Linkage Disequilibrium and Age of HLA Region SNPs in Relation to Classic HLA Gene Alleles within Europe

Walter Bodmer, Irina Evseeva, Kristin Nicodemus, Carolina Bonilla, Susan Tonks

Linkage Disequilibrium and Age of HLA Region SNPs in Relation to Classic HLA Gene Alleles within Europe

Irina Evseeva[1*], Kristin K. Nicodemus[1,2*], Carolina Bonilla[1], Susan Tonks[1], Walter F. Bodmer[1,3§]

1. Department of Clinical Pharmacology, Old Road Campus Research Building, University of Oxford, Off Roosevelt Drive, Oxford OX3 7DQ, United Kingdom

2. Wellcome Trust Centre for Human Genetics, University of Oxford, Roosevelt Road, Oxford OX3 7BN, United Kingdom

3. Cancer and Immunogenetics Laboratory, Weatherall Institute of Molecular Medicine, John Radcliffe Hospital, Oxford OX3 9DS, United Kingdom

§ To whom correspondence and reprint requests may be addressed:

Sir Walter Bodmer FRCPath, FRS
Cancer and Immunogenetics Laboratory
Weatherall Institute of Molecular Medicine
John Radcliffe Hospital
Oxford OX3 9DS
United Kingdom
Email: walter.bodmer@hertford.ox.ac.uk
Telephone: +44 (0)1865 222 422
Fax: +44 (0)1865 222 431

* These authors contributed equally to this work.

Abstract

The HLA region on chromosome 6 is gene-rich and under selective pressure due to the high proportion of immunity-related genes. Linkage disequilibrium (LD) patterns and allele frequencies in this region are highly differentiated across broad geographical populations, making it a region of interest for population genetics and immunity-related disease studies. We examined LD in this important region of the genome among 6 European populations using 166 putatively neutral SNPs and the classical HLA-A, -B and –C gene alleles. We found the pattern of association between the classical HLA gene alleles and SNPs implied that most of the SNPs predated the origin of the classic HLA gene alleles. The SNPs most strongly associated with HLA gene alleles were in some cases highly predictive of HLA allele carrier status (misclassification rates ranged from less than 1% to 27%) in independent populations using 5 or fewer SNPs, a much smaller number than tagSNP panels previously proposed and often with similar accuracy, showing our approach may be a viable solution to designing novel HLA prediction panels. To describe the LD within this region, we developed a novel haplotype clustering method/software based on r^2, which may be more appropriate for use within regions of strong LD. Haplotype blocks created using this proposed method, classic HLA gene alleles and SNPs were predictive of northern versus southern European population membership (misclassification error rates ranged from 0-23% depending on which independent population was used for prediction), indicating this region may be a rich source of ancestry informative markers.

Keywords: HLA, population genetics, Europe, LD, haplotype

The HLA region on chromosome 6 is an important region of interest for both population genetics and immunity-related disease studies. Due to the selective pressure associated with immune functions, linkage disequilibrium patterns and allele frequencies are highly differentiated across populations. Because HLA typing of the classical HLA gene alleles is expensive and time-consuming, although necessary for transplantation matching and detailed analysis of disease-associations, we assessed whether a simple novel method could ascertain SNPs that were informative of HLA allele carrier status. One important consideration in finding such SNPs is the age of the SNP in question, and whether it is likely that it arose before or after the origin of the classical HLA gene alleles. SNPs that arose on a particular classical HLA gene allele haplotype background may be more informative for prediction of carrier status for that allele than SNPs which pre-date the origin of the allele. It is also of interest to examine the LD structure of the HLA region, especially within closely-related populations. Previous approaches for the creation of haplotype blocks have generally relied on the LD metric D', which may not be as sensitive within these high LD regions as the alternate LD metric r^2. We developed a novel haplotype blocking strategy based on r^2, and applied this method to SNP data in the HLA region across six European populations. Finally, we tested whether haplotype blocks, HLA alleles and SNPs were useful for differentiating European populations using logistic regression and unsupervised clustering algorithm approaches.Using these approaches we identify novel SNPs in the HLA region that may be useful as ancestry informative markers (AIMs) for European populations.

Material and Methods

Populations sampled

The project involved genotyping 657 DNA samples from unrelated representatives of 9 European populations: English (n=77, mainly from Birmingham), Orcadians from the Orkney Islands, UK (n=88), Catalans (n=66), Italians from Bergamo (n=82), Piedmont (n=59), and Sicily (n=59), French Basques (n=76), Finns (n=71) and unrelated individuals from the CEPH reference families in Utah (n=79). Nine European populations were available for study, of which we used 6 populations for logistic regression and clustering analyses. These populations were collapsed into northern European (Orkney) and southern European (Bergamo, Piedmont, Sicily, Catalan). This choice was based on the assumption that the Basques and the Finns are generally considered to be outlier populations, while the European CEPH are an unknown mixture, though retrospectively they are mostly Northern European. This is a limited sample of European populations, but serves to illustrate methods of analysis and does reveal major differences between northern and southern European populations.

SNP selection

Two sets of markers were selected:

1. One hundred and eighty-eight SNPs within the HLA region defined as putatively neutral by the following criteria: in introns, pseudogenes, intergenic regions, or synonymous changes. The SNPs were selected based on their chromosome position to provide, as far as possible, an even distribution across the 3.9MB HLA region (chromosome 6, 29587512 – 33516520, National Center for Biotechnology Information, Build 36.3) at an average density of 1 SNP per 18Kb. In addition, extra SNPs were chosen to cover areas around presumed recombination hotspots[1]. Only SNPs with minor allele frequency reported on NCBI as greater that 0.05 were included.

