (2009). Overview of the MHC fine mapping data. Diabetes Obes. Metab. *11* (*Suppl 1*), 2–7.
20. Price, A.L., Patterson, N.J., Plenge, R.M., Weinblatt, M.E., Shadick, N.A., and Reich, D. (2006). Principal components analysis corrects for stratification in genome-wide association studies. Nat. Genet. *38*, 904–909.
21. Brown, M.A., Pile, K.D., Kennedy, L.G., Calin, A., Darke, C., Bell, J., Wordsworth, B.P., and Cornélis, F. (1996). HLA class I associations of ankylosing spondylitis in the white population in the United Kingdom. Ann. Rheum. Dis. *55*, 268–270.
22. Reveille, J.D., Sims, A.M., Danoy, P., Evans, D.M., Leo, P., Pointon, J.J., Jin, R., Zhou, X., Bradbury, L.A., Appleton, L.H., et al.; Australo-Anglo-American Spondyloarthritis Consortium (TASC) (2010). Genome-wide association study of ankylosing spondylitis identifies non-MHC susceptibility loci. Nat. Genet. *42*, 123–127.
23. Tiilikainen, A., Lassus, A., Karvonen, J., Vartiainen, P., and Julin, M. (1980). Psoriasis and HLA-Cw6. Br. J. Dermatol. *102*, 179–184.
24. Nair, R.P., Stuart, P.E., Nistor, I., Hiremagalore, R., Chia, N.V.C., Jenisch, S., Weichenthal, M., Abecasis, G.R., Lim, H.W., Christophers, E., et al. (2006). Sequence and haplotype analysis supports HLA-C as the psoriasis susceptibility 1 gene. Am. J. Hum. Genet. *78*, 827–851.
25. Ho, P.Y.P.C., Barton, A., Worthington, J., Thomson, W., Silman, A.J., and Bruce, I.N. (2007). HLA-Cw6 and HLA-DRB1*07 together are associated with less severe joint disease in psoriatic arthritis. Ann. Rheum. Dis. *66*, 807–811.
26. Karlson, E.W., Chibnik, L.B., Kraft, P., Cui, J., Keenan, B.T., Ding, B., Raychaudhuri, S., Klareskog, L., Alfredsson, L., and Plenge, R.M. (2010). Cumulative association of 22 genetic variants with seropositive rheumatoid arthritis risk. Ann. Rheum. Dis. *69*, 1077–1085.
27. Morrison, A.C., Bare, L.A., Chambless, L.E., Ellis, S.G., Malloy, M., Kane, J.P., Pankow, J.S., Devlin, J.J., Willerson, J.T., and Boerwinkle, E. (2007). Prediction of coronary heart disease risk using a genetic risk score: the Atherosclerosis Risk in Communities Study. Am. J. Epidemiol. *166*, 28–35.
28. Meigs, J.B., Shrader, P., Sullivan, L.M., McAteer, J.B., Fox, C.S., Dupuis, J., Manning, A.K., Florez, J.C., Wilson, P.W.F., D'Agostino, R.B., Sr., and Cupples, L.A. (2008). Genotype score in addition to common risk factors for prediction of type 2 diabetes. N. Engl. J. Med. *359*, 2208–2219.
29. Price, P., Witt, C., Allcock, R., Sayer, D., Garlepp, M., Kok, C.C., French, M., Mallal, S., and Christiansen, F. (1999). The genetic basis for the association of the 8.1 ancestral haplotype (A1, B8, DR3) with multiple immunopathological diseases. Immunol. Rev. *167*, 257–274.

30. Pettersen, E.F., Goddard, T.D., Huang, C.C., Couch, G.S., Greenblatt, D.M., Meng, E.C., and Ferrin, T.E. (2004). UCSF Chimera—a visualization system for exploratory research and analysis. J. Comput. Chem. *25*, 1605–1612.
31. Pereyra, F., Jia, X., McLaren, P.J., Telenti, A., de Bakker, P.I., Walker, B.D., Ripke, S., Brumme, C.J., Pulit, S.L., Carrington, M., et al.; International HIV Controllers Study (2010). The major genetic determinants of HIV-1 control affect HLA class I peptide presentation. Science *330*, 1551–1557.
32. Gregersen, P.K., Kosoy, R., Lee, A.T., Lamb, J., Sussman, J., McKee, D., Simpfendorfer, K.R., Pirskanen-Matell, R., Piehl, F., Pan-Hammarstrom, Q., et al. (2012). Risk for myasthenia gravis maps to a (151) Pro→Ala change in TNIP1 and to human leukocyte antigen-B*08. Ann. Neurol. *72*, 927–935.
33. Shi, J., Knevel, R., Suwannalai, P., van der Linden, M.P., Janssen, G.M.C., van Veelen, P.A., Levarht, N.E.W., van der Helm-van Mil, A.H.M., Cerami, A., Huizinga, T.W.J., et al. (2011). Autoantibodies recognizing carbamylated proteins are present in sera of patients with rheumatoid arthritis and predict joint damage. Proc. Natl. Acad. Sci. USA *108*, 17372–17377.
34. Boire, G., Ménard, H.A., Gendron, M., Lussier, A., and Myhal, D. (1993). Rheumatoid arthritis: anti-Ro antibodies define a non-HLA-DR4 associated clinicoserological cluster. J. Rheumatol. *20*, 1654–1660.

The American Journal of Human Genetics, Volume *94*

Supplemental Data

Fine Mapping Seronegative and Seropositive Rheumatoid Arthritis to Shared and Distinct HLA Alleles by Adjusting for the Effects of Heterogeneity

Buhm Han, Dorothée Diogo, Steve Eyre, Henrik Kallberg, Alexandra Zhernakova, John Bowes, Leonid Padyukov, Yukinori Okada, Miguel A. González-Gay, Solbritt Rantapää-Dahlqvist, Javier Martin, Tom W.J. Huizinga, Robert M. Plenge, Jane Worthington, Peter K. Gregersen, Lars Klareskog, Paul I.W. de Bakker, and Soumya Raychaudhuri

Supplemental Figures

Figure S1: Frequency of HLA-DRβ1 Val11 in seronegative RA cases. In the US and UK cohorts, we measured the seronegative case frequency of HLA-DRβ1 Val11, a major risk factor for seropositive RA, as we increased the level of stringency of anti-CCP cutoff (i.e. as we reduced the cut-off). As we reduced the cut-off from the default values (5.0 in UK and 20.0 in US), we observed decreasing trend in the Val-11 frequency (Spearman P=6.9×10^{-5}). This suggested that uncertainties in anti-CCP testing might have caused possible confounding from seropositive RA.

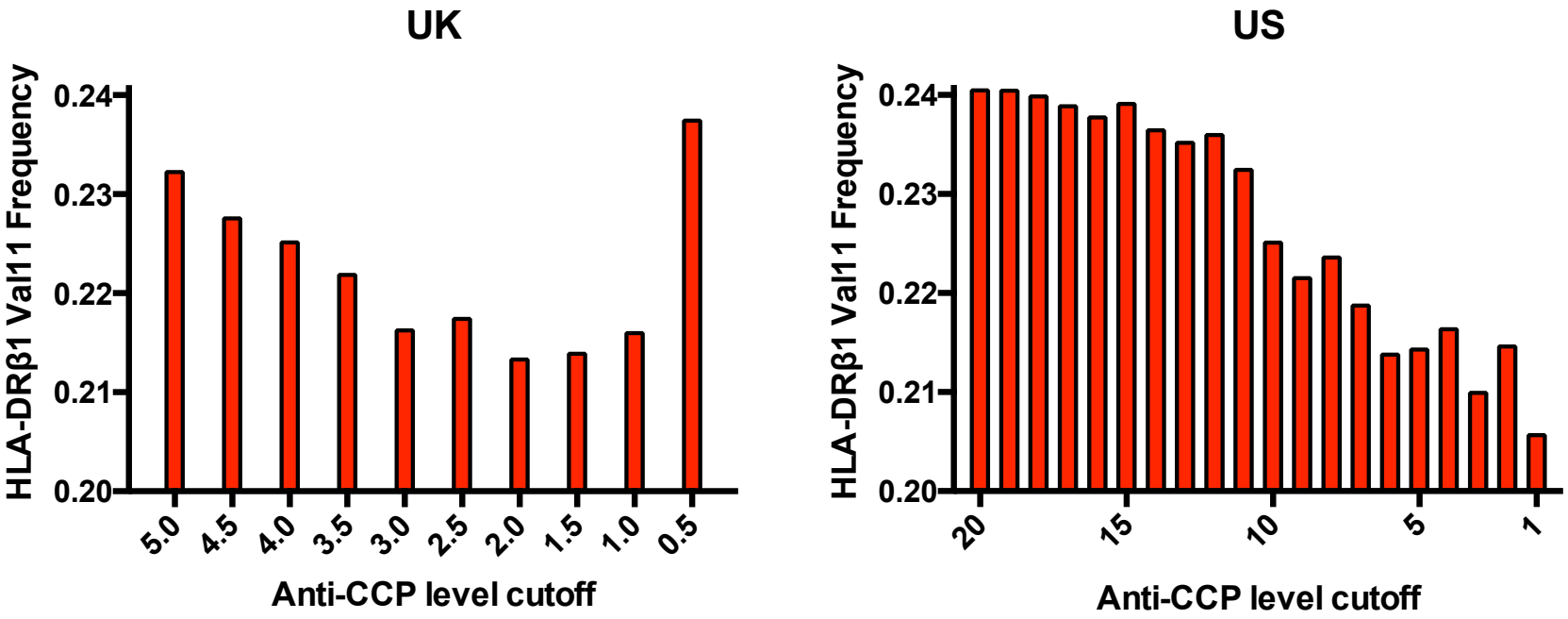


Figure S2: Effects of individual amino acids within HLA proteins on seronegative and seropositive RA. For each amino acid position, we listed the allele frequencies in cases (red) and controls (blue) along with univariate odds ratios (odds ratio of a residue taking the other residues as reference). Newly identified positions are in bold faces. For seronegative RA, we estimated odds ratios of Ser11+Leu11 of HLA-DRβ1 and Asp9 of HLA-B by conditioning on each other. For seropositive RA, we estimated odds ratios at each position by conditioning on previous positions in forward search; the effects of HLA-B are conditioned on the classical *HLA-DRB1* alleles, the effects of HLA-DPβ1 are conditioned on the *HLA-DRB1* alleles and HLA-B position 9, and the effects of HLA-A are conditioned on the *HLA-DRB1* alleles, HLA-B position 9, and HLA-DPβ1 position 9.

