nature genetics

# Five amino acids in three HLA proteins explain most of the association between MHC and seropositive rheumatoid arthritis

Soumya Raychaudhuri[1–4], Cynthia Sandor[1–4], Eli A Stahl[1,2,4], Jan Freudenberg[5], Hye-Soon Lee[6], Xiaoming Jia[1,4,7], Lars Alfredsson[8], Leonid Padyukov[9], Lars Klareskog[9], Jane Worthington[10], Katherine A Siminovitch[11], Sang-Cheol Bae[6], Robert M Plenge[1,2,4], Peter K Gregersen[5] & Paul I W de Bakker[1,4,12,13]



**The genetic association of the major histocompatibility complex (MHC) to rheumatoid arthritis risk has commonly been attributed to alleles in *HLA-DRB1*. However, debate persists about the identity of the causal variants in *HLA-DRB1* and the presence of independent effects elsewhere in the MHC. Using existing genome-wide SNP data in 5,018 individuals with seropositive rheumatoid arthritis (cases) and 14,974 unaffected controls, we imputed and tested classical alleles and amino acid polymorphisms in *HLA-A*, *HLA-B*, *HLA-C*, *HLA-DPA1*, *HLA-DPB1*, *HLA-DQA1*, *HLA-DQB1* and *HLA-DRB1*, as well as 3,117 SNPs across the MHC. Conditional and haplotype analyses identified that three amino acid positions (11, 71 and 74) in HLA-DRβ1 and single–amino-acid polymorphisms in HLA-B (at position 9) and HLA-DPβ1 (at position 9), which are all located in peptide-binding grooves, almost completely explain the MHC association to rheumatoid arthritis risk. This study shows how imputation of functional variation from large reference panels can help fine map association signals in the MHC.**

Rheumatoid arthritis is a systemic autoimmune disease characterized by intra-articular inflammation[1]. About 70% of affected individuals have antibodies against cyclic citrullinated peptide (anti-CCP–positive rheumatoid arthritis)[2]. Previously, the strong association of the MHC to anti-CCP–positive rheumatoid arthritis[3,4] was explained by the presence of consensus amino acid sequences (QRRAA, RRRAA and QKRAA) spanning positions 70–74 in the β1 subunit of the HLA-DR molecule. The classical haplotypes encoding these sequences in the corresponding *HLA-DRB1* gene define the 'shared epitope' alleles[5]. The shared epitope association was historically defined by exploring structural differences between *HLA-DRB1*04* alleles using immunological reagents that leveraged allospecific T cell recognition[6,7]. These reagents focused attention on sequence determinants on the exposed α-helical rim of the HLA-DR molecule, where the shared epitope is located, but left allelic differences at the inaccessible base of the binding groove largely unexplored.

Despite serving as the foundation for genetic studies of rheumatoid arthritis, the shared epitope hypothesis does not fully explain the association at *HLA-DRB1*; studies have suggested additional independent associations to rheumatoid arthritis within the MHC in addition to that at *HLA-DRB1* (refs. 3,8–11). However, pinpointing the associated loci has been challenging, in part because of the complexity and cost of complete HLA genotyping and the broad linkage disequilibrium (LD) across the MHC[12].

To define the association across the region and identify functional and potentially causal variants, we obtained SNP genotype data for a total of 19,992 individuals from six independent genome-wide datasets (**Supplementary Table 1**)[13], including 5,018 cases with anti-CCP–positive rheumatoid arthritis and 14,974 controls of European descent. We used a large reference panel of 2,767 individuals of European descent[14] to impute classical allele genotypes for *HLA-A*, *HLA-B*, *HLA-C*, *HLA-DPA1*, *HLA-DPB1*, *HLA-DQA1*, *HLA-DQB1* and *HLA-DRB1*, their corresponding amino acid sequences and SNPs within the MHC[15]. In total, we tested 99 classical *HLA* alleles at two-digit resolution, 164 classical *HLA* alleles four-digit resolution, 372 polymorphic amino acid positions and 3,117 SNPs across the region for association with logistic regression. To control for population stratification, we included as covariates the first five principal components

[1]Division of Genetics, Brigham and Women's Hospital, Harvard Medical School, Boston, Massachusetts, USA. [2]Division of Rheumatology, Brigham and Women's Hospital, Harvard Medical School, Boston, Massachusetts, USA. [3]Partners HealthCare Center for Personalized Genetic Medicine, Boston, Massachusetts, USA. [4]Program in Medical and Population Genetics, Broad Institute of Harvard and Massachusetts Institute of Technology, Cambridge, Massachusetts, USA. [5]Robert S. Boas Center for Genomics and Human Genetics, The Feinstein Institute for Medical Research, North Shore–Long Island Jewish Health System, Manhasset, New York, USA. [6]Department of Rheumatology, Hanyang University Hospital for Rheumatic Diseases, Seoul, South Korea. [7]Harvard–Massachusetts Institute of Technology Division of Health Sciences and Technology, Boston, Massachusetts, USA. [8]Institute of Environmental Medicine, Karolinska Institutet, Stockholm, Sweden. [9]Rheumatology Unit, Department of Medicine, Karolinska Institutet at Karolinska University Hospital Solna, Stockholm, Sweden. [10]Arthritis Research UK Epidemiology Unit, Manchester Academic Health Sciences Centre, University of Manchester, Manchester, UK. [11]Department of Medicine, University of Toronto, Mount Sinai Hospital and University Health Network, Toronto, Ontario, Canada. [12]Julius Center for Health Sciences and Primary Care, University Medical Center Utrecht, Utrecht, The Netherlands. [13]Department of Medical Genetics, University Medical Center Utrecht, Utrecht, The Netherlands. Correspondence should be addressed to S.R. (soumya@broadinstitute.org) or P.I.W.d.B. (pdebakker@rics.bwh.harvard.edu).