2. Seventy-eight SNPs in exons 2 and 3 of the HLA-A, B and C genes providing a low to medium resolution HLA Class I allele typing with a total of 69 alleles (locus A - 18, B - 31, C – 20) all with frequencies greater that 0.05 in Caucasians in the Allele*Frequencies in World Populations Database (http://www.allelefrequencies.net).

Genotyping

Thirty-five SNPs were successfully genotyped in house using ARMS-PCR (Amplification Refractory Mutation System) with KCl buffer and 15 ng of DNA in 6.5 µl PCR reaction. The detection of the product was done using AMDI (Alkaline-mediated differential interaction). One hundred and fifty-three SNPs were typed by the Centre National de Genotypage (Ivry, France), using a customized Illumina Beadarray Platform[2]. One hundred and thirty-one of these gave successful results and were included in the analysis, resulting in a total of 166 SNPs. HLA Class I allele typing was performed based on the 12th International Histocompatibility Workshop Protocol[3], but using a SNP based approach, as discussed above. The 78 'diagnostic' SNPs were typed by ARMS-PCR with $MgCl_2$ buffer and 15 ng of DNA in 6.5 µl PCR reaction, followed by AMDI detection[4]. All typing results were checked for Hardy-Weinberg equilibrium using a cut off p-value of 0.05 to exclude abberrant results. Internal quality control with 94 duplicates gave 100% concordance. Fourteen individuals had 5 or more missing genotypes and were removed from further analysis, thus the available N was 643.

Statistical Methods

Prediction of classic HLA alleles by SNPs

We tested whether the 166 SNPs predicted individual classic HLA-A, B and C gene alleles by deriving a binary variable for each HLA allele and each SNP. The binary

variable represented the presence or absence of the haplotype or minor allele, respectively, in an individual: thus, if an individual carried the haplotype or minor allele they were assigned a 1, otherwise a 0. We then used Fisher's exact test to test the 2x2 associations between each SNP and HLA allele. We corrected for multiple testing using a Bonferroni correction for the total number of tests (166 SNPs * 56 observed HLA alleles = 9, 296 tests). We ranked pairwise associations between SNPs and HLA alleles by $-\log_{10}$(p-value), and, using all SNPs passing Bonferroni correction, obtained counts of the number of SNPs where minor alleles were present in HLA allele carriers and non-carriers. In addition, we formally tested whether these SNPs were able to predict HLA carrier status using a split-half cross-validation approach. We estimated a logistic regression model on the training set containing the 5 most strongly associated SNPs (with the exception of HLA-B*44, which only had 4 SNPs in the Bonferroni-corrected set) to predict HLA allele carrier status. We used only the top 5 most strongly associated SNPs instead of the full set due to multicollinearity. We then used the test set to predict HLA allele carrier status. We validated the ability of this model built on the training data to predict HLA allele carrier status using the independent CEPH population. The logistic regression model was of the form:

$$\ln\left(\frac{p}{(1-p)}\right) = \sum_{i=1}^{N} \beta_i x_i$$

where the probability of carrying a particular HLA allele was the outcome to be predicted by the linear combination of x_i of N markers in the equation, which were binary-coded as the presence or absence of the minor frequency allele carried by that individual. We then calculated the sensitivity (number of predicted carriers/number of true carriers) and specificity (number of predicted non-carriers/number of true non-carriers) for the test set; since the association between the SNPs and HLA alleles was

calculated using the full set of 6 populations, this may overestimate both sensitivity and specificity. We therefore also calculated the same quantities for the independent CEPH population.

`r2blocks`*: a haplotype blocking algorithm based on* r^2

We implemented a new algorithm to define correlated clusters or blocks of SNPs based on the LD metric r^2 (`r2blocks`) and compared this algorithm to blocks defined by the program Haploview[5], which are based on evidence for historical recombination using D'[6]. D' is less sensitive in tightly-correlated LD regions than r^2. The clustering algorithm of `r2blocks` accepts genotype-level data or phased haplotypes as input and allows users to impute missing genotypes, set an r^2 threshold for defining blocks and set a maximum number (M) of SNPs to skip that do not pass the threshold while continuing to build a block. Briefly, starting with the highest r^2 value for all pairs of SNPs that are separated by at most M SNPs, it then calculates pair-wise r^2 measures for all SNPs within M SNPs of the first pair and continues to grow the block in either direction as long as one pair-wise r^2 value within M SNPs of any SNP within the current block is above the threshold, omitting SNPs that do not pass the threshold (Figure 1). It then considers the next SNP within the block and all pair-wise r^2 values for SNPs within M SNPs, growing the block until no additional SNPs remain or until no additional SNPs pass the threshold value. We evaluated two threshold values for the creation of haplotype blocks: $r^2 \geq 0.70$ and $r^2 \geq 0.5$ with M set to 4. The blocking algorithm is implemented in a freely-available contributed package `r2blocks` for the R statistical computing environment (www.r-project.org).

175 ***Population differentiation: HWE, F_{ST}, association, prediction and clustering***

176 Because population differentiation can cause departures from Hardy Weinberg
177 equilibrium (HWE), we tested for departures from HWE in the pooled population and in
178 the northern and southern population separately using Fisher's exact test.

179 We used Weir and Cockerham's[7] estimate of F_{ST} as implemented in the R
180 package `Geneland`[8] to assess genetic differentiation using the 166 SNPs and using the
181 top 20 SNPs, individual `r2blocks` blocks or two-locus HLA haplotypes that were
182 most strongly associated with north-south status. In addition, we calculated F_{ST} for
183 individual SNPs, `r2blocks` haplotype blocks, HLA alleles and HLA haplotypes.
184 Three-locus HLA haplotypes were estimated using PHASE v.2.1.1[9-10] with parent-
185 independent mutation; two-locus HLA haplotypes were derived from those estimates.