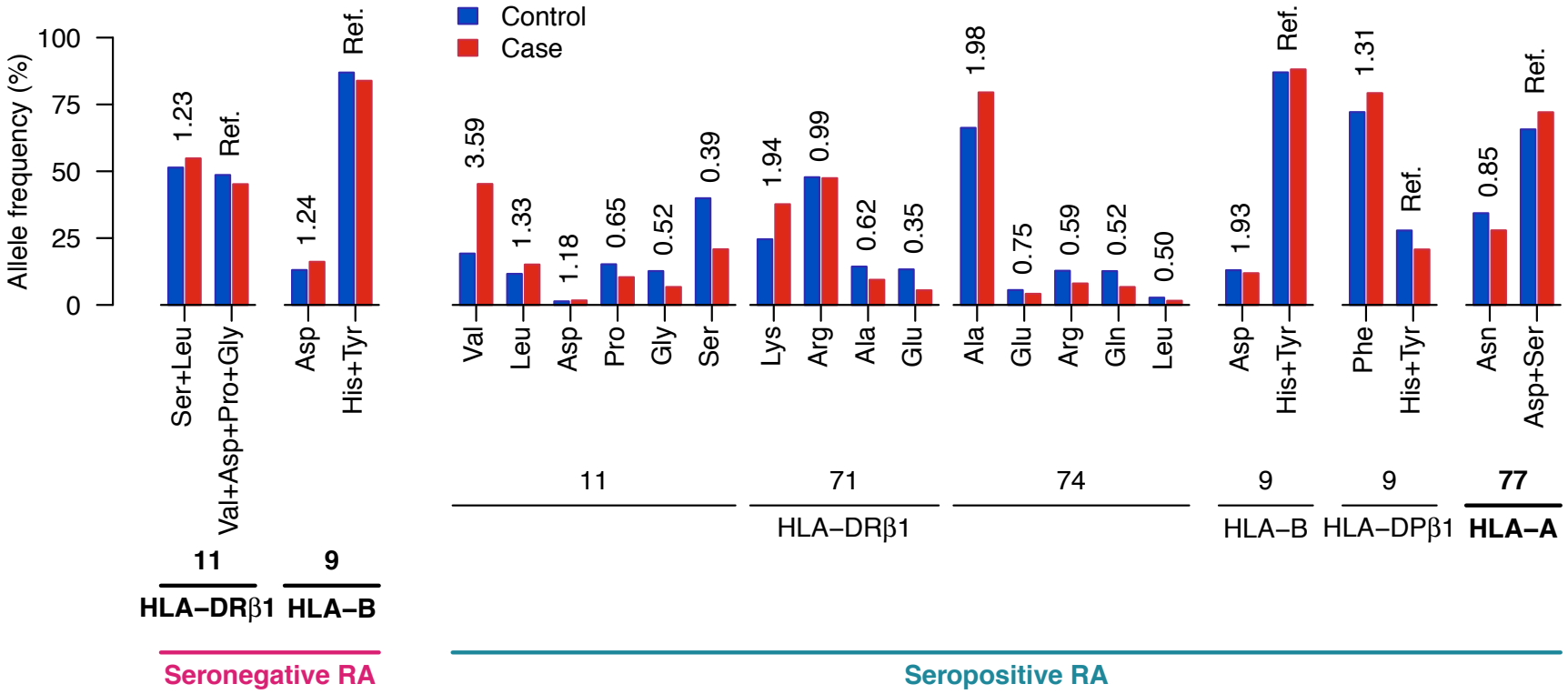


Figure S3: Comparison of confounding proportion estimates when using full GRS and non-overlapping loci GRS. For each cohort, we estimated the confounding proportion from ACPA+ RA (the proportion of samples that actually have ACPA+ RA) using genetic risk scores (GRS). First, we used GRS from the full list of known risk loci for ACPA+ RA (MHC loci in addition to 47 non-MHC loci). Then we used GRS from the selected list of loci that approximates non-overlapping loci between ACPA+ RA and ACPA- RA (38 loci that are not associated to ACPA- RA) (See Table S2). The mean estimate over the five cohorts were 26.3% for full GRS and 28.3% for non-overlapping loci GRS. Vertical lines denote 95% C.I..

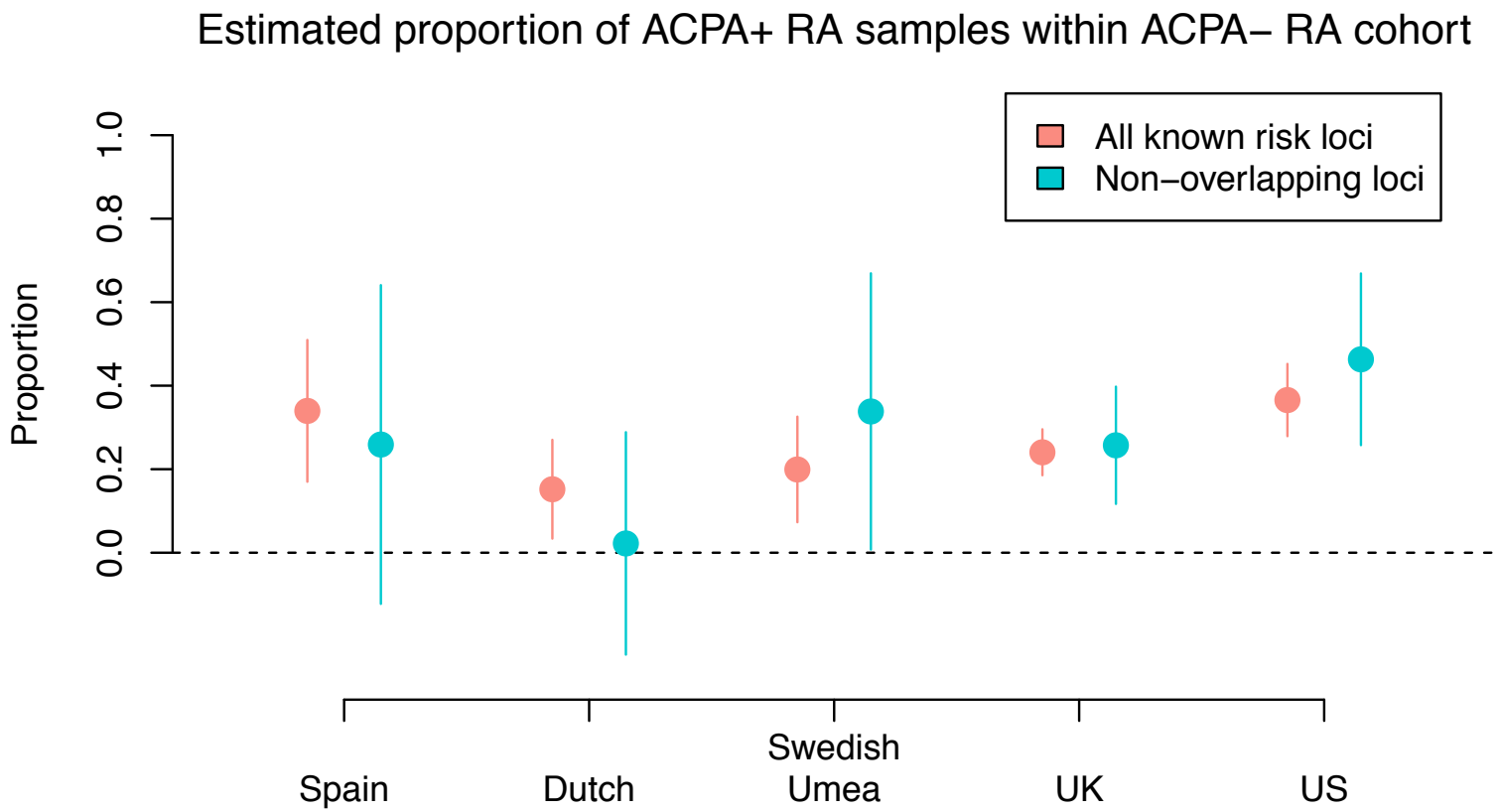


Figure S4. Association results within the MHC to seropositive rheumatoid arthritis. (**A**) We observed the most significant association in *HLA-DRB1*. (**B**) Conditioning on all *HLA-DRB1* alleles revealed an independent association at HLA-B Asp9. (**C**) Conditioning on all *HLA-DRB1* alleles and amino acids at HLA-B position 9 revealed an independent association at HLA-DPβ1 Phe9. (**D**) Conditioning on *HLA-DRB1* alleles and amino acids at HLA-B position 9 and HLA-DPβ1 position 9 revealed an independent association at HLA-A Asn77. (**E**) Conditioning on *HLA-DRB1* alleles and amino acids at HLA-B position 9, HLA-DPβ1 position 9, and HLA-A position 77 did not reveal any convincingly significant association within MHC (P>1.9×10^{-6}).