Received 30 September 2011; accepted 12 December 2011; published online 29 January 2012; doi:10.1038/ng.1076

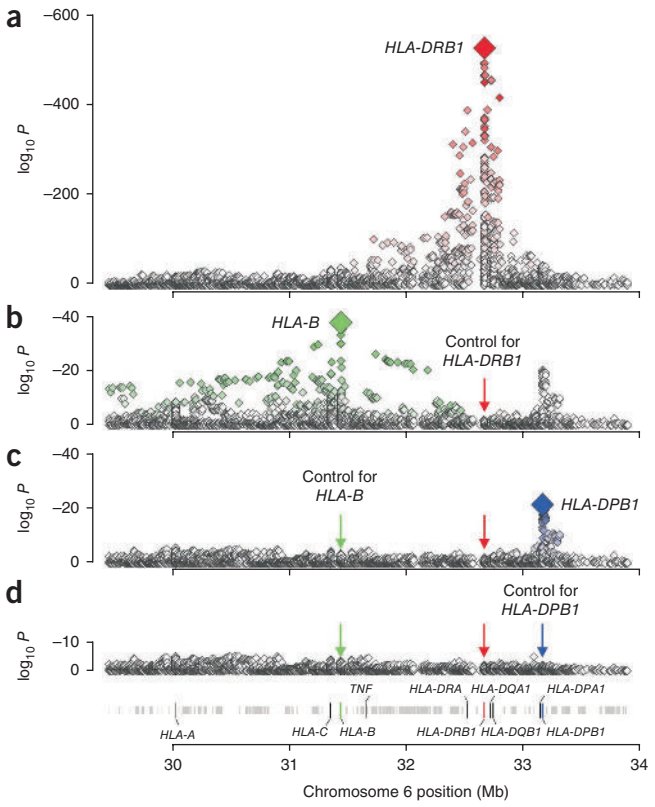


**Figure 1** Association tests within the MHC to rheumatoid arthritis. (**a**) The major genetic determinants of rheumatoid arthritis risk map to *HLA-DRB1*. (**b**) Subsequent conditional analyses controlling for all classical *HLA-DRB1* alleles revealed an independent association at *HLA-B* corresponding to the *HLA-B*08* allele, or Asp9 in the corresponding protein. (**c**) Subsequent analyses that conditioned on *HLA-DRB1* alleles and *HLA-B*08* revealed an independent association for the HLA-DPβ1 Phe9 variant. (**d**) After controlling for *HLA-DRB1*, HLA-B Asp9 and HLA-DPβ1 Phe9, no significant association signal was observed.

from the genome-wide SNP genotypes for each of the six datasets[16] (genomic control inflation factor ($\lambda_{gc}$) = 1.06; **Supplementary Note**).

First, to assess imputation accuracy, we compared imputed *HLA-DRB1* classical alleles to the genotyped alleles in a subset of 1,403 individuals from two datasets genotyped to four-digit resolution (**Supplementary Table 2a**). The imputations were 95.8% accurate for alleles at a two-digit resolution and 84.0% accurate at a four-digit resolution (**Supplementary Note**). We observed high accuracy in the frequency estimates and the imputation quality for alleles with >2.5% frequency in the reference set (**Supplementary Fig. 1a**). We observed similar accuracies at four other classical loci in a subset of samples from the 1958 British Birth Cohort that were part of the Wellcome Trust Case Control Consortium (WTCCC) control group (**Supplementary Table 2b,c**). We note that the WTCCC samples have the sparsest SNP coverage across the MHC of all the samples tested and that these accuracies probably represent a lower bound (**Supplementary Fig. 1b**).

Next, we compared the allelic odds ratios of the imputed *HLA-DRB1* haplotypes in our data with recently reported allelic odds ratios for *HLA-DRB1* haplotypes in a large study of anti-CCP–positive rheumatoid arthritis[17]. Except for the combination of the rare *HLA-DRB1*11:02* and *HLA-DRB1*11:03* haplotypes (which has a frequency of <1%), the effect sizes for each of the *HLA-DRB1* classical haplotypes from our study were entirely consistent with the recently reported results (**Supplementary Fig. 2** and **Supplementary Table 3**).

Having shown the validity of our analytic approach, we next tested SNPs and *HLA* alleles across the MHC for association to rheumatoid arthritis. The most significant allele was the A nucleotide at rs17878703, a quadrallelic SNP in the second nucleotide of *HLA-DRB1* codon 11 (odds ratio (OR) = 3.7, $P < 10^{-526}$; **Fig. 1** and **Supplementary Table 4**). This allele codes for Val11 or Leu11 in HLA-DRβ1. Thus, the strongest MHC signal mapped to amino acid 11 of HLA-DRβ1 and not to any of the shared epitope positions (amino acids 70–74).

We then tested each of the amino acid positions within HLA-DRβ1 for association by grouping classical *HLA-DRB1* haplotypes according to the specific amino acid carried at each position (**Supplementary Table 5**). Amino acid position 11 showed the strongest association ($P < 10^{-581}$; **Fig. 2**). Of the six possible amino acids at this position, the aliphatic residues Val11 (OR = 3.8) and Leu11 (OR = 1.3) conferred a high risk of rheumatoid arthritis, whereas other residues at this position conferred less risk of disease (**Fig. 3** and **Supplementary Table 4**). In fact, the polar Ser11 residue is highly protective against disease (OR = 0.38). Amino acid position 13 showed a similarly statistically significant association ($P < 10^{-574}$); the six alleles at this position are in tight LD with those at position 11. Conditioning on position 11 eliminated the effect of position 13 ($P = 0.57$), but conditioning on position 13 did not eliminate the effect of position 11 ($P = 3.5 \times 10^{-8}$). Although these results favor the role of position 11 over position 13 as causing the association of *HLA-DRB1* to rheumatoid arthritis, the tight LD between the two positions makes it difficult to unambiguously assign causality to one position at the exclusion of the other (**Table 1**). After conditioning on the shared epitope status, amino acid positions 11 and 13 both remained highly significantly associated with disease ($P < 10^{-70}$ and $P < 10^{-63}$, respectively) and were more strongly associated than all of the other polymorphic HLA-DRβ1 amino acid positions.

To replicate these HLA-DRβ1 effects without imputed genotypes, we analyzed an independent South Korean dataset of 616 cases with anti-CCP–positive rheumatoid arthritis and 675 controls using genome-wide SNP data[18] and sequencing-based classical *HLA-DRB1* genotypes at four-digit resolution[19]. We used the first five principal components as covariates to correct for population stratification ($\lambda_{gc} = 1.01$). Of all the amino acids tested in HLA-DRβ1, the strongest associations mapped to amino acid positions 11 ($P = 6.1 \times 10^{-36}$) and 13 ($P = 3.1 \times 10^{-36}$) (**Supplementary Tables 3** and **6**). As both positions are in tight LD, the effects are statistically indistinguishable; adding one of these two positions to a model including the other does not result in a significantly improved fit ($P > 0.08$). Thus, amino acids 11 and 13 in HLA-DRβ1 had the strongest associations with rheumatoid arthritis in two different continental populations.