186 Allele- and genotype-based associations were tested between individual SNPs,
187 `r2blocks`-defined haplotypes, HLA alleles and 2 and 3-locus HLA-A, B and C
188 haplotypes, and north-south status using χ^2 tests or Fisher's exact test, when appropriate.
189 Allele-based tests tested the association between each allele and north-south status
190 (sample size = 2N); genotype-based tests tested the number of minor alleles versus
191 north-south status (sample size = N). We set the p-value threshold to 0.05.

192 We ranked association tests between north-south status and SNPs, haplotype
193 blocks and 2-loci HLA haplotypes by $-\log_{10}$(p-value) and considered the top 20 most
194 strongly associated predictors in a leave-one-population-out approach to validate
195 predictive ability on an independent southern population. We could not perform leave-
196 one-out analyses with the northern set due to small sample size; instead, we used the
197 CEPH sample for prediction. The logistic regression model used was of the form:

198 $$\ln\left(\frac{p}{(1-p)}\right) = \sum_{i=1}^{N} \beta_i x_i$$

199 where the probability of population membership is the outcome to be predicted by the
200 linear combination of x_i of N markers/haplotypes in the equation, which are coded as the
201 presence/absence of particular alleles/haplotypes carried by that individual. Prediction
202 of individual population assignment for the CEPH population and the removed southern
203 population was performed by calculating the probability of being northern European
204 using each individual's observed genotypes in the model. Misclassification rates were
205 calculated by taking the number of individuals misclassified given their 'true'
206 north/south label/total number of individuals in the independent population.

207 Current approaches to population differentiation detection using genomewide
208 sets of biallelic markers often apply unsupervised clustering algorithms, such as
209 principal components analysis (PCA, e.g., EIGENSTRAT[11], KPCA from the R package
210 `kernlab`[12]) or Bayesian methods such as implemented in STRUCTURE[13-14] or
211 BAPS[15-16]. We applied EIGENSTRAT, KPCA and BAPS to data from the 166 SNPs.
212
213 **Results**

214 ***Association of classic HLA alleles by genotyped SNPs***

215 Nearly all of the classic HLA-A, B and C gene alleles were strongly associated
216 with at least one SNP. Table 1 lists the top 20 most strongly associated SNP-HLA allele
217 pairs and Figure 2 shows the $-\log_{10}$(p-values) for those SNPs passing Bonferroni
218 correction (see Table S1 for a full list of all pairs passing Bonferroni correction). As
219 expected, most SNPs in strong LD with one or more alleles at a particular HLA locus
220 are physically proximal to the corresponding locus with peaks observed close to the

physical location of each gene, although LD patterns extended across most of the region (see Figure 2). Only a few very low-frequency HLA alleles (4 HLA-A alleles; 10 HLA-B alleles; 2 HLA-C alleles) were not observed to show Bonferroni-corrected association with genotyped SNPs. Fifty-five (33.1%) SNPs did not show corrected association with HLA alleles; of the 111 significantly associated SNPs, 72 (64.8%) showed association with $\leq$ 3 alleles. One SNP showed strong association with 7 HLA alleles: rs1265059 (HLA-A*29, HLA-B*07, HLA-C*0702, *06, *16, *0302 and *0303). The HLA-A*29/C*16 haplotype has a frequency of 2.8% in Northern Ireland and the HLA-B*07/C*0702 haplotype is frequent in the same population (17.0%) (frequencies from allelefrequencies.net), indicating this SNP may be tagging common haplotypes. Note the position of HLA-DRA1 is between SNPs 122-123 and that of HLA-DRB1, HLA-DQA1 and HLA-DQB1 are between SNPs 129-130, possibly explaining the strong association observed on the far right hand side of Figure 2.

Histograms of counts of the number of minor alleles carried at SNPs passing Bonferroni correction in HLA allele carriers and non-carriers clearly showed a bimodal distribution (Figure 3; also see Figures S1-S10). Using the top 5 most strongly associated SNPs and the HLA haplotypes observed with > 1% frequency in either northern or southern European populations (Table S2), we observed high sensitivity and specificity in predicting whether an individual carried a particular classic HLA allele in both the test set and in the independent CEPH set for most HLA alleles tested (Table 2). In particular, the overall misclassification rate for HLA-A*29 in the independent CEPH set was less than 0.01, with sensitivity near 1.0 (0.997) and 94.3% specificity. In addition, HLA-A*01 showed a less than 5% misclassification rate in the CEPH population (4.5%) and had 93.5% sensitivity and 99.9% specificity, indicating that

genotyping even a small number of SNPs can provide information about HLA allele carrier status, although not as complete information as direct HLA typing.