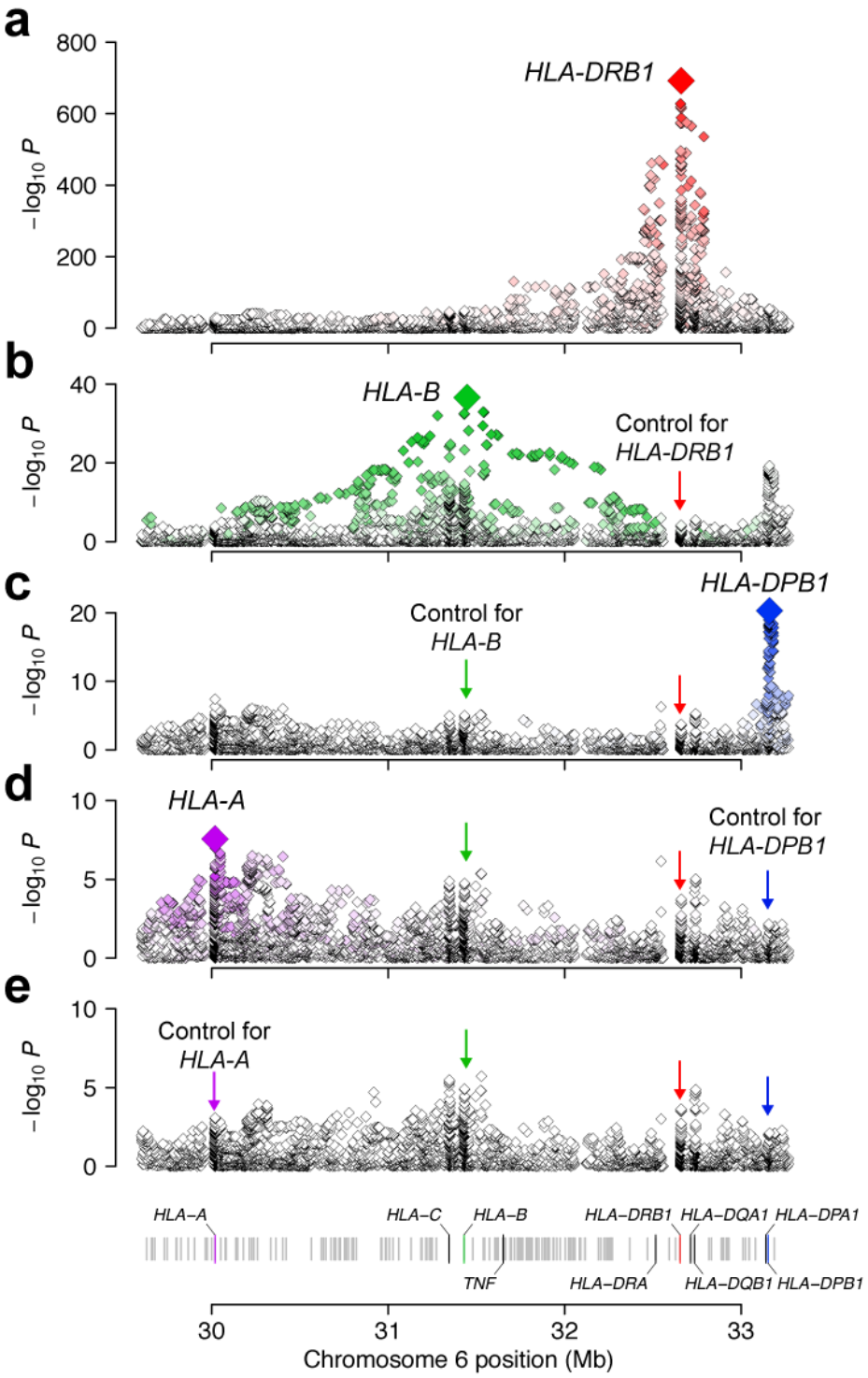


Figure S5: Overview of associated amino acid positions to seronegative and seropositive RA in three dimensional models. All associated positions are in binding grooves. Cyan/Green colors indicate associated positions to both diseases but having distinct effects depending on residues. Orange colors indicate associated positions to both diseases with shared effect size direction. Magenta colors indicate associated positions uniquely to seropositive RA. We used Protein Data Bank (PBD) entries 3pdo (HLA-DR), 3lqz (HLA-DP), 2bvp (HLA-B), and 1x7q (HLA-A) with UCSF Chimera to prepare the figure.

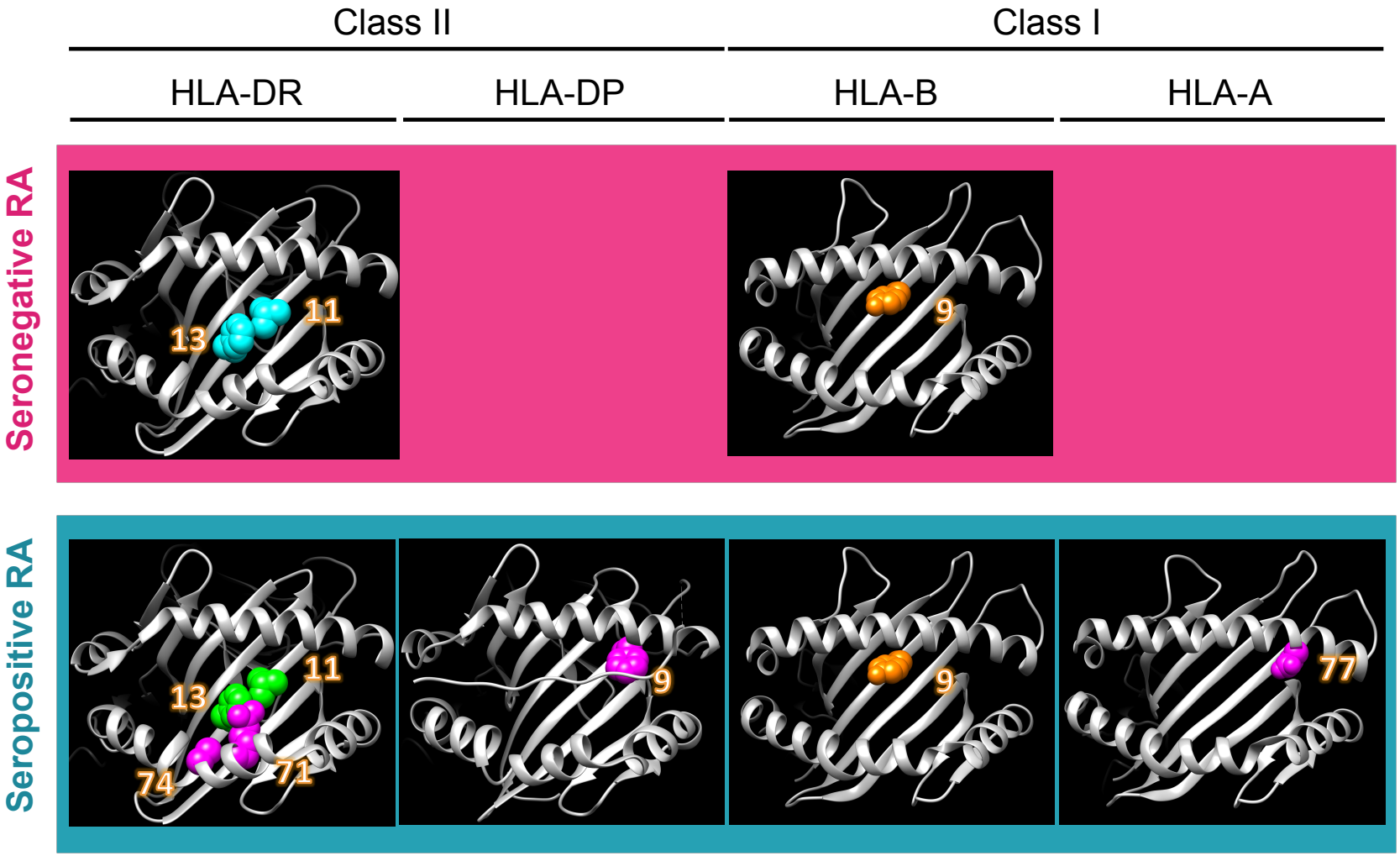


Figure S6: Replication of individual effect sizes of amino acid residues at HLA-DRβ1 position 11. We plot the univariate odds ratio (odds ratio with respect to the other residues as reference) of six residues along with 95% confidence interval, in discovery analysis versus replication analysis. **(A)** When we accounted for possible heterogeneity using risk score corrections in the discovery analysis (See **Methods**), the individual effects were well replicated. The p-value for discordance test was not significant (*P*=0.44). Green line indicates the fitted regression line, which ideally should follow the diagonal line. **(B)** If we do not adjust for risk scores in the discovery analysis, the individual effects were much less concordant to replication (discordance *P*=0.0045).