Given the polymorphic nature of *HLA-DRB1*, we evaluated whether a similarly significant result could emerge by chance by 'tagging' classical alleles of differential risk. To test this possibility, we preserved the classical HLA genotypes and case-control status in all samples and permuted the amino acid sequence defined by each classical *HLA-DRB1* allele 10,000 times. We found that a single amino acid position only rarely resulted in a better goodness of fit for the model (measured by the deviance) as compared to that produced by amino acid position 11 in the actual data ($P = 0.0002$; **Fig. 2b**). Therefore, the degree to which the six alleles at amino acid position 11 divide the classical alleles of *HLA-DRB1* into differential risk groups is extremely unlikely to occur by chance.

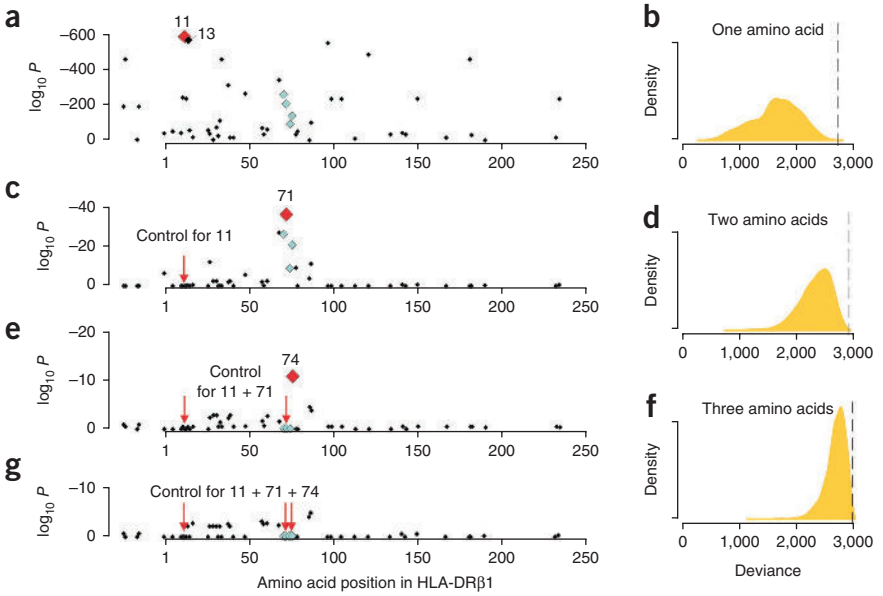


**Figure 2** Association results for amino acids in HLA-DRβ1. (**a**) Amino acid position 11 showed the strongest association with rheumatoid arthritis ($P < 10^{-581}$), followed by position 13 ($P < 10^{-574}$). Shared epitope positions (70–74) are indicated with light-blue diamonds. (**b**) Distribution of deviance in 10,000 permutations of amino acid sequences across classical *HLA-DRB1* alleles, where deviance is calculated as –2 times the log-likelihood for the best amino acid position. The vertical dashed line indicates the deviance for position 11 in the actual data ($P = 0.0002$). (**c**) Controlling for position 11, position 71 was significantly associated with rheumatoid arthritis ($P = 5.6 \times 10^{-38}$). (**d**) Deviance of the best two amino acid positions in 10,000 permutations. The vertical dashed line indicates the deviance for positions 11 and 71 in the actual data ($P = 0.0002$). (**e**) Controlling for positions 11 and 71, position 74 was significantly associated with rheumatoid arthritis ($P = 1.5 \times 10^{-11}$). (**f**) Deviance of the best three amino acid positions in 10,000 permutations. The vertical dashed line indicates the deviance for positions 11, 71 and 74 in the actual data ($P = 0.004$). (**g**) After controlling for positions 11, 71 and 74, no amino acid position was significant ($P > 8 \times 10^{-4}$).

After accounting for the effects of amino acid 11 in HLA-DRβ1 using a conditional haplotype analysis, we observed an independent association at position 71 ($P < 10^{-37}$; **Fig. 2c** and **Supplementary Table 5a**). We tested all possible pairs of polymorphic amino acid positions in HLA-DRβ1; of the 1,275 pairs of amino acid positions tested, none achieved a better goodness of fit than the position 11 and 74 pair ($P = 4 \times 10^{-615}$). Using the same permutation strategy described above, we found that the degree to which amino acid positions 11 and 71 divide the classical alleles of *HLA-DRB1* into differential risk groups is unlikely to occur by chance ($P = 0.0002$) (**Fig. 2d**). At HLA-DRβ1 position 71, the positively charged Lys71 and Arg71 residues confer greater odds of disease (OR = 2.0 and OR = 0.97, respectively) than the small aliphatic Ala71 residue (OR = 0.59); the negatively charged Glu71 confers the least odds of disease of all the four residues at this position (OR = 0.32; **Fig. 3**).

Conditioning on positions 11 and 71 revealed an additional association at position 74 ($P = 1.5 \times 10^{-11}$; **Fig. 2e** and **Supplementary Table 5a**). When we tested all possible combinations of three amino acid positions in HLA-DRβ1, we found that only one combination of amino acids sites (37, 67 and 74; $P = 2 \times 10^{-624}$) out of the 20,825 combinations tested outperformed the combination of sites 11, 71 and 74 ($P = 1.6 \times 10^{-622}$). However, even that combination did not outperform the 11, 71 and 74 combination by a statistically significant margin ($P > 0.01$). As before, we permuted the amino acid sequences, and only rarely were we able to pick three amino acid positions that obtained a better goodness of fit in the permuted data than positions 11, 71 and 74 did in the actual data ($P = 0.004$; **Fig. 2f**). The addition of each of these three amino acid positions to the model yielded an improved model fit, even after accounting for the increased number of parameters with each addition (**Supplementary Table 5b**). We observed no residual association at other HLA-DRβ1 amino acids after conditioning on positions 11, 71 and 74 ($P > 8 \times 10^{-4}$; **Fig. 2g** and **Supplementary Table 5a**).