Comparison of haplotype blocking algorithms

Using an r^2 threshold of 0.70 and a window size (M) of 3, `r2blocks` creates 12 blocks across the HLA region in the pooled sample of European populations. Reducing the r^2 threshold to 0.50 leads to an additional 7 blocks and 14 additional SNPs being assigned to blocks (Figure 4; Table S3). The Gabriel block method using default parameters (lower bound D' confidence interval $\geq$ 0.7 and upper bound confidence interval $\geq$ 0.98) defines 18 blocks. The two additional blocking methods (the four gamete rule (FGR) and solid spine of LD with default D' threshold of 0.7) both create twice as many blocks over the region as compared to `r2blocks` and the Gabriel method, and assign nearly half of the HLA genomic region to haplotype blocks. In northern European populations, `r2blocks` using an r^2 threshold of 0.50 creates 20 blocks over the HLA region and in southern European populations 19 blocks; block regions were generally consistent across the two sets of European populations. Using `r2blocks` with an r^2 threshold of 0.50 leads to a similar number of blocks and number of SNPs assigned to blocks as the Gabriel method, although the block boundaries are often different. Not surprisingly, the use of `r2blocks` with the higher r^2 (0.70) threshold leads to a more similar percentage of genome covered as the Gabriel method than the lower r^2 threshold. `r2blocks` with both r^2 thresholds and the Gabriel method show that LD is slightly lower in southern versus northern European populations, but with similar average block size, number of SNPs assigned to blocks and percent of the

genome assigned to blocks in northern Europeans. Of the 20 blocks assigned using `r2blocks` with an r^2 threshold of 0.50, more than half (11; 55%) are in genic regions.

Hardy Weinberg Equilibrium, F_{ST}, association tests and clustering of markers/haplotypes and northern-southern European status

Excess deviations from HWE versus expected numbers of deviations can be induced by population structure. Of 166 SNPs tested, 15 (9.0%) were out of HWE at the $\alpha = 0.05$ level, almost double the number expected by chance alone (8.3) (Table S4 shows HWE, association test results and F_{ST} for all markers). In separate analyses of northern/southern populations the number of SNPs out of HWE was much nearer the expected value (8 and 10, respectively), suggesting that the excess when using the combined populations is probably due to population structure.

Overall F_{ST} values using the 20 SNPs, haplotypes and classic HLA gene alleles most strongly associated with north-south status were modest, as expected within European populations (Table S5; see Table 3 for full list of markers). The largest F_{ST} value (0.056) was observed between northern and southern populations using this panel of 20 alleles/markers/haplotypes, which were selected to highlight north-south differences; similarly, the second largest F_{ST} (0.050) was observed between southern populations and the CEPH sample. The smallest F_{ST} (0.0024) was found between northern populations and the CEPH sample. Single SNP F_{ST} values were strongly negatively correlated with both allelic ($r = -0.41$, $-\log_{10}$(p-value) = 7.42) and genotypic ($r = -0.42$, $-\log_{10}$(p-value) = 7.81) association test p-values for north-south status, indicating that the allelic/genotypic tests are similar measures to F_{ST} (Table S4). We note that single SNP F_{ST} estimates are approximately distributed as χ^2_1 and thus have large variances. The largest single SNP F_{ST} for differentiation between northern and

294 southern European populations was for rs411136 in SYNGAP1 (F_{ST} = 0.29);
295 interestingly, this same SNP showed the largest pair-wise F_{ST} between any two
296 populations, namely for for southern Europeans versus CEPH (F_{ST} = 0.51). The same
297 comparison between northern Europeans and CEPH, however, produced a F_{ST} of 0.015,
298 suggesting rs411136 may be an important AIM within European populations.
299 Seventy-four (44.6%) of the 166 SNPs were significantly associated with north-
300 south status using the uncorrected allelic or genotypic test, and after Bonferroni
301 correction for the 332 allelic and genotypic tests, 25 (15.1%) still showed significant
302 associations (Table 3; see also Table S4). The strongest association was observed with
303 rs411136 in SYNGAP1, with an allelic test $-\log_{10}$(p-value) of 32.42. As expected, the
304 largest 3-locus haplotype frequency difference between northern and southern European
305 populations was for haplotype HLA-A*01-HLA-B*08-HLA-C*0701 with frequencies
306 of 0.097 versus 0.034 respectively ($-\log_{10}$(p-value) = 4.16; Table S2)[17].
307 Association tests between the 20 haplotype blocks defined using `r2blocks`
308 with an r^2 threshold of 0.50 and north-south status showed in nearly all cases the
309 haplotype block was more strongly associated than the individual SNPs comprising the
310 block. Of the 20 haplotype blocks, 15 were associated with north-south status. Six of the
311 significantly associated blocks did not contain any individual SNPs that were
312 significantly associated with north-south status. Of the 9 blocks containing at least 1
313 significantly associated SNP, 6 of them showed stronger association with the haplotype
314 block containing that SNP than with all individual SNPs (Table S4; Figure S11). In fact,
315 of the 20 strongest-associated HLA allele haplotypes, SNPs or block-based haplotypes,
316 4 were haplotypes created using our novel methodology (Table 3).

KPCA and EIGENSTRAT analysis of single SNPs did not reveal tight clustering by north-south designation (Figure S12). Clustering of individuals using BAPS on single SNPs resulted in a best-fitting solution of 11 clusters; inspection of the proportion of individuals from northern and southern European populations in each cluster revealed a mixture of both; none of the clusters were comprised of purely northern or southern European samples (data not shown). However, BAPS clustering of the 6 population samples, instead of individuals, led to a best-fitting solution of 2 clusters exactly matching our northern and southern designations, confirming our *a priori* clustering of individuals.

Population membership prediction using HLA alleles, SNPs and blocks

We used logistic regression to test the predictive ability of the 20 most significantly associated HLA SNPs by using leave-one-population-out validation to predict north-south population membership on the independent population and on the CEPH population. To avoid multicollinearity induced by LD, we removed predictors that were strongly correlated with other predictors, retaining the predictor that was more strongly associated with north-south status, to thin the model from 20 variables to a model containing 7 predictors: rs411136, rs1265160, rs3096702, rs2256328, rs2855453, block 8 and HLA-B*08/C*0701. Even using such a limited number of predictors provided perfect prediction for the Piedmont population and low prediction error for the Sicily population (8.5%), although the error rates for the Bergamo (13.0%) and Catalan (18.0%) were higher (Table 4). The CEPH population prediction error rate was steady across all models at ~22%; however, given the genetic background of this population is not clearly defined, this may indicate some evidence for southern European admixture.