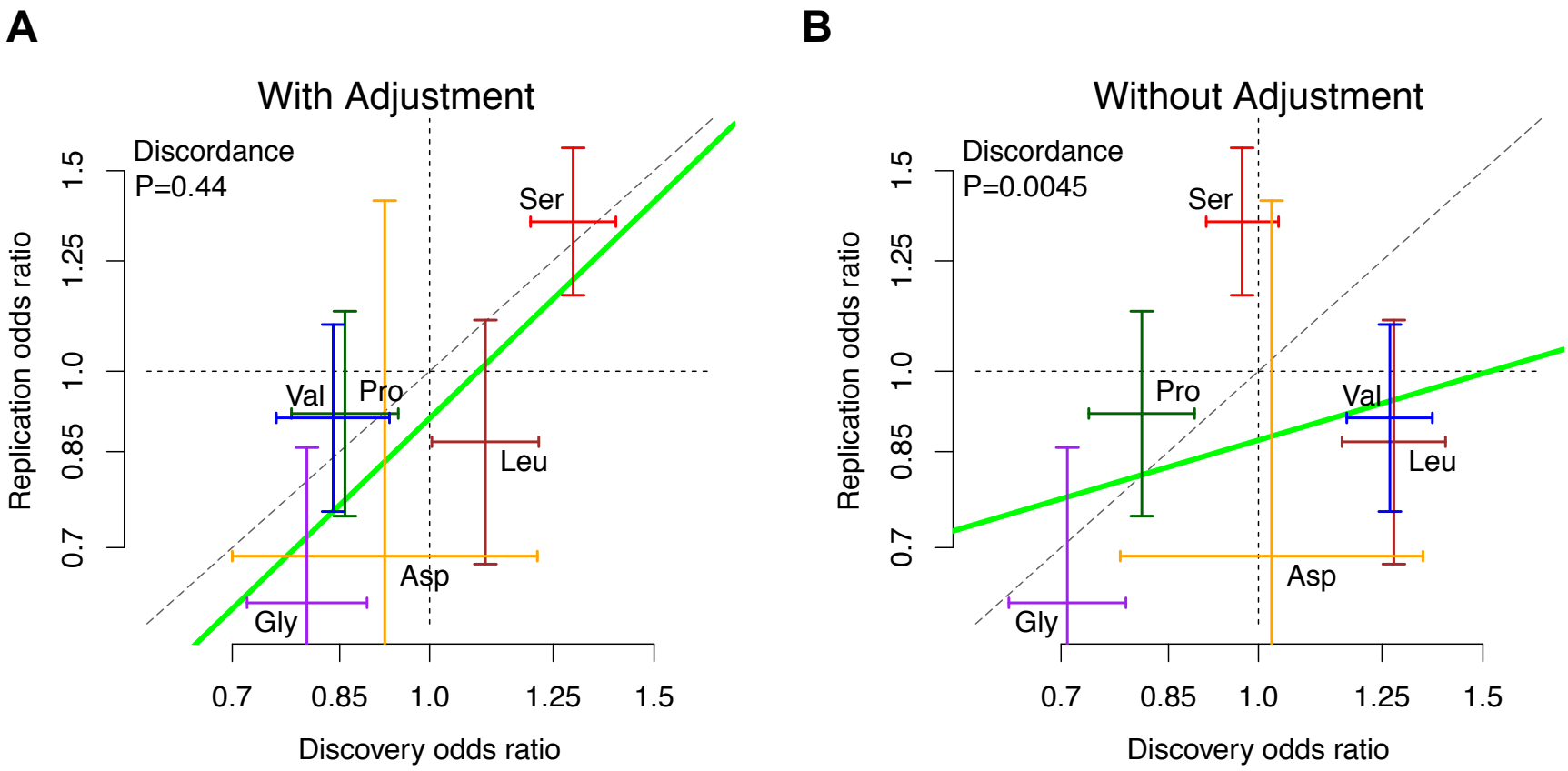
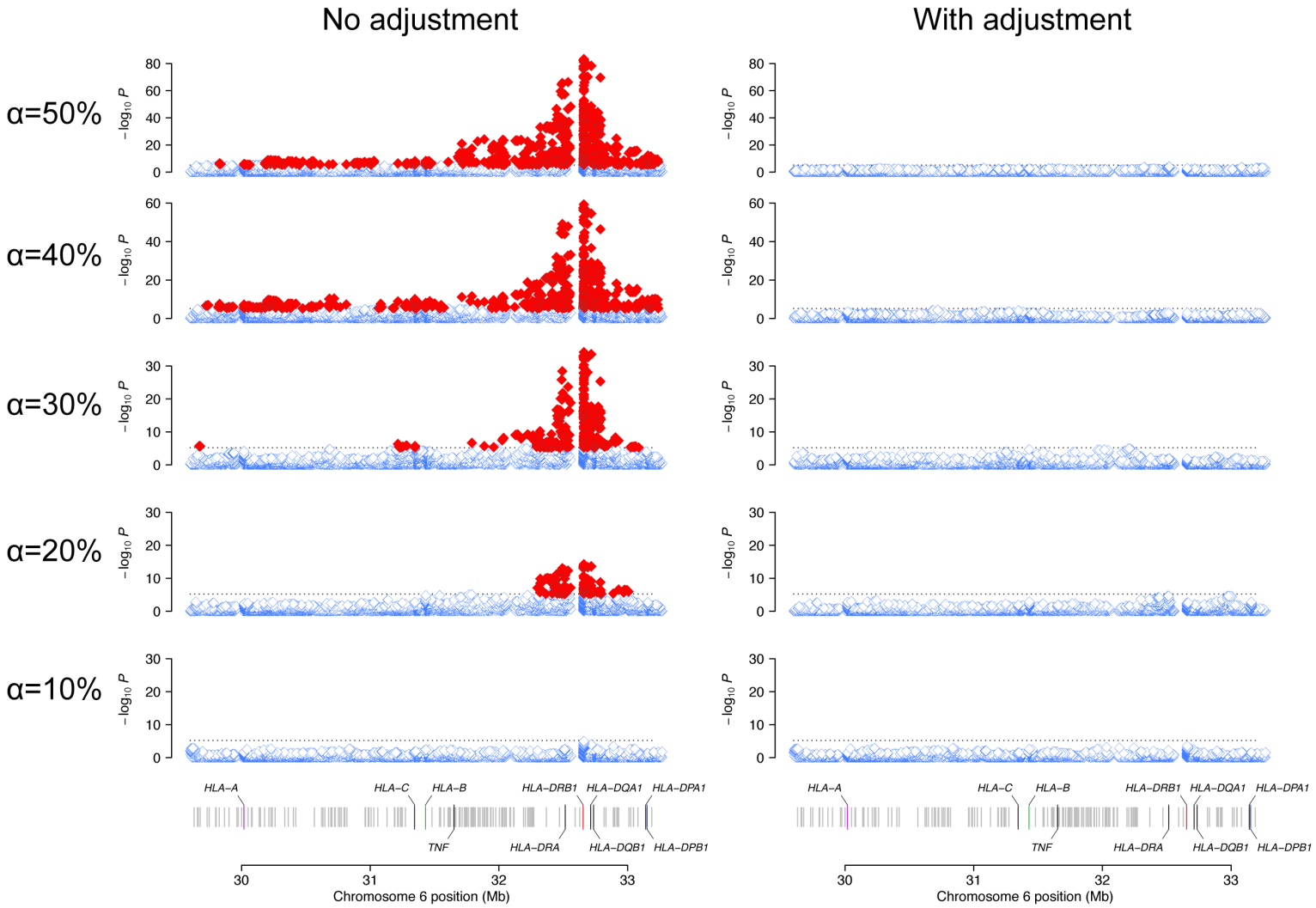


Figure S7: Simulations under the null disease model. We performed a simple simulation that splits UK controls to half and half as null cases and controls, and replaces *α*% of null cases with randomly sampled ACPA+ RA cases to simulate confounding. **(A)** Spurious associations due to the confounding exacerbated with increasing *α* (Left pane). Red diamonds are spurious associations with *P*<6E-6. After we adjust for risk scores, spurious associations disappeared (Right pane). The dotted horizontal line is the threshold 6E-6. **(B)** We approximated the sample proportion of confounding disease using risk score. Vertical lines denote 95% C.I..

A.



B.