The amino acids at positions 11, 71 and 74 in HLA-DRβ1 define 16 haplotypes (**Table 1**). The individual disease risk predicted by a full model in which each classical *HLA-DRB1* allele confers its own unique risk and that predicted by a simpler model where risk is defined by amino acid positions 11, 71 and 74 are nearly perfectly correlated ($r = 0.994$). Hence, the model based on the amino acid residues at positions 11, 71 and 74 provides a parsimonious explanation for the effects of the classical *HLA-DRB1* haplotypes and suggests a key role for these amino acids in the function of HLA-DRβ1 in rheumatoid arthritis etiology. This is underscored by the central location of these positions in the peptide-binding groove of the HLA-DR structure (**Fig. 4**). Positions 11 and 13 are located on the β-sheet floor with their side chains oriented into the peptide-binding groove. Positions 71 and 74 are separated by a single turn along the α helix, and their side chains are spatially close to those of positions 11 and 13.

To assess whether there were other independent MHC associations outside of *HLA-DRB1*, we conditioned on HLA-DRβ1 amino acids 11, 71 and 74 and tested all MHC SNPs and *HLA* alleles. We observed the most significant association at *HLA-B* in the class I MHC region ($P < 2 \times 10^{-37}$; **Fig. 1b**). This association maps to Asp9 in HLA-B (OR = 2.12 relative to His9 or Tyr9; **Table 1**, **Fig. 3** and **Supplementary Table 4**), although we could not statistically distinguish this effect from that

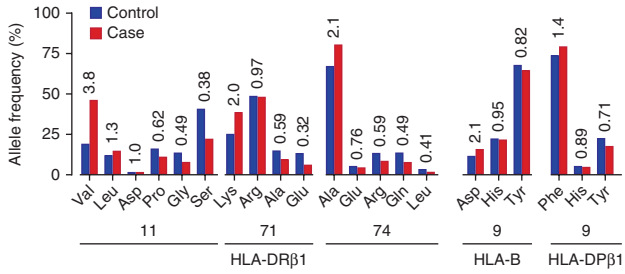


**Figure 3** Effects of individual amino acids within HLA proteins. For amino acid positions 11, 71 and 74 in HLA-DRβ1, position 9 in HLA-B and position 9 in HLA-DPβ1, the allele frequencies in cases (red) and controls (blue) are plotted, and the univariate odds ratios are listed. The effects of HLA-B and HLA-DPβ1 are adjusted for the classical *HLA-DRB1* alleles.

**Table 1 Effect estimates for the five amino acids associated with risk of rheumatoid arthritis**

| HLA-DRβ1 amino acid at position | | | | Multivariate | | Unadjusted allele frequency | | |
|---|---|---|---|---|---|---|---|---|
| 11 | 13 | 71 | 74 | OR | 95% CI | Controls | Cases | Classical *HLA-DRB1* alleles |
| Val | His | Lys | Ala | 4.44 | 4.02–4.91 | 0.106 | 0.316 | ***04:01*** |
| Val | His or Phe | Arg | Ala | 4.22 | 3.75–4.75 | 0.056 | 0.141 | ***04:08, *04:05, *04:04, *10:01*** |
| Leu | Phe | Arg | Ala | 2.17 | 1.94–2.42 | 0.109 | 0.143 | ***01:02, *01:01*** |
| Pro | Arg | Arg | Ala | 2.04 | 1.59–2.62 | 0.013 | 0.012 | *16:01 |
| Val | His | Arg | Glu | 1.65 | 1.24–2.19 | 0.010 | 0.009 | *04:03, *04:07 |
| Asp | Phe | Arg | Glu | 1.65 | 1.29–2.10 | 0.011 | 0.013 | *09:01 |
| Val | His | Glu | Ala | 1.43 | 1.04–1.96 | 0.011 | 0.006 | *04:02 |
| Ser | Ser | Lys | Ala | 1.04 | 0.76–1.41 | 0.012 | 0.006 | *13:03 |
| Pro | Arg | Ala | Ala | 1.00 | Reference | 0.142 | 0.092 | *15:01, *15:02 |
| Gly | Tyr | Arg | Gln | 0.91 | 0.80–1.03 | 0.133 | 0.064 | *07:01 |
| Ser | Ser or Gly | Arg | Ala | 0.88 | 0.77–1.00 | 0.103 | 0.049 | *11:01, *11:04, *12:01 |
| Ser | Ser | Arg | Glu | 0.84 | 0.67–1.05 | 0.025 | 0.012 | *14:01 |
| Leu | Phe | Glu | Ala | 0.73 | 0.42–1.27 | 0.004 | 0.002 | *01:03 |
| Ser | Gly | Arg | Leu | 0.71 | 0.57–0.89 | 0.028 | 0.013 | *08:01, *08:04 |
| Ser | Ser | Lys | Arg | 0.63 | 0.54–0.73 | 0.128 | 0.083 | *03:01 |
| Ser | Ser | Glu | Ala | 0.59 | 0.51–0.68 | 0.112 | 0.041 | *11:02, *11:03, *13:01, *13:02 |
| HLA-B amino acid at position 9 | | | | | | | | Classical *HLA-B* allele |
| Asp | | | | 2.12 | 1.89–2.38 | 0.118 | 0.130 | *08 |
| His, Tyr | | | | 1.00 | Reference | 0.882 | 0.870 | *07, *13, *14, *15, *18, *27, *35, *37, *38, *39, *40, *41, *44, *45, *47, *49, *50, *51, *52, *53, *55, *56, *57, *58, *73 |
| HLA-DPβ1 amino acid at position 9 | | | | | | | | Classical *HLA-DPB1* alleles |
| Phe | | | | 1.40 | 1.31–1.50 | 0.728 | 0.799 | *02:01, *02:02, *04:01, *04:02, *05:01, *16:01, *19:01, *23:01 |
| His, Tyr | | | | 1.00 | Reference | 0.272 | 0.201 | *01:01, *03:01, *06:01, *09:01, *10:01, *11:01, *13:01, *14:01, *15:01, *17:01, *20:01 |

Estimated effects for haplotypes of *HLA-DRB1*, *HLA-B* and *HLA-DPB1*. Classical alleles of *HLA-DRB1* are grouped based on the amino acid residues present at positions 11 (or 13), 71 and 74 within DRβ1. The classical shared epitope alleles are shown in bold. 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). All effects are conditional on Asp9 in HLA-B and Phe9 in HLA-DPβ1. Unadjusted haplotype frequencies are given for cases and controls. *HLA-DRB1* haplotypes in aggregate explain 9.7% of the phenotypic variance of rheumatoid arthritis. The multivariate effect sizes, allele frequencies and classical alleles corresponding to Asp9 in HLA-B and Phe9 in HLA-DPβ1 are also listed. 95% CI, 95% confidence interval.



of the classical *HLA-B*08* allele ($P > 0.68$). Similar to positions 11, 71 and 74 in HLA-DRβ1, position 9 in HLA-B is also located in the peptide-binding groove (**Fig. 4**). Many of the previously described associations across the MHC, including markers in the *TNF* region, are in LD with Asp9 (ref. 10).