Discussion

Due to the time-consuming and expensive process needed to perform full classical HLA allele typing, we developed a simple strategy to identify inexpensively and easily-genotyped SNP combinations that were able to predict classic HLA gene allele carrier status in leave-one-population-out cross-validation using logistic regression models. We have also described a novel haplotype blocking method and software based on r^2, which is probably more appropriate than a D'-based method in regions of strong linkage disequilibrium. In addition, we have shown that haplotype blocks created using the novel haplotype blocking method, classic HLA gene alleles and neutral HLA region SNPs were useful for the differention of northern versus southern European populations, in agreement with previous work[18], and suggested particular SNPs that may be useful as AIMs.

An examination of the 15 associations in Table S1 that show the opposite classical HLA allele-SNP pattern to that observed in the other 304 associations, namely, the frequency of individuals carrying the HLA allele but *not* carrying the minor allele at the SNP is greater than the frequency of individuals *not* carrying the HLA allele but carrying the minor allele at the SNP, shows that the 4 SNPs associated with > 1 HLA allele are associated with known HLA haplotypes of high frequency in Europeans. rs404240 shows this pattern with HLA-A*01, HLA-B*08 and HLA-C*0701, which is the most frequent 3-locus haplotype in northern Europeans (0.034 in southern Europeans and 0.097 in northern Europeans); two additional SNPs, rs2001009 and rs2249099, also show this pattern with HLA-A*01 and HLA-C*0701. rs1800684 shows this pattern with HLA-B*07 and HLA-C*0702, a high-frequency haplotype in Europe

(0.060 in southern Europeans and 0.16 in northern Europeans). rs404240 and rs1800684 are both synonymous, and rs404240 and rs2249099 are physically proximal to the HLA-A gene, whereas rs2001099 is physically proximal to HLA-DRA. Because the opposite pattern is much more frequently observed (that individuals who *do not* carry the classical HLA *do* carry the minor allele at a particular SNP), this implies that most of these SNPs are older than the classical HLA alleles. We searched dbSNP build 36.3 (http://www.ncbi.nlm.nih.gov/projects/SNP/) for presumed ancestral (*Pan troglodytes*) alleles for these polymorphisms and found 26/166 (15.7%) were in regions not able to be aligned with the chimpanzee sequence (indicated in Table S4). Most of these SNPs were unassociated with north-south status, although 3 of the 26 were strongly associated with north-south status (rs2256328, rs2857205 and rs2747479). However, of these 26 SNPs, only 6 (23.1%) did not show Bonferroni-corrected association with at least one classic HLA gene allele, a smaller percentage than in the full set of 166 SNPs (31.1%) suggesting the existence of human-specific SNPs, that probably arose on a particular HLA allele haplotype background, and which might be more informative of classic HLA gene allele carrier status. The fact that these data suggest that most SNPs are older than the classic HLA gene alleles, most of which are common to humans and chimpanzees, suggests that they are not likely to be good markers for LD based associations. It seems likely that this is a problem shared by a high proportion of SNPs in the commonly-used SNP databases, since the SNPs used in this study were selected only by location. The age of the SNPs could account for the emphasis on building haplotype blocks using very high LD thresholds. Only in those cases where LD is very high and the SNPs are very closely linked will such blocking give meaningful results for relatively old SNPs, given that the average rate of decay of LD between two SNPs is 1-r

per generation, where r is the recombination fraction between the SNPs. For example, for a distance of 1000 bp, corresponding, on average, to $r = 10^{-5}$, the LD would decay by a factor of 0.0034 in 500, 000 generations, and so to negligible levels in the separation time between humans and chimpanzees. The SNPs may show associations with more recent variation, as with HLA alleles, but these associations will be incomplete.

We show that, even with highly ancestry informative markers, unsupervised clustering algorithms were not able to detect substructure with our limited number of SNPs. Clustering algorithms should be used with caution when genome-wide data are unavailable, even if the SNPs selected are informative of ancestry.

The bimodal distribution of minor allele carriers of sets of SNPs that are strongly associated with classic HLA gene alleles and the use of logistic regression to predict HLA allele carrier status are computationally efficient and simple methods that do not require particular 'tag SNPs'[19] or prior database-based information[20] and thus may be preferred when no previous data exist on a particular population. Even though our method and previously-described methods show relatively high sensitivity and specificity for prediction of classic HLA gene allele carrier status, none of the proposed methods, including ours, is as accurate as HLA allele typing. Even with this caveat, our method may be helpful in pre-selecting a subset of individuals for full classic HLA gene allele typing in disease association studies, thus reducing genotyping time and costs.

Higher-order associations, such as associations between blocks of SNPs, may more accurately describe genetic diversity and historical recombination patterns of a particular region of the genome, and may be helpful for the assignment of classic HLA gene allele carrier status than previously-reported approaches[17, 19-20]. Future work will develop a novel meta-blocking algorithm to perform higher-level blocking using blocks

created by `r2blocks`, to be used as input to unsupervised and/or supervised clustering algorithms for the detection of population stratification based on the example of the HLA region in closely-related populations. This meta-blocking algorithm may also be useful in prediction of classic HLA gene allele carrier status.