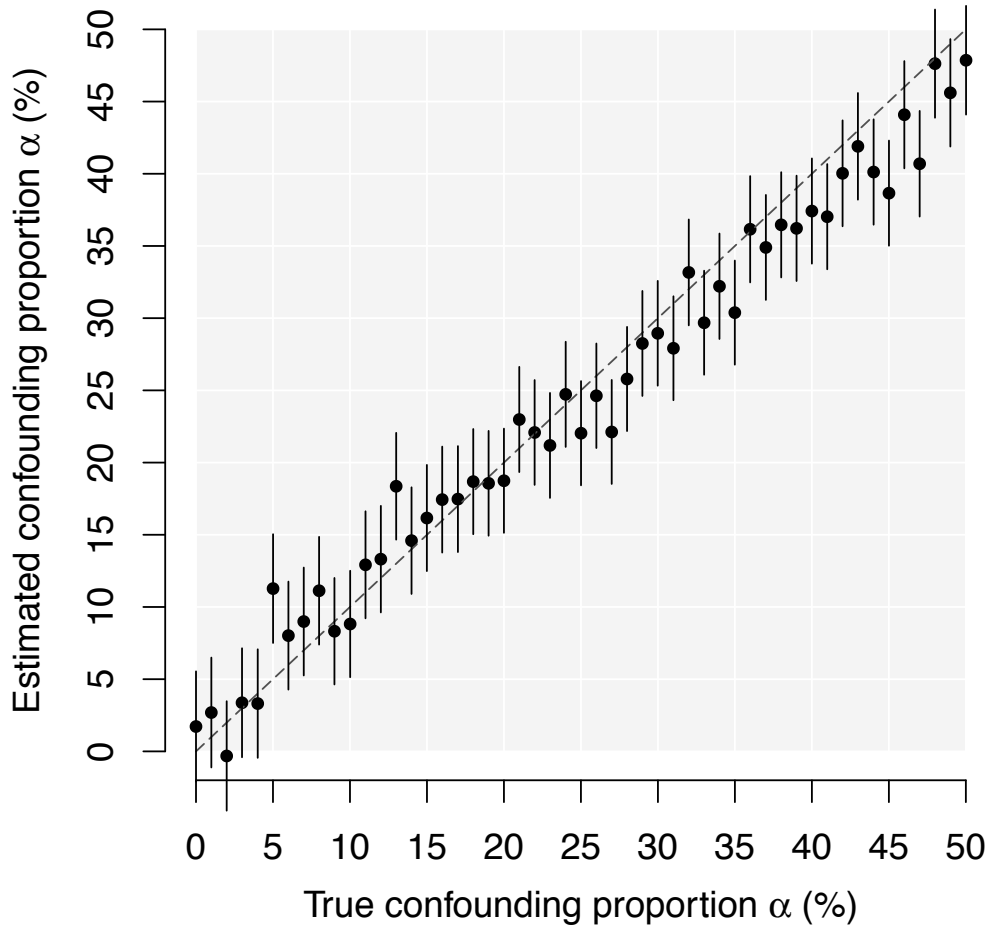
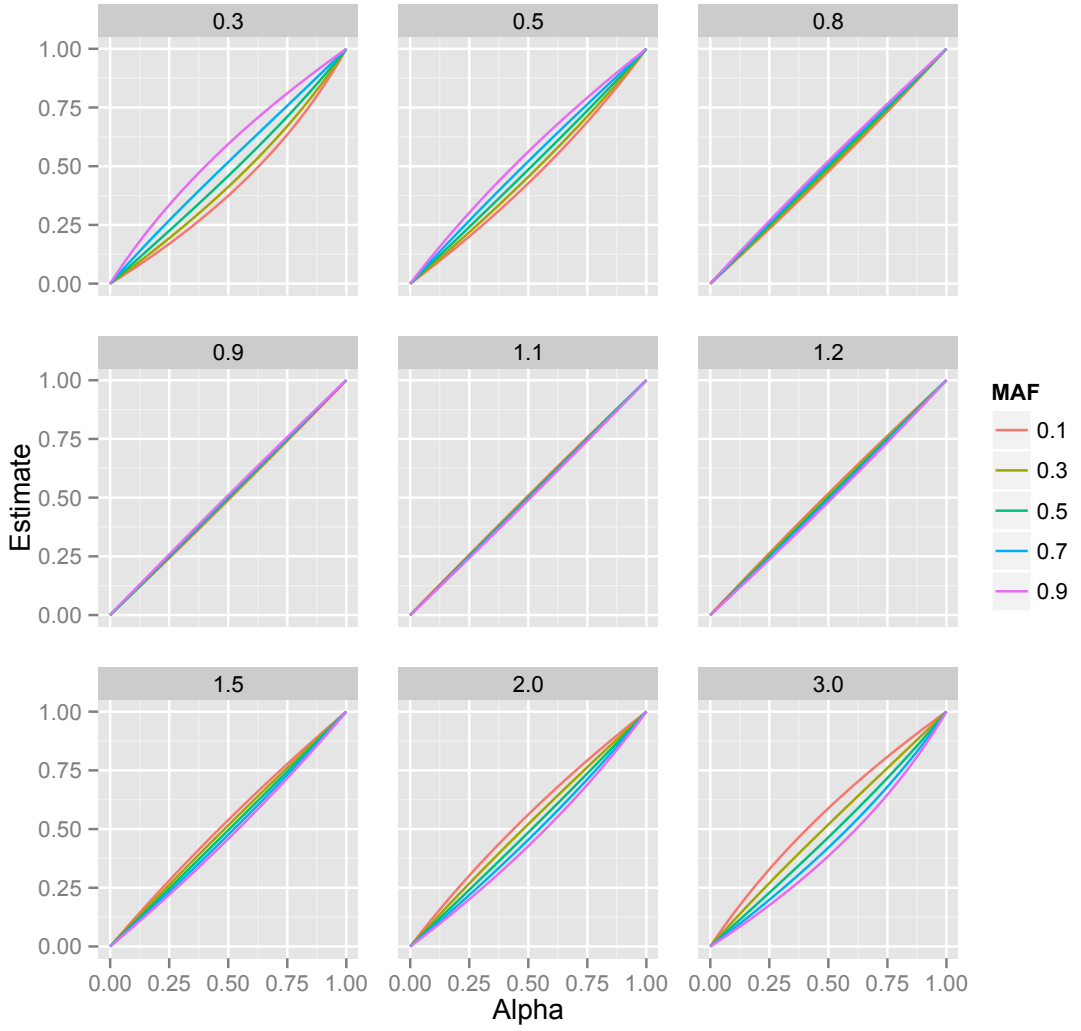

Estimated confounding proportion α (%)
0
5
10
15
20
25
30
35
40
45
50
True confounding proportion α (%)

Figure S8: Approximate relationship between confounding proportion and estimated proportion in logistic regression. We assume one locus with MAF and specific odds ratio. Then we plot for each confounding proportion α (x-axis), the expected value of the estimated proportion in logistic regression (y-axis). The numbers in grey boxes are odds ratios. Unless the odds ratio is large and the MAF is very low or high, the estimated α approximates true α well.



Supplemental Tables

Table S1: Sample Collections. (A) For seronegative RA analysis, we collected five cohorts for the discovery analysis and an independent cohort for replication. Within five cohorts, we confirmed the seronegative status of samples using the conventional anti-CCP testing (yellow colors). In replication cohort, in order to stringently define seronegative samples, we additionally applied newly developed ACPA-specific sensitive testing (blue color) (**Material and Methods**). (B) For seropositive RA analysis, we collected six cohorts. We confirmed seropositive status of these samples using the conventional anti-CCP testing (yellow colors). We re-used the control samples for both seronegative and seropositive analyses.

A. Sample collections for seronegative RA analysis

Cohorts	Case	Control
Discovery analysis		
UK	1096	8430
US	551	2134
Dutch	301	2004
Swedish Umea	242	963
Spanish	216	399
Discovery analysis Total	2406	13930
Replication		
Swedish EIRA	427	1691
Discovery+replication Total	2833	15621

B. Sample collections for seropositive RA analysis

Cohorts	Case	Control
UK	2463	8430
US	1803	2134
Dutch	330	2004
Swedish Umea	524	963
Spanish	397	399
Swedish EIRA	1762	1940
Total	7279	15870

Table S2. List of known associated SNPs used for defining risk scores for ACPA+ RA and AS. (A) We used 47 RA associated loci reported in Eyre et al. 2012 to build genetic risk scores (GRS). We estimated the effect sizes of these loci with respect to ACPA+ RA using leave-one-out approach. In two-step approach, we used loci not associated to ACPA- RA (P>0.01, last column) to build GRS. **(B)** We obtained the list of 24 AS associated loci from the Cortes et al. 2013. We used 19 loci that passed QC in our collections (right most column) in addition to HLA-B*27 to build GRS. We used odds ratios reported in Cortes et al. 2013.

A. 47 RA associated loci (Eyre et al. 2012)

SNP	Gene	Chromosome	Position of proxy	Proxy in Immunochip	r^2 to proxy	ACPA+ association P	ACPA- association P	ACPA- P > 0.01?
rs2843401	TNFRSF14	1	2528133	rs2843401	1	6.57E-09	6.02E-01	TRUE
rs2240336	PADI4	1	17674402	rs2240336	1	5.98E-09	2.83E-02	TRUE
rs883220	INPP5B	1	38616871	rs883220	1	1.01E-04	6.66E-02	TRUE
rs2476601	PTPN22	1	114377568	rs2476601	1	6.99E-77	1.74E-04	FALSE
rs11586238	IGSF2	1	117263138	rs11586238	1	3.42E-03	6.83E-01	TRUE
rs2228145	IL6R	1	154426970	rs2228145	1	1.58E-07	2.44E-02	TRUE
rs12746613	FCGR2B	1	161467042	rs12746613	1	6.91E-05	5.01E-02	TRUE
rs10919563	PTPRC	1	198700442	rs10919563	1	2.88E-04	5.90E-01	TRUE
rs34695944	REL	2	61124850	rs34695944	1	2.75E-08	2.89E-01	TRUE
rs1858036	SPRED2	2	65598241	rs1858036	1	1.04E-06	7.81E-01	TRUE
rs11676922	AFF3	2	100806940	rs11676922	1	2.27E-08	1.64E-02	TRUE
rs13426947	STAT4	2	191933254	rs13426947	1	7.44E-09	2.67E-03	FALSE
rs1980422	CD28	2	204610396	rs1980422	1	2.64E-07	4.72E-01	TRUE
rs11571302	ICOS	2	204742934	rs11571302	1	4.48E-08	1.21E-01	TRUE
rs13315591	PXK	3	58555895	rs9813011	1	1.72E-01	5.00E-01	TRUE
rs12506688	RBPJ	4	26104113	rs12506688	1	2.55E-10	4.18E-02	TRUE
rs6822844	IL21	4	123509421	rs6822844	1	4.04E-02	9.90E-02	TRUE
rs71624119	ANKRD55	5	55440730	rs71624119	1	1.20E-11	5.21E-12	FALSE