As previously observed associations of *HLA-B*08* to autoimmune diseases, including to rheumatoid arthritis, have been attributed specifically to the long ancestral 8.1 haplotype, which contains *HLA-B*08* on the *HLA-DRB1*03* background[9,11], we tested whether the *HLA-B*08*-Asp9 effect is common to all *HLA-DRB1* backgrounds. Because *HLA-B*08* and *HLA-DRB1*03* are not in perfect LD and are both seen independent of the 8.1 haplotype, we were able to apply a conditional haplotype analysis to show that *HLA-B*08*-Asp9 increases disease risk roughly twofold regardless of the *HLA-DRB1* background (**Fig. 5**). Therefore, this risk effect is not restricted to the 8.1 haplotype. Risk alleles for *HLA-B* and *HLA-DRB1* contribute risk additively (on a log-odds scale) even though they are in strong (but incomplete) LD.

Conditioning on the effects of *HLA-DRB1* and *HLA-B*, we observed the most significant association at *HLA-DPB1* in the class II HLA region ($P < 10^{-20}$; **Fig. 1c**), which corresponds to Phe9 in HLA-DPβ1 (OR = 1.40 relative to His9 or Tyr9; **Table 1**, **Fig. 3** and **Supplementary Table 4**). The Phe9 effect was significantly stronger than that of any two- or four-digit *HLA-DPB1* classical allele; Phe9 is in LD with and is indistinguishable from the Val8 allele in *HLA-DPB1*. Amino acid position 9 is within the binding groove of HLA-DP (**Fig. 4**).

We observed no residual signals across the MHC after conditioning on the effects of *HLA-DRB1*, *HLA-B*08* Asp9 in HLA-B and Phe9 in HLA-DPβ1 ($P > 3 \times 10^{-6}$; **Fig. 1d**). We also did not observe any evidence of epistatic interactions between known risk loci[13,20,21] and any of the HLA alleles described here ($P > 0.0003$; **Supplementary Note**).

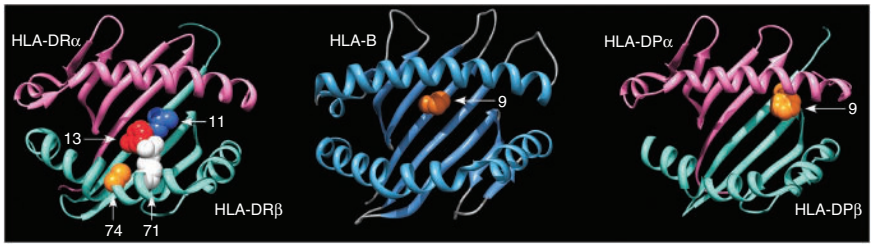


**Figure 4** Three-dimensional ribbon models for the HLA-DR, HLA-B and HLA-DP proteins. These structures are based on Protein Data Bank entries 3pdo, 2bvp and 3lqz, respectively, with a direct view of the peptide-binding groove. Key amino acid positions identified by the association analysis are highlighted. This figure was prepared using UCSF Chimera[25].

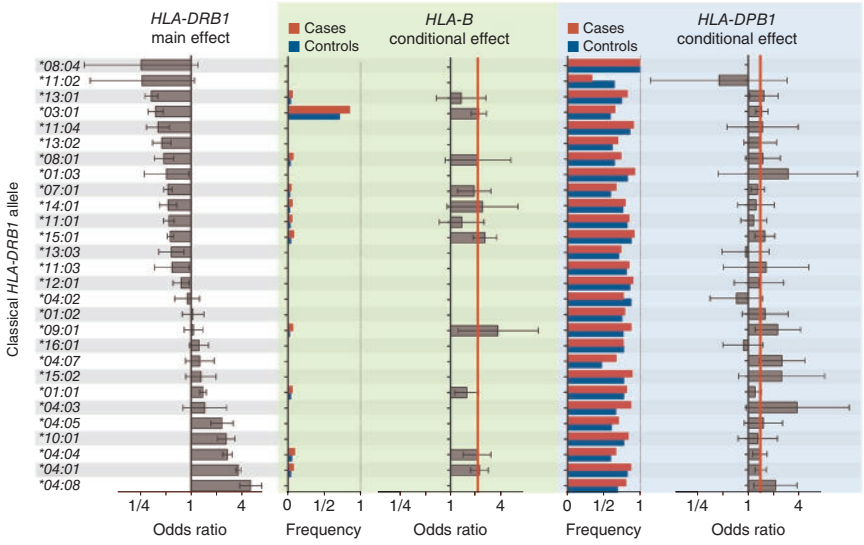


**Figure 5** Conditional haplotype analysis. Each row refers to a single classical *HLA-DRB1* allele. In the left box, the main (univariate) effect is plotted as an odds ratio (with 95% confidence intervals) for each *HLA-DRB1* allele (compared to not having that allele), sorted in order of anti-CCP–positive rheumatoid arthritis risk. In the middle box (in green), case and control allele frequencies and odds ratios are plotted for the HLA-B Asp9 allele. In the right box (in blue), case and control allele frequencies and odds ratios (with 95% confidence intervals) are plotted for the HLA-DPβ1 Phe9 allele. The red vertical lines indicate the aggregate effects for HLA-B and HLA-DPβ1 across all *HLA-DRB1* haplotypes. The Asp9 allele in HLA-B and the Phe9 allele in HLA-DPβ1 both have consistent effects across all *HLA-DRB1* haplotype backgrounds. This suggests that these three effects are additive and independent and are not the consequence of any individual extended haplotype.