Acknowledgements

The Project was funded by European Union (Linkage Disequilibrium in European Populations, 2001-2005, CT-2001-00916), the Wellcome Trust (support for KKN) and Cancer Research UK (support for CB and ST). DNA samples were provided by Project partners : Prof. Howard Cann (Fondation Jean Dausset-CEPH, Paris, France), Prof. Laurent Excoffier (Computational and Molecular Population Genetics Lab, Zoological Institute, University of Bern, Switzerland), Prof. Antti Sajantila (Department of Forensic Medicine, Laboratory of Forensic Biology, University of Helsinki, Finland), Prof. Alberto Piazza (Dipartimento di Genetica, Biologia e Biochimica, Universita di Torino, Italy), Prof. Silvana Santachiara (Department of Genetics and Microbiology, University of Pavia, Italy), Prof. Jaume Bertranpetit (Biologia Evolutiva, CEXs, Universitat Pompeu Fabra, Barcelona, Spain)

Conflict of Interest: The authors declare no conflict of interest.

References

1. Miretti MM, Walsh EC, Ke X, *et al*: A high-resolution linkage-disequilibrium map of the human major histocompatibility complex and first generation of tag single-nucleotide polymorphisms. *Am J Hum Genet* 2005; **76**: 634–646.

2. Shen R, Fan JB, Campbell D, *et al*: High-throughput SNP genotyping on universal bead arrays. *Mutat Res* 2005; **573**: 70-82.

3. Tonks S, Marsh S, Bunce M, Bodmer JG Molecular typing for HLA class I using ARMS-PCR: Further development following the 12th International Histocompatibility Workshop. *Tissue Antigens* 1999; **53**: 175-183.

4. Bartlett S, Straub J, Tonks S, Wells RS, Bodmer JG, Bodmer, WF Alkaline-mediated differential interaction (AMDI): A simple automatable single-nucleotide polymorphism assay. *Proc Natl Acad Sci USA* 2001; **98**: 2694-2697.

5. Barrett JC, Fry B, Maller J, Daly MJ Haploview: analysis and visualization of LD and haplotype maps. *Bioinformatics* 2005; **21**: 263-265.

6. Gabriel SB, Schaffner SF, Nguyen H, *et al*: The structure of haplotype blocks in the human genome. *Science* 2002; **296**: 2225-2229.

7. Weir BS, Cockerham CC Estimating F-statistics for the analysis of population structure. *Evolution* 1984; **38**: 1358-1370.

8. Guillot G, Mortier F, Estoup A Geneland: A program for landscape genetics. *Mol Ecol Notes* 2005; **5**: 1261-1280.

9. Stephens M, Smith N, Donnelly P A new statistical method for haplotype reconstruction from population data. *Am J Hum Genet* 2001; **68**: 978-989.

10. Stephens M, Donnelly P A comparison of Bayesian methods for haplotype reconstruction from population genotype data. *Am J Hum Genet* 2003; **73**: 1162-1169.

11. Price AL, Patterson NJ, Plenge RM, Weinblatt ME, Shadick NA, Reich D Principal components analysis corrects for stratification in genome-wide association studies. *Nat Genet* 2006; **38**: 904-909.

12. Karatzoglou A, Smola A, Hornik K kernlab: Kernel-based Machine Learning Lab. 2008 R package version 0.9-8.

13. Pritchard JK, Stephens M, Donnelly P Inference of population structure using multilocus genotype data. *Genetics* 2000; **155**: 945-959.

14. Falush D, Stephens M, Pritchard JK Inference of population structure using multilocus genotype data: linked loci and correlated allele frequencies. *Genetics* 2003; **164**: 1567-1587.

15. Corander J, Waldman P, Silanpää MJ Bayesian analysis of genetic differentiation between populations. *Genetics* 2003; **163**: 367-374.

16. Corander J, Marttinen P, Siren J, Tang J Enhanced Bayesian modeling in BAPS software for learning genetic structures of populations. *BMC Bioinformatics* 2008; **9**: 539.

17. Bodmer JG. The HLA System: The HLA-DR antigens and HLA haplotypes in 2 populations; in Eriksson E (ed): Population Structure and Genetic Disorders. Acad Press, 1980, pp 211-238.

18. Cavalli-Sforza LL, Menozzi P, Piazza A. The history and geography of human genes. Princeton University Press, 1994, Princeton.

18. de Bakker PI, McVean G, Sabeti PC, *et al*: A high-resolution HLA and SNP haplotype map for disease association studies in the extended human MHC. *Nat Genet* 2006; **38**: 1166-1172.

19. Leslie S, Donnelly P, McVean G A statistical method for predicting classical HLA alleles from SNP data. *Am J Hum Genet* 2008; **82**:48-56.

Titles and legends to figures

Figure 1. Schematic of the `r2blocks` algorithm. HAPLOVIEW plot of pair-wise r^2 values between a set of 7 simulated SNPs; the block shading shows strength of correlation. A, Assuming a window size (M) of 4 and an r^2 threshold of 0.70, the `r2blocks` algorithm begins with the highest pair-wise LD value, here between SNPs 3 and 5, which are in perfect LD ($r^2 = 1.0$). Starting with SNP 3, consider r^2 values with SNPs 1, 2 and 4. Only SNP 1 passes the r^2 threshold; add SNP 1 to the block. Discontinue growing the block to the left. Consider r^2 values between SNP 5 and SNPs 4, 6 and 7 and add SNP 6. Move to SNP 6, consider r^2 values between SNP 6 and 7, which is below the threshold. Terminate growing block to the right, creating block 1 of SNPs 1, 3, 5 and 6. Now consider r^2 values between SNPs not assigned to blocks: SNPs 2, 4 and 7; none of the pair-wise r^2 values are above the threshold so the algorithm terminates, leaving these three SNPs as singletons. B. The resulting haplotype block from A.