rs2561477	PAM	5	102608924	rs2561477	1	2.74E-05	6.20E-05	FALSE
rs548234	PRDM1	6	106568034	rs548234	1	1.57E-02	5.18E-01	TRUE
rs10499194	TNFAIP3	6	138002637	rs10499194	1	8.15E-07	3.36E-01	TRUE
rs6920220	TNFAIP3	6	138006504	rs6920220	1	2.30E-13	3.76E-02	TRUE
rs58721818	TNFAIP3	6	138243739	rs58721818	1	5.99E-12	1.14E-01	TRUE
rs212389	TAGAP	6	159489791	rs212389	1	2.95E-06	7.15E-01	TRUE
rs59466457	CCR6	6	167537754	rs59466457	1	2.91E-10	6.41E-01	TRUE
rs3807306	IRF5	7	128580680	rs3807306	1	1.90E-07	2.22E-02	TRUE
rs10488631	IRF5	7	128594183	rs10488631	1	2.02E-03	4.44E-04	FALSE
rs2736340	BLK	8	11343973	rs2736340	1	1.94E-04	4.92E-04	FALSE
rs951005	CCL21	9	34743681	rs951005	1	3.82E-02	3.73E-01	TRUE
rs2269060	TRAF1	9	123683569	rs2269060	1	5.58E-06	9.46E-02	TRUE
rs10795791	IL2RA	10	6108340	rs10795791	1	4.75E-06	6.24E-02	TRUE
rs4750316	PRKCQ	10	6393260	rs4750316	1	4.54E-04	1.00E-01	TRUE
rs2275806	GATA3	10	8095340	rs2275806	1	1.45E-05	3.50E-02	TRUE
rs12764378	ARID5B	10	63800004	rs12764378	1	1.68E-06	6.93E-01	TRUE
rs540386	TRAF6	11	36509189	rs5030485	0.93	4.47E-02	3.16E-01	TRUE
rs595158	CD5	11	60909581	rs595158	1	3.88E-05	4.07E-03	FALSE
rs10892279	DDX6	11	118611781	rs10892279	1	2.13E-06	8.05E-01	TRUE
rs1678542	KIF5A	12	57968715	rs1678542	1	1.04E-03	6.17E-01	TRUE
rs8043085	RASGRP1	15	38828140	rs8043085	1	1.36E-10	3.70E-01	TRUE
rs8026898	TLE3	15	69991417	rs8026898	1	1.27E-10	3.64E-03	FALSE
rs13330176	IRF8	16	86019087	rs13330176	1	3.85E-08	7.76E-01	TRUE
rs2872507	IKZF3	17	38040763	rs2872507	1	1.28E-06	2.15E-01	TRUE
rs34536443	TYK2	19	10463118	rs34536443	1	2.24E-14	1.16E-02	TRUE
rs4810485	CD40	20	44747947	rs4810485	1	1.45E-09	9.19E-01	TRUE

rs2834512	RCAN1	21	35911599	rs2834512	1	2.16E-04	5.82E-01	TRUE
rs9979383	RUNX1	21	36715761	rs9979383	1	3.76E-05	1.02E-04	FALSE
rs3218253	IL2RB	22	37544810	rs3218253	1	2.55E-07	3.16E-01	TRUE

B. 24 AS associated loci (Cortes et al. 2013)

SNP	Gene	Chromosome	Position	Risk allele/non-Risk allele	Odds ratio	QC passed in our dataset
rs11209026	IL23R	1p31	67478546	G/A	1.62	O
rs1801274	FCGR2A	1q23	159746369	T/C	1.11	O
rs4129267	IL6R	1q21	152692888	C/T	1.14	X
rs41299637	GPR25-KIF21B	1q32	199144473	T/G	1.19	O
rs6600247	RUNX3	1p36	25177701	C/T	1.15	O
rs12615545	UBE2E3	2q31	181756697	C/T	1.12	O
rs4676410	GPR35	2q37	241212412	T/C	1.13	X
rs6759298	Intergenic	2p15	62421949	C/G	1.29	O
rs12186979	PTGER4	5p13	40560617	G/A	1.08	O
rs30187	ERAP1	5q15	96150086	T/C	1.29	O
rs6871626	IL12B	5q33	158759370	A/C	1.1	O
rs17765610	BACH2	6q15	90722494	G/A	1.15	O
rs1128905	CARD9	9q34	138373660	C/T	1.1	X
rs11190133	NKX2-3	10q24	101268715	C/T	1.15	O
rs1250550	ZMIZ1	10q22	80730323	G/T	1.11	O
rs11065898	SH2B3	12q24	110346958	T/C	1.11	O
rs1860545	LTBR-TNFRSF1A	12p13	6317038	C/T	1.13	O
rs11624293	GPR65	14q31	87558574	C/T	1.2	O
imm_16_28525386	IL27-SULT1A1	16p11	28525386	A/G	1.11	X
rs2531875	NOS2	17q11	23172294	G/T	1.12	O
rs9901869	NPEPPS-TBKBP1-TBX21	17q21	42930205	A/G	1.14	O
rs35164067	TYK2	19p13	10386181	G/A	1.14	O
rs2836883	Intergenic	21q22	39388614	G/A	1.18	O

rs7282490	ICOSLG	21q22	44440169	G/A	1.11	X

Table S3. Imputation accuracy in current dataset and previous dataset (Raychaudhuri et al. 2012). To measure imputation accuracy, we typed *HLA-A, HLA-B, HLA-C, HLA-DQB1,* and *HLA-DRB1* in 918 individuals in UK cohort. Then we calculated imputation accuracy as the proportion of the alleles correctly imputed (Online Methods). For each gene, we only used individuals whose four-digit typing was successful.

Gene	Two digit alleles			Four digit alleles		
	Accuracy in previous dataset, A_{prev}	Accuracy in current dataset, A_{new}	Error reduction ratio, $(1-A_{prev})/(1-A_{new})$	Accuracy in previous dataset, A_{prev}	Accuracy in current dataset, A_{new}	Error reduction ratio, $(1-A_{prev})/(1-A_{new})$
HLA-A	0.972	0.988	2.43	0.967	0.983	2.00
HLA-B	0.945	0.982	3.07	0.936	0.972	2.30
HLA-C	0.968	0.985	2.07	0.957	0.978	1.92
HLA-DQB1	0.964	0.994	6.20	0.829	0.987	13.32
HLA-DRB1	0.943	0.974	2.16	0.870	0.926	1.76
Average	**0.959**	**0.985**	**2.70**	**0.912**	**0.969**	**2.87**

Table S4: List of binary markers defined within MHC and the association results at these markers.

Described in a separate Excel file.

Table S5. Exhaustive pairwise search results for associations in *HLA-DRB1* and *HLA-B*. In order to find the best pair of markers explaining the associations to seronegative RA, we tested every possible pair of binary markers between *HLA-DRB1* and *HLA-B*. The total number of tests was 383,130 (495 markers in DRB1 × 774 markers in B). We tested each pair by including them in the logistic regression and comparing the deviance to chi-square distribution with 2 degrees of freedom. We show the top 20 pairs. The best pair was Ser11+Leu11 from *HLA-DRB1,* and from *HLA-B*, one of B*0801, B*08, HLA-B Asp9, and rs2596492 at the first base position of the codon at position 9. The four markers at *HLA-B* are in almost perfect LD ($r^2 \geq 0.997$) and statistically indistinguishable.