These results are consistent with a disease model in which classical HLA genes and proteins are the dominant factors in rheumatoid arthritis pathogenesis, with only a minor contribution from non-HLA loci in the MHC.

A key finding of this study is the major influence of amino acids 11 and 13 within HLA-DRβ1 but outside of the well-described shared epitope region. It is possible that one position is driving the effect and the other is in tight LD with it. Alternatively, there may be a joint effect involving both amino acids that is driven by combined selection. This option is plausible given the key role of natural selection[22] in the MHC and the physical proximity of these two positions. To disentangle these effects, larger studies that include individuals of multiple ethnicities as well as many more examples of alleles where the LD between positions 11 and 13 is discordant will be necessary. Alternatively, if candidate rheumatoid arthritis auto-antigens can be determined, then these effects might be disentangled by comparing T cell responses to these antigens presented in the context of HLA-DRB1 molecules engineered to contain distinct combinations of amino acids at positions 11 and 13.

This study implicates three amino acid positions in HLA-DRβ1 and two additional amino acid positions in HLA-B and HLA-DP in conferring risk to anti-CCP–positive rheumatoid arthritis. These variants account for 12.7% of the phenotypic variance of seropositive rheumatoid arthritis risk, whereas common validated alleles outside the MHC explain ~4% of this variance[13] (**Supplementary Note**). The location of these positions within the peptide-binding grooves implies a functional impact on antigenic peptide presentation to T cells, either during early thymic development or during peripheral immune responses. The presence of class I and II MHC alleles implicates both $CD8^+$ cytotoxic and $CD4^+$ helper T cells in rheumatoid arthritis pathogenesis. In addition to rheumatoid arthritis, type 1 diabetes has also been shown to have strong MHC class I and II associations[23]. We also note that the *HLA-B*08* allele, which carries Asp9, has been documented in many autoimmune diseases, including myasthenia gravis, immunoglobulin A deficiency and systemic lupus erythematosus[24].

The pathogenic auto-antigens in the majority of autoimmune disorders remain under debate. For rheumatoid arthritis, these results could facilitate the evaluation of specific citrullinated polypeptides with molecular modeling and binding assays and, in doing so, guide our understanding of how HLA risk alleles influence the immune repertoire and disease susceptibility.

## 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 thank the Type 1 Diabetes Genetics Consortium and the Wellcome Trust Case Control Consortium for data access. We acknowledge use of the HLA genotyping data performed by the Juvenile Diabetes Research Foundation/Wellcome Trust Diabetes and Inflammation Laboratory on the British 1958 Birth Cohort DNA collection. This project is supported by grants from the US National Institutes of Health (K08AR055688 (S.R.), R01-AR44422 (P.K.G.), R01-AR057108 and U01-GM092691 (R.M.P.)), the Korea Healthcare Technology Research and Development Project (A102065 and A111218-11-GM01 (H.-S.L. and S.-C.B.)), the Burroughs Wellcome Fund (R.M.P.), the Arthritis Foundation (S.R.) and by the Eileen Ludwig Greenland Center for Rheumatoid Arthritis (P.K.G.).


### AUTHOR CONTRIBUTIONS

S.R. and P.I.W.d.B. conceptualized and coordinated the study, oversaw the statistical analyses and wrote the initial version of the manuscript. S.R., P.I.W.d.B., C.S., E.A.S., J.F. and X.J. conducted all the statistical analyses. H.-S.L., S.-C.B., L.A., L.P., L.K., J.W., K.A.S., R.M.P. and P.K.G. organized and contributed subject samples and collected genome-wide SNP data. S.-C.B., H.-S.L. and P.K.G. provided the classical HLA genotype data. All authors contributed to writing the final manuscript.

### 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://www.nature.com/reprints/index.html.



1. Isenberg, D. *Oxford Textbook of Rheumatology*, 1278 (Oxford University Press, Oxford, New York, USA, 2004).
2. Klareskog, L., Catrina, A.I. & Paget, S. Rheumatoid arthritis. *Lancet* **373**, 659–672 (2009).
3. Ding, B. *et al.* 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 (2009).
4. van der Woude, D. *et al.* Quantitative heritability of anti-citrullinated protein antibody-positive and anti-citrullinated protein antibody-negative rheumatoid arthritis. *Arthritis Rheum.* **60**, 916–923 (2009).
5. Gregersen, P.K., Silver, J. & Winchester, R.J. The shared epitope hypothesis. An approach to understanding the molecular genetics of susceptibility to rheumatoid arthritis. *Arthritis Rheum.* **30**, 1205–1213 (1987).
6. Stastny, P. HLA-D and IA antigens in rheumatoid arthritis and systemic lupus erythematosus. *Arthritis Rheum.* **21**, S139–S143 (1978).
7. Reinsmoen, N.L. & Bach, F.H. Five *HLA-D* clusters associated with *HLA-DR4*. *Hum. Immunol.* **4**, 249–258 (1982).