Figure 2. Association of HLA-A, B and C alleles by genotyped SNPs. Plot shows $-\log_{10}$(p-values) of Fisher's exact tests for association between classic HLA gene alleles and genotyped SNPs passing Bonferroni correction. Association with the classic HLA gene alleles only are plotted in primary colours (HLA-A = blue, HLA-B = dark yellow, HLA-C = red); association with 2 classic HLA gene alleles are shown as secondary colours (HLA-A (blue) and HLA-B (yellow) = green, HLA-A (blue) and HLA-C (red) = violet, HLA-B (yellow) and HLA-C (red) = orange); association with all 3 classic

HLA gene alleles = black. The position of HLA-DRA1 is between SNPs 122-123 and that of HLA-DRB1, HLA-DQA1 and HLA-DQB1 are between SNPs 129-130.

Figure 3. Histograms of the number of minor allele carriers at associated SNPs by HLA allele carrier status. A. HLA-A*01 (18 SNPs), B. HLA-A*03 (12 SNPs), C. HLA-B*08 (27 SNPs); y-axis = frequency, x-axis = number of SNPs where individuals carry at least one minor allele. Blue = HLA allele non-carrier; red = HLA allele carrier.

Figure 4. HLA region haplotype blocks in European populations defined by `r2blocks` and the Gabriel, four gamete rule and solid spine of LD methods. Plots show LD heatmap of pair-wise r^2 values for SNPs. Top bar represents physical spacing of SNPs. Triangles show location of haplotype blocks defined by each method. Methods are indicated on the left hand side of each plot. A: pooled European populations; red triangles show blocks added by reducing the r^2 threshold from 0.7 to 0.5 using `r2blocks`. B: blocks obtained using `r2blocks` with r^2 threshold 0.70 in northern European populations (top) and southern European populations (bottom) C: blocks obtained using `r2blocks` with r^2 threshold 0.50 in northern European populations (top) and southern European populations (bottom). D: blocks obtained using the Gabriel method in northern European populations (top) and southern European populations (bottom).

Table 1. Counts, odds ratios (ORs) and $-\log_{10}$(p-values) for association tests for the top 20 pairs of Bonferroni-corrected significantly positively associated 'tagging' pairs of HLA alleles and SNPs.

HLA Allele	SNP	+ +[1]	+ -	- +	- -	OR	-log10 (p-value)[2]
HLA-A*03	rs3121593	82	9	3	337	946.2	73.74
HLA-A*02	rs6909253	202	14	64	151	33.66	59.8
HLA-C*06	rs10484554	72	0	52	307	425.1	48.92
HLA-C*05	rs2524160	67	3	37	324	191.1	47.44
HLA-A*03	rs6921921	88	3	74	266	104.2	43.41
HLA-A*01	rs1150741	105	8	82	236	37.43	43.18
HLA-B*08	rs3094014	68	7	57	299	50.23	36.76
HLA-A*03	rs2734925	89	2	102	238	102.9	36.2
HLA-C*06	rs2523619	72	0	110	249	163	33.25
HLA-A*03	rs1737043	88	3	116	224	56.2	31.14
HLA-A*24	rs2394186	66	17	63	285	17.4	29.88
HLA-B*51	rs2523685	69	3	107	252	53.71	28.93
HLA-C*06	rs3130473	72	0	128	231	129.9	28.46
HLA-A*24	rs1150741	79	4	108	240	43.53	28.06
HLA-A*11	rs2076177	50	7	58	316	38.41	27.98
HLA-B*08	rs3094216	66	9	80	276	25.06	27.95
HLA-A*11	rs29226	50	7	62	312	35.5	27.64
HLA-C*1203	rs10484554	46	0	78	307	181.1	27.56
HLA-B*07	rs3093993	87	12	88	244	19.94	27.3
HLA-A*01	rs404240	46	67	4	314	53.25	26.9

[1] + + refers to counts of co-occurrence of the HLA haplotype indicated and a minor allele at the SNP indicated; + - refers to the presence of the HLA haplotype but no copies of the minor allele at the SNP; - + refers to no copies of the HLA haplotype and at least one copy of the minor allele at the SNP; - - refers to no copies of the HLA haplotype indicated or the minor allele at the SNP. [2] $-\log_{10}$(p-value) is for the association test between copies of HLA alleles and minor alleles carried.

Table 2. Sensitivity, specificity and misclassification rates for 5-SNP logistic regression models predicting HLA allele carrier status.