	HLA-DRB1		***HLA-B***			
Rank	**SNP2HLA ID**	**Marker**	**SNP2HLA ID**	**Marker**	**Deviance**	**P-value**
1	AA_DRB1_11_32660115_SL	Ser11+Leu11	HLA_B_0801	HLA-B*0801	92.255	9.30E-21
2	AA_DRB1_11_32660115_SL	Ser11+Leu11	HLA_B_08	HLA-B*08	92.251	9.30E-21
3	AA_DRB1_11_32660115_SL	Ser11+Leu11	AA_B_9_31432689_D	Asp9	92.007	1.00E-20
4	AA_DRB1_11_32660115_SL	Ser11+Leu11	SNP_B_31432690_C	rs2596492(C_vs_G+T)	92.007	1.00E-20
5	AA_DRB1_11_32660115_SLD	Ser11+Leu11+Asp11	HLA_B_0801	HLA-B*0801	89.967	2.90E-20
6	AA_DRB1_11_32660115_SLD	Ser11+Leu11+Asp11	HLA_B_08	HLA-B*08	89.962	2.90E-20
7	AA_DRB1_13_32660109_SFG	Ser13+Phe13+Gly13	HLA_B_0801	HLA-B*0801	89.921	3.00E-20
8	AA_DRB1_13_32660109_SFG	Ser13+Phe13+Gly13	HLA_B_08	HLA-B*08	89.916	3.00E-20
9	AA_DRB1_11_32660115_SLD	Ser11+Leu11+Asp11	AA_B_9_31432689_D	Asp9	89.713	3.30E-20
10	AA_DRB1_11_32660115_SLD	Ser11+Leu11+Asp11	SNP_B_31432690_C	rs2596492(C_vs_G+T)	89.713	3.30E-20
11	AA_DRB1_13_32660109_SFG	Ser13+Phe13+Gly13	AA_B_9_31432689_D	Asp9	89.663	3.40E-20
12	AA_DRB1_13_32660109_SFG	Ser13+Phe13+Gly13	SNP_B_31432690_C	rs2596492(C_vs_G+T)	89.663	3.40E-20
13	AA_DRB1_11_32660115_SL	Ser11+Leu11	SNP_B_31430769	rs2523607	86.743	1.50E-19
14	AA_DRB1_11_32660115_SL	Ser11+Leu11	SNP_B_31431395	rs2596495	86.688	1.50E-19
15	AA_DRB1_11_32660115_SL	Ser11+Leu11	SNP_B_31431485	rs4990036	86.619	1.60E-19
16	AA_DRB1_11_32660115_SL	Ser11+Leu11	AA_B_97_31432180_SNV	Ser97+Val97+Asn97	84.906	3.70E-19
17	AA_DRB1_11_32660115_SL	Ser11+Leu11	AA_B_97_31432180_SV	Ser97+Val97	84.652	4.10E-19
18	AA_DRB1_11_32660115_SL	Ser11+Leu11	SNP_B_31432582	rs9266178	84.514	4.40E-19
19	AA_DRB1_11_32660115_SL	Ser11+Leu11	SNP_B_31432583	rs9266179	84.514	4.40E-19
20	AA_DRB1_11_32660115_SL	Ser11+Leu11	AA_B_45_31432581_EG	Glu45+Gly45	84.486	4.50E-19

Table S6. RF status-stratified analysis results in the UK cohort. We obtained rheumatoid factor (RF) data for the cases in the UK cohort. We stratified the cases into two groups based on the RF status and examined association results in each group, controlling for heterogeneity due to possible confounding from ACPA+ RA and AS.

CCP and RF status	# Cases	HLA-DRβ1 Ser11+Leu11		HLA-B Asp9		Estimated confounding proportion	
		P-value	OR (CI95)	P-value	OR (CI95)	ACPA+ RA	AS
All CCP-	1096	5.7E-11	1.38 (1.26-1.53)	2.3E-5	1.31 (1.16-1.48)	0.241	0.099
CCP- / RF+	470	6.2E-9	1.53 (1.32-1.77)	8.1E-7	1.56 (1.32-1.85)	0.255	0.090
CCP- / RF-	546	1.8E-4	1.29 (1.13-1.47)	0.08	1.17 (0.99-1.39)	0.233	0.118

Table S7. Forward conditional haplotype analysis on individual HLA-DRβ1 amino acid residues for ACPA+ RA. For each amino acid position in HLA-DRβ1 (column 1), we partitioned the classical alleles into groups based on the amino acid residues and performed omnibus association testing where the degree of freedom (**df**) is the number of partitions minus one. We included the signal peptide in the test (negative positions). If multiple amino acid positions are statistically the same (give the exactly same partitioning all the time), we only kept the position with the lowest position number. The most significant amino acid positions were 11 and 13 (highlighted), which were statistically indistinguishable ($P > 0.03$). Then we performed conditional analysis where given the partitioning defined on position 11, if further partitioning by additional amino acid gives significant p-value. Conditioning on 11, we found 71 is significant (highlighted), and conditioning on 11 and 71, we found 74 was significant (highlighted). Conditioning on 11, 71, and 74, the most significant was position 70 (highlighted).

Condition:	**On Nothing**			**On Position 11**			**On Positions 11 and 71**			**On positions 11, 71 and 74**		
Amino acid position	**df**	χ^2	$\log_{10}P$	**df**	χ^2	$\log_{10}P$	**df**	χ^2	$\log_{10}P$	**df**	χ^2	$\log_{10}P$
-29	1	124.26	-28.13	1	0.84	-0.44	1	3.30	-1.16	1	10.34	-2.89
-25	2	1274.28	-276.71	2	5.31	-1.15	2	15.67	-3.40	2	19.79	-4.30
-24	2	2669.56	-579.69	2	5.31	-1.15	2	15.67	-3.40	2	19.79	-4.30
-17	2	149.22	-32.40	1	0.84	-0.44	1	3.30	-1.16	1	10.34	-2.89
-16	2	1274.28	-276.71	2	5.31	-1.15	2	15.67	-3.40	2	19.79	-4.30
-1	2	355.64	-77.23	2	15.15	-3.29	1	3.30	-1.16	1	10.34	-2.89
1	1	1.64	-0.70	1	5.12	-1.63	1	4.95	-1.58	1	4.14	-1.38
4	2	251.49	-54.61	1	5.12	-1.63	1	4.95	-1.58	1	4.14	-1.38
9	2	207.59	-45.08	0	0.00	0.00	0	0.00	0.00	0	0.00	0.00
10	2	1572.99	-341.57	1	4.48	-1.47	1	12.67	-3.43	1	8.35	-2.41
11	5	3551.51	-766.45	0	0.00	0.00	0	0.00	0.00	0	0.00	0.00
12	1	1504.66	-328.42	0	0.00	0.00	0	0.00	0.00	0	0.00	0.00
13	5	3489.63	-753.02	2	7.29	-1.58	2	15.01	-3.26	2	9.65	-2.10
14	1	313.74	-69.48	0	0.00	0.00	0	0.00	0.00	0	0.00	0.00
16	1	111.92	-25.43	1	2.81	-1.03	1	1.68	-0.71	1	1.17	-0.55
26	2	224.95	-48.85	2	75.19	-16.33	2	24.89	-5.41	2	10.57	-2.29

28	2	15.31	-3.33	1	15.64	-4.12	1	23.79	-5.97	1	9.60	-2.71
30	5	498.85	-104.85	2	14.94	-3.24	2	22.79	-4.95	2	10.28	-2.23
31	2	218.07	-47.35	1	4.48	-1.47	1	12.67	-3.43	1	8.35	-2.41
32	1	708.24	-155.32	1	4.62	-1.50	1	11.02	-3.04	1	56.73	-13.30
33	1	2639.72	-575.02	1	4.48	-1.47	1	12.67	-3.43	1	8.35	-2.41
37	4	1864.09	-401.81	4	11.68	-1.70	4	12.36	-1.83	4	73.74	-14.43
38	2	144.08	-31.29	2	14.13	-3.07	2	21.95	-4.77	2	9.88	-2.15
40	1	125.01	-28.30	1	4.48	-1.47	1	12.67	-3.43	1	8.35	-2.41
47	1	1642.11	-358.28	1	1.99	-0.80	1	2.60	-0.97	1	9.29	-2.64
57	3	425.20	-91.11	3	10.40	-1.81	3	9.87	-1.71	3	43.16	-8.64
58	1	241.01	-53.63	1	2.57	-0.96	1	1.47	-0.65	1	33.10	-8.06
60	2	374.12	-81.24	2	10.40	-2.26	2	9.56	-2.08	2	41.65	-9.04
67	2	2107.81	-457.70	2	143.52	-31.16	2	2.11	-0.46	2	84.35	-18.32
70	2	1610.74	-349.77	2	85.29	-18.52	2	2.33	-0.51	2	93.66	-20.34
71	3	1279.84	-276.46	3	222.78	-47.30	0	0.00	0.00	0	0.00	0.00
73	1	574.88	-126.31	1	52.92	-12.46	1	4.10	-1.37	0	0.00	0.00
74	4	782.79	-167.39	3	139.68	-29.35	3	97.96	-20.37	0	0.00	0.00
77	1	203.84	-45.52	1	52.92	-12.46	1	4.10	-1.37	0	0.00	0.00
78	1	250.09	-55.61	0	0.00	0.00	0	0.00	0.00	0	0.00	0.00
85	1	17.88	-4.63	1	0.05	-0.08	1	0.06	-0.09	1	2.89	-1.05
86	1	605.20	-132.91	1	35.85	-8.67	1	18.20	-4.70	1	17.32	-4.50
95	1	2.93	-1.06	1	4.81	-1.55	1	4.76	-1.54	1	4.84	-1.56
96	4	3272.65	-707.43	2	9.29	-2.02	2	17.47	-3.79	2	13.18	-2.86
98	2	1381.39	-299.97	2	9.29	-2.02	2	17.47	-3.79	2	13.18	-2.86
102	1	0.03	-0.07	1	0.25	-0.21	1	0.59	-0.36	1	0.58	-0.35
104	2	1377.57	-299.14	2	4.74	-1.03	2	13.28	-2.88	2	8.93	-1.94
112	2	64.54	-14.01	2	0.51	-0.11	2	1.22	-0.26	2	39.31	-8.54