8. Vignal, C. *et al.* Genetic association of the major histocompatibility complex with rheumatoid arthritis implicates two non-*DRB1* loci. *Arthritis Rheum.* **60**, 53–62 (2009).
9. Lee, H.S. *et al.* Several regions in the major histocompatibility complex confer risk for anti-CCP-antibody positive rheumatoid arthritis, independent of the *DRB1* locus. *Mol. Med.* **14**, 293–300 (2008).
10. Newton, J.L., Harney, S.M., Wordsworth, B.P. & Brown, M.A. A review of the MHC genetics of rheumatoid arthritis. *Genes Immun.* **5**, 151–157 (2004).
11. Jawaheer, D. *et al.* Dissecting the genetic complexity of the association between human leukocyte antigens and rheumatoid arthritis. *Am. J. Hum. Genet.* **71**, 585–594 (2002).
12. 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).
13. Stahl, E.A. *et al.* Genome-wide association study meta-analysis identifies seven new rheumatoid arthritis risk loci. *Nat. Genet.* **42**, 508–514 (2010).
14. Brown, W.M. *et al.* Overview of the MHC fine mapping data. *Diabetes Obes. Metab.* **11** (suppl. 1), 2–7 (2009).
15. Pereyra, F. *et al.* The major genetic determinants of HIV-1 control affect HLA class I peptide presentation. *Science* **330**, 1551–1557 (2010).
16. Price, A.L. *et al.* Principal components analysis corrects for stratification in genome-wide association studies. *Nat. Genet.* **38**, 904–909 (2006).
17. van der Woude, D. *et al.* Protection against anti-citrullinated protein antibody-positive rheumatoid arthritis is predominantly associated with *HLA-DRB1*1301*: a meta-analysis of *HLA-DRB1* associations with anti-citrullinated protein antibody-positive and anti-citrullinated protein antibody-negative rheumatoid arthritis in four European populations. *Arthritis Rheum.* **62**, 1236–1245 (2010).
18. Freudenberg, J. *et al.* Genome-wide association study of rheumatoid arthritis in Koreans: population-specific loci as well as overlap with European susceptibility loci. *Arthritis Rheum.* **63**, 884–893 (2011).
19. Lee, H.S. *et al.* Microsatellite typing for *DRB1* alleles: application to the analysis of HLA associations with rheumatoid arthritis. *Genes Immun.* **7**, 533–543 (2006).
20. Raychaudhuri, S. Recent advances in the genetics of rheumatoid arthritis. *Curr. Opin. Rheumatol.* **22**, 109–118 (2010).
21. Zhernakova, A. *et al.* Meta-analysis of genome-wide association studies in celiac disease and rheumatoid arthritis identifies fourteen non-HLA shared loci. *PLoS Genet.* **7**, e1002004 (2011).
22. Trowsdale, J. The MHC, disease and selection. *Immunol. Lett.* **137**, 1–8 (2011).
23. Nejentsev, S. *et al.* Localization of type 1 diabetes susceptibility to the MHC class I genes *HLA-B* and *HLA-A*. *Nature* **450**, 887–892 (2007).
24. Price, P. *et al.* The genetic basis for the association of the 8.1 ancestral haplotype (A1, B8, DR3) with multiple immunopathological diseases. *Immunol. Rev.* **167**, 257–274 (1999).
25. Pettersen, E.F. *et al.* UCSF Chimera—a visualization system for exploratory research and analysis. *J. Comput. Chem.* **25**, 1605–1612 (2004).

## ONLINE METHODS

**Sample collections.** All cases met the 1987 American College of Rheumatology diagnostic criteria[26], were diagnosed by a board-certified rheumatologist and were confirmed as being positive for antibodies to CCP. Samples came from multiple studies, each of which received approval from the appropriate institutional review boards[13,18]; all participants signed informed consent.

For the primary analysis, we used six sample collections (**Supplementary Table 1**) from the UK (WTCCC), Sweden (Epidemiological Investigation of Rheumatoid Arthritis (EIRA)), Canada (CANADA), the United States (North American Rheumatoid Arthritis Consortium I (NARAC-I) and NARAC-III) and Boston, Massachusetts (Brigham and Women's Hospital Rheumatoid Arthritis Sequential Study (BRASS)) from a recent rheumatoid arthritis GWAS meta-analysis[13]. We followed the quality control steps outlined in the original publication[13]. Additionally, we excluded cases from the WTCCC sample that were not confirmed as being positive for antibodies to CCP ($n$ = 797 individuals), WTCCC shared controls that had previously been used to study other phenotypes and individuals that failed the *HLA-DRB1* phasing ($n$ = 57 individuals). All subjects were self-described white and were of European descent. In total, there were 5,018 cases and 14,974 controls.

For the secondary analysis, we used a South Korean collection of 616 cases and 675 controls recruited at the Hanyang University Hospital for Rheumatic Diseases in Seoul, South Korea, described in detail elsewhere[18]. Our study followed the quality control steps outlined in the original publication[18]. We excluded cases not confirmed as being positive for antibodies to CCP and individuals not successfully genotyped for *HLA-DRB1* classical alleles.

For all samples, we had access to genome-wide SNP data. The European samples were genotyped on different platforms (**Supplementary Table 1**). The South Korean samples were genotyped using Illumina HumanHap550v3 or HumanHap 660W platforms. All South Korean samples, a subset of the WTCCC samples ($n$ = 700, which was all the controls) and a subset of the NARAC-I samples ($n$ = 450) had full genotype data to a four-digit resolution at the *HLA-DRB1* locus. The South Korean samples were genotyped using PCR sequence-based typing; the NARAC-I samples were genotyped using sequence-specific oligonucleotide genotyping[9,19]. Some WTCCC controls were part of the 1958 British Birth Cohort[27] and were HLA typed at the Juvenile Diabetes Research Foundation/Wellcome Trust Diabetes and Inflammation Laboratory; the resulting data from that typing were made available through the European Genome-phenome Archive.

**Imputing HLA genotypes.** As previously described[15], we imputed classical HLA alleles and the corresponding amino acid sequences using reference data collected by the Type 1 Diabetes Genetics Consortium (T1DGC). This reference dataset contains genotype data for 2,537 SNPs, selected to tag the entire MHC, and classical types for *HLA-A*, *HLA-B*, *HLA-C*, *HLA-DRB1*, *HLA-DQA1*, *HLA-DQB1*, *HLA-DPA1* and *HLA-DPB1* at a four-digit resolution in 2,767 unrelated individuals of European descent[14]. As the data for the GWAS datasets were collected with different genotyping platforms, the overlapping SNPs between the GWAS and the T1DGC samples varied and ranged from 219 to 674 (**Supplementary Table 1**). We encoded all variants in the reference panel as biallelic markers, which facilitated the application of BEAGLE for imputation (using default parameters)[28]. For each dataset, we imputed cases and controls together.