120	2	2858.50	-620.72	1	0.25	-0.21	1	0.59	-0.36	1	0.58	-0.35
133	2	192.39	-41.78	1	0.25	-0.21	1	0.59	-0.36	1	0.58	-0.35
140	2	190.04	-41.27	1	0.25	-0.21	1	0.59	-0.36	1	0.58	-0.35
142	2	192.39	-41.78	1	0.25	-0.21	1	0.59	-0.36	1	0.58	-0.35
149	2	1504.92	-326.79	1	0.25	-0.21	1	0.59	-0.36	1	0.58	-0.35
166	2	125.04	-27.15	2	4.74	-1.03	2	13.28	-2.88	2	8.93	-1.94
180	2	2658.12	-577.20	2	4.79	-1.04	2	13.21	-2.87	2	10.53	-2.29
181	2	220.16	-47.81	2	4.79	-1.04	2	13.21	-2.87	2	10.53	-2.29
182	1	80.78	-18.60	1	0.31	-0.24	1	0.77	-0.42	1	2.18	-0.85
188	1	105.05	-23.92	1	0.12	-0.14	1	2.80	-1.03	1	9.62	-2.72
189	2	105.46	-22.90	2	3.19	-0.69	2	5.08	-1.10	2	8.98	-1.95
190	1	104.05	-23.70	1	0.07	-0.10	1	2.49	-0.94	1	8.74	-2.51
231	2	224.61	-48.77	2	4.55	-0.99	2	14.90	-3.24	2	18.19	-3.95
233	2	1502.31	-326.22	2	3.19	-0.69	2	5.08	-1.10	2	8.98	-1.95
234	1	103.53	-23.59	1	0.04	-0.07	1	2.29	-0.89	1	8.30	-2.40
236	1	99.34	-22.67	1	0.01	-0.04	1	1.46	-0.64	1	5.74	-1.78

Table S8. Effect estimates for the amino acids associated with risk of ACPA- and ACPA+ rheumatoid arthritis.
(A) Effect estimates for ACPA- RA. We estimated the effect size of HLA-DRβ1 Ser11+Leu11 taking Val11+Asp11+Pro11+Gly11 as reference and HLA-B Asp9 taking His9+Tyr9 as reference. The effect size at HLA-DRβ1 Ser11+Leu11 is conditioned on HLA-B Asp9, and the effect size at HLA-B Asp9 is conditioned on HLA-DRβ1 Ser11+Leu11. We also show the unadjusted case/control allele frequencies and the classical alleles of *HLA-DRB1* and *HLA-B* corresponding to the amino acids. **(**B) Effect estimates for ACPA+ RA. For HLA-DRβ1, We defined haplotypes based on the amino acid residues present at position 11, 13, 71, and 74. For each haplotype, the multivariate effect is given as an odds ratio (OR), taking the most frequent haplotype (ProArgAlaAla) in the control samples as the reference (that is, giving that haplotype an OR of 1). The effects are conditioned on the remaining associated loci: position 9 in HLA-B, position 9 in HLA-DPβ1, and position 77 in HLA-A. We show the unadjusted allele frequencies and the classical alleles corresponding to each haplotype. We also list the effect sizes, allele frequencies and classical alleles corresponding to HLA-B Asp9, HLA-DPβ1 Phe9, and HLA-A Asn77. The effects of each of these positions were estimated conditioned on the remaining loci; e.g. the effects in HLA-B were conditioned on *HLA-DRB1* alleles, position 9 in HLA-DPβ1, and position 77 in HLA-A. 95% CI, 95% confidence interval.

A. ACPA- RA					
HLA-DRβ1 amino acid at position 11	OR	95% CI	Unadjusted allele frequencies		Classical *HLA-DRB1* alleles
			Controls	Cases	
Ser+Leu	1.23	1.15-1.32	0.514	0.548	**01, *03, *08, *11, *12, *13, *14*
Val+Asp+Pro+Gly	Reference		0.486	0.452	**04, *07, *09, *10, *15, *16*
HLA-B amino acid at position 9					Classical *HLA-B* alleles
Asp	1.24	1.14-1.36	0.131	0.161	**08*
His, Tyr	Reference		0.869	0.839	**07, *13, *14, *15, *18, *27, *35, *37, *38, *39, *40, *41, *42, *44, *45, *46, *47, *48, *49, *50, *51, *52, *53, *54, *55, *56, *57, *58, *73, *81*

B. ACPA+ RA								
HLA-DRβ1 amino acid at position				Multivariate OR	95% CI	Unadjusted allele frequencies		Classical *HLA-DRB1* alleles
11	13	71	74			Controls	Cases	
Val	Phe	Arg	Ala	4.65	3.80-5.70	0.007	0.021	**10:01*
Val	His	Lys	Ala	4.03	3.72-4.37	0.110	0.292	**04:01, *04:09*
Val	His	Arg	Ala	3.63	3.29-4.01	0.054	0.123	**04:04, *04:05, *04:08, *04:10*
Leu	Phe	Arg	Ala	2.11	1.94-2.31	0.104	0.146	**01:01, *01:02*
Asp	Phe	Arg	Glu	1.82	1.52-2.18	0.013	0.017	**09:01*
Pro	Arg	Arg	Ala	1.58	1.26-1.99	0.009	0.010	**16:01, *16:02*
Val	His	Arg	Glu	1.29	1.06-1.57	0.016	0.012	**04:03, *04:06, *04:07, *04:11*
Ser	Gly	Arg	Ala	1.04	0.86-1.25	0.018	0.013	**12:01, *12:02*
Val	His	Glu	Ala	1.03	0.71-1.50	0.005	0.003	**04:02, *04:37*
Pro	Arg	Ala	Ala	1.00	Reference	0.143	0.094	**15:01, *15:02, *15:03*
Gly	Tyr	Arg	Gln	0.92	0.83-1.02	0.127	0.067	**07:01*
Ser	Ser	Lys	Ala	0.87	0.66-1.14	0.009	0.005	**13:03*

Ser	Gly	Arg	Leu	0.83	0.70-0.98	0.027	0.016	**08:01, *08:02, *08:03, *08:04, *08:06, *14:15*
Ser	Ser	Arg	Glu	0.77	0.64-0.94	0.023	0.011	**14:01, *14:05, *14:07*
Ser	Ser	Arg	Ala	0.76	0.67-0.86	0.067	0.034	**11:01, *11:04, *11:06, *11:08, *13:05, *14:02, *14:06*
Leu	Phe	Glu	Ala	0.71	0.55-0.93	0.012	0.005	**01:03*
Ser	Ser	Lys	Arg	0.67	0.60-0.76	0.127	0.081	**03:01, *03:02, *03:04*
Ser	Ser	Glu	Ala	0.60	0.54-0.67	0.115	0.046	**11:02, *11:03, *13:01, *13:02, *13:04*
Ser	Gly	Arg	Glu	0.49	0.26-0.91	0.003	0.001	**14:04*
HLA-B amino acid at position 9				Classical *HLA-B* alleles				
Asp				2.13	1.91-2.37	0.130	0.118	**08*
His, Tyr				1.00	Reference	0.870	0.882	**07, *13, *14, *15, *18, *27, *35, *37, *38, *39, *40, *41, *42, *44, *45, *46, *47, *48, *49, *50, *51, *52, *53, *54, *55, *56, *57, *58, *73, *81*
HLA-DPβ1 amino acid at position 9				Classical *HLA-DPB1* alleles				
Phe				1.31	1.24-1.39	0.721	0.793	**02, *04, *05, *16, *19, *23, *34*
His, Tyr				1.00	Reference	0.279	0.207	**01, *03, *06, *09, *10, *11, *13, *14, *15, *17, *18, *20, *21, *26, *30, *35*
HLA-A amino acid at position 77				Classical *HLA-A* alleles				
Asn				0.85	0.81-0.90	0.343	0.279	**01, *23, *24, *26, *29, *30, *36, *80*
Asp, Ser				1.00	Reference	0.657	0.721	**02, *03, *11, *25, *30, *31, *32, *33, *34, *66, *68, *69, *74*

Table S9. Frequency of ancestral haplotype in six cohorts. We calculated the frequency of ancestral 8.1 haplotype using the best guess imputation data that was phased across the MHC region.

Cohort	Ancestral Haplotype Frequency
UK	0.131
US	0.106
Dutch	0.165
Swedish Umea	0.095
Spanish	0.052
Swedish EIRA	0.136

Table S10. Estimated proportions of confounding diseases in seronegative RA dataset. Using risk scores built from the known associated loci to ACPA+ RA and ankylosing spondylitis (AS), we estimated the proportion of ACPA+ RA and AS samples within ACPA- (**Online Methods**). We applied two different approaches; (A) We used logistic regression that includes the ACPA+ risk score and AS risk score, and no candidate associations to ACPA- RA. (B) We used logistic regression that includes not only the ACPA+ risk score and AS risk score but also the two variables that are putatively associated to ACPA- RA, Ser+Leu-11 of HLA-DRβ1 and Asp-9 of HLA-B. 95% C.I., 95% confidence interval.

Cohort	**(A) Regression using risk scores only**				**(B) Regression using risk scores and putative associations in ACPA- RA**			
	ACPA+	95% C.I.	AS	95% C.I.	ACPA+	95% C.I.	AS	95% C.I.
UK	0.241	0.185-0.296	0.099	0.055-0.142	0.282	0.223-0.340	0.100	0.056-0.144
US	0.366	0.279-0.452	0.041	-0.032-0.115	0.409	0.320-0.498	0.048	-0.025-0.122
Dutch	0.152	0.034-0.270	0.082	-0.017-0.181	0.184	0.063-0.304	0.088	-0.011-0.188
Swedish Umea	0.199	0.073-0.326	0.108	0.030-0.185	0.256	0.127-0.384	0.112	0.034-0.190
Spanish	0.340	0.170-0.510	0.079	-0.066-0.225	0.381	0.210-0.552	0.073	-0.073-0.220