**Statistical framework for association testing.** To test markers for an effect on risk that was fixed (consistent across datasets) and additive on the log-odds scale, we used logistic regression. To account for population stratification, we included as covariates five principal components for each individual dataset (**Supplementary Note**). We also included five indicator variables to account for cohort-specific effects or differences in the proportion of cases and controls between the GWAS datasets. This resulted in the following logistic regression model:

$$\log(\text{odds}_i) = \theta + \sum_{a=1\ldots m-1} \beta_a g_{a,i} + \sum_{j \in \text{collection}} \delta_{i,j}\left(\gamma_j + \sum_{k=1\ldots 5} \pi_{j,k} p_{i,k}\right)$$

where $a$ indicates the specific allele being tested, and $g_{a,i}$ is the dosage (imputed or genotyped) of allele $a$ in individual $i$. The $\beta_a$ parameter represents the additive effect per allele. For testing a multi-allelic locus with $m$ possible alleles (for example, amino acid residues at a specific position), we included $m$–1 $\beta$ parameters, one for each allele, where one allele was arbitrarily selected as the reference allele. We used the most frequent allele in the controls as the reference allele. Here, $\delta_{i,j}$ is an indicator variable that is equal to 1 only if individual $i$ is in the case-control collection $j$. The $\gamma_j$ parameter is the effect for the $j^{\text{th}}$ case collection and was set to 0 for one arbitrarily selected reference cohort. The $\pi_{j,k}$ parameter is the effect for each of the principal components, and $p_{i,k}$ is the value for individual $i$ for the $k^{\text{th}}$ principal component. The $\theta$ parameter represents a constant background rate (the logistic regression intercept).

**Testing across the MHC locus.** We defined a series of binary markers across the region using SNPs, classical HLA alleles and amino acid residues[15], as listed in **Supplementary Table 4**. For biallelic SNPs, the binary marker was the alternate (minor) allele. For classical HLA alleles, the binary marker was the presence of the allele or the absence of the allele. For binary amino acid residues, the binary marker was the presence of the less frequent amino acid in lieu of the more frequent one. For multi-allelic amino acid positions and SNP residues, we defined composite markers for testing for which each possible individual allele and combination of alleles was tested for association. For example, a biallelic SNP induces a single variable, a triallelic SNP induces three variables and a quadrallelic SNP induces six variables. Across the MHC, we applied the logistic regression framework above to test each of these binary markers for association, controlling for collection effects and population stratification. For each marker, we used probabilistic genotypes that took any uncertainty in imputation into account.

**Conditional analysis outside of the *HLA-DRB1* locus.** To assess whether there were independent effects outside of the *HLA-DRB1* locus, we used the same additive logistic regression approach described above to test all markers across the MHC. We included *HLA-DRB1* alleles as covariates, using either all four-digit classical *HLA-DRB1* alleles (which is more conservative) or the *HLA-DRB1* haplotypes defined by amino acid positions 11, 71 and 74 (**Table 1**). Both approaches yielded very similar results. If we identified other independently associated markers, we included them as covariates in our subsequent conditional analyses to identify additional independent effects.

**Analysis of *HLA-DRB1* amino acid sites.** To test the effects of amino acids in *HLA-DRB1*, we applied a conditional haplotype analysis. We tested each single amino acid position by first identifying the $m$ amino acid residues occurring at that position and then partitioning the classical alleles into $m$ groups of alleles with identical residues at that position. We estimated the effect of each of the $m$ groups using a logistic regression model (including covariates, as described above) and calculated the log-likelihood improvement in model fit over a null model. We assessed the significance of the improvement in fit by calculating the deviance (defined as −2 × the log likelihood), which follows a $\chi^2$ distribution with $m$–1 degrees of freedom. This is equivalent to testing a single multi-allelic locus for association with $m$ alleles.

For the conditional analyses, we assumed that the null model consisted of haplotypes as defined by residues at previously defined amino acid positions. Addition of another position with $m$ residues, if the amino acid is independent, may result in $k$ additional unique haplotypes. We tested whether the addition of those amino acid positions, and the creation of $k$ additional haplotype groups, improved on the previous set. We assessed the significance of the improvement in the log-likelihood value over the previous model (with fewer haplotype groupings) by calculating the deviance (which is distributed as $\chi^2$ with $k$ degrees of freedom).

We also used logistic regression with probabilistic dosages of amino acids, taking into account imputation uncertainty, and confirmed that the same amino acids emerged in the exact same order using this method as they did using the previous method.

***HLA* allele permutations to determine significance.** Given the polymorphic nature of the *HLA* genes and the large *HLA-DRB1* effect sizes, we wanted to assess whether the observed associations at positions 11, 71 and 74 in HLA-DRβ1 could emerge by chance by tagging classical alleles of differential risk. To test this, we repeatedly reassigned amino acid sequences to each of the classical *HLA-DRB1* alleles (as defined in the standard HLA dictionary[29]).

In each permutation, we selected amino acids sequentially and assessed the improvement in deviance after the addition of this amino acid. We conducted 10,000 such permutations, in each case selecting three polymorphic amino acids sequentially that most improved the model deviance. We compared the improvements achieved using these permutations by fitting randomized amino acid sequences to the observed improvement by fitting the actual data.

**Exhaustively testing combinations of amino acids.** We tested all possible amino acid pairs and triplets for association to disease risk. For each set of amino acid positions, we defined groups of classical *HLA-DRB1* alleles with consistent residues at those positions. We used those groups to predict rheumatoid arthritis risk and calculated for each of these models the log-likelihood improvement in risk prediction (and its significance) over the null model.

**Conditional haplotype analysis.** We were concerned that some of the multiple effects observed across the MHC region might be driven by LD of the associated alleles to other classical *HLA-DRB1* alleles. We obtained fully phased haplotypes across the MHC from the imputed data. Using the statistical framework and covariates as defined above, we individually tested each of the classical *HLA-DRB1* alleles. For each *HLA-DRB1* allele, we included a variable that represented its dosage (0, 1 or 2). We also included a variable that indicated the dosage of Asp9 (or *HLA-DRB1*08*) alleles of HLA-B in phase with the classical *HLA-DRB1* allele being tested and, similarly, included a variable that indicated the dosage of the Phe9 allele of HLA-DPβ1 in phase with the *HLA-DRB1* classical allele being tested.

**Availability of software.** Available from authors on request.


26. Arnett, F.C. *et al.* The American Rheumatism Association 1987 revised criteria for the classification of rheumatoid arthritis. *Arthritis Rheum.* **31**, 315–324 (1988).
27. Power, C. & Elliott, J. Cohort profile: 1958 British birth cohort (National Child Development Study). *Int. J. Epidemiol.* **35**, 34–41 (2006).
28. Browning, B.L. & Browning, S.R. 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 (2009).
29. Robinson, J. *et al.* The IMGT/HLA database. *Nucleic Acids Res.* **39**, D1171–D1176 (2011).





doi:10.1038/ng.1076