HLA Allele	Mean Test Set Sensitivity (95% CI)	Mean Test Set Specificity (95% CI)	Number of CEPH carriers (%)	Mean CEPH Sensitivity (95% CI)	Mean CEPH Specificity (95% CI)	Mean CEPH Misclassification (95% CI)
HLA-A*01	0.953 (0.901, 1.00)	0.869 (0.745, 0.989)	23 (35.4)	0.935 (0.925, 0.944)	0.999 (0.996, 1.00)	0.0453 (0.0385, 0.0521)
HLA-A*02	0.878 (0.770, 0.987)	0.762 (0.695, 0.829)	40 (61.5)	0.937 (0.765, 1.00)	0.823 (0.776, 0.882)	0.147 (0.124, 0.170)
HLA-A*03	0.974 (0.953, 0.996)	0.949 (0.867, 1.00)	18 (27.7)	0.936 (0.875, 0.997)	0.982 (0.924, 1.00)	0.0540 (0.00501, 0.103)
HLA-A*29	0.977 (0.945, 1.00)	0.826 (0.677, 0.976)	3 (4.6)	0.997 (0.978, 1.00)	0.943 (0.678, 1.00)	0.00748 (0.000201, 0.0351)
HLA-B*07	0.907 (0.863, 0.952)	0.793 (0.656, 0.930)	19 (29.2)	0.863 (0.923, 0.903)	0.834 (0.730, 0.937)	0.147 (0.124, 0.168)
HLA-B*08	0.952 (0.921, 0.983)	0.859 (0.702, 1.00)	17 (26.2)	0.882 (0.850, 0.913)	0.914 (0.785, 1.00)	0.114 (0.0862, 0.142)
HLA-B*44[1]	0.818 (0.726, 0.910)	0.686 (0.508, 0.863)	23 (35.4)	0.722 (0.590, 0.853)	0.938 (0.758, 1.00)	0.265 (0.135, 0.394)
HLA-B*57	0.954 (0.926, 0.982)	0.708 (0.496, 0.920)	0 (0)	---	---	---
HLA-C*0701	0.828 (0.784, 0.873)	0.818 (0.690, 0.946)	22 (33.8)	0.828 (0.812, 0.843)	0.765 (0.687, 0.844)	0.191 (0.161, 0.220)
HLA-C*0702	0.889 (0.840, 0.938)	0.753 (0.608, 0.898)	3 (4.6)	0.972 (0.949, 0.995)	0.106 (0.0347, 0.177)	0.225 (0.178, 0.272)
HLA-C*05	0.964 (0.928, 1.00)	0.773 (0.638, 0.908)	4 (6.2)	0.924 (0.922, 0.927)	0.0 (0.0, 0.0)	0.247 (0.221, 0.273)
HLA-C*06	0.997 (0.992, 1.00)	0.946 (0.829, 1.00)	0 (0)	---	---	---
HLA-C*16	0.977 (0.955, 0.999)	0.870 (0.375, 1.00)	3 (4.6)	0.972 (0.960, 0.985)	0.828 (0.275, 1.00)	0.0405 (0.00680, 0.0742)

[1]This model contains 4 SNPs.

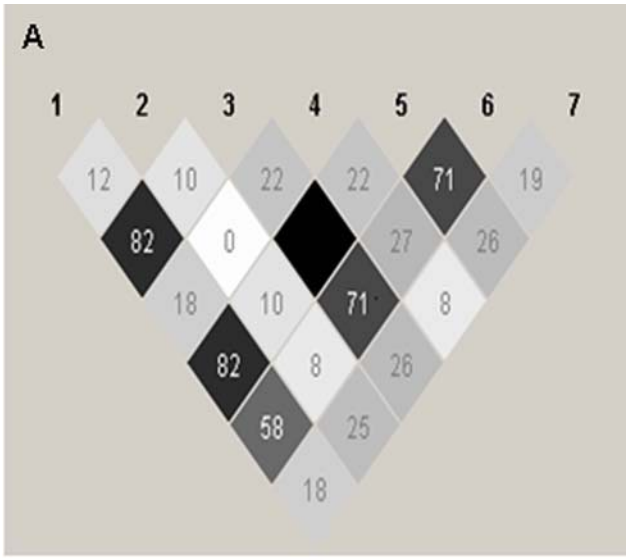
Table 3. Association tests for Hardy-Weinberg equilibrium, minor allele frequencies and F_{ST} for 20 most strongly associated HLA alleles, blocks and haplotypes with north-south status.

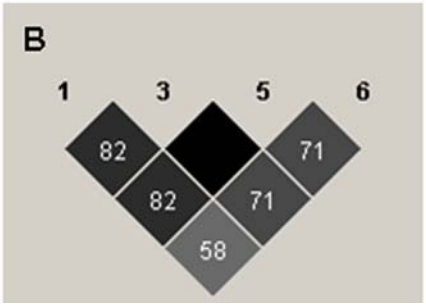
Marker	Location	Gene Symbol	Function	Combined MAF	Combined HWE $-\log_{10}$ p-value	North-South Single SNP Association – $\log_{10}$ p-value[1]	North-South Single Block Association – $\log_{10}$ p-value ($r^2 \geq 0.5$)	North-South F_{ST}	North MAF	North HWE $-\log_{10}$ p-value	South MAF	South HWE $-\log_{10}$ p-value
rs411136	33516520	SYNGAP1	Ser556Ser	0.15	4.00	32.42 (28.57)	---	0.29	0.33	0.37	0.034	0.00
rs1265160	31246350	POU5F1	Phe3Phe	0.070	0.66	11.15 (11.55)	---	0.099	0.15	**1.44**	0.023	0.00
rs3096702	32300309	NOTCH4	5' region	0.29	0.08	9.20 (8.29)	---	0.087	0.42	0.39	0.22	0.02
rs2256328	31489616	MICA	intron	0.24	0.33	8.87 (9.10)	---	-0.0021	0.12	0.93	0.31	0.24
rs2256594	32294850	NOTCH4	intron	0.13	0.00	8.46 (8.11)	---	0.078	0.21	0.36	0.073	0.10
rs2535318	31159367	---	---	0.49	0.06	7.48 (6.66)	---	0.070	0.37	0.00	0.44	0.19
rs659445	31972283	EHMT2	intron	0.28	0.57	7.07 (5.91)	---	0.066	0.39	0.16	0.22	0.03
Block 20	33438959-33454164	---	1-1 haplotype	0.40	0.23	---	6.36 (3.96)	0.030	0.42	0.35	0.60	0.09
Block 10	31159367-31186474	C6orf15	2-1 haplotype	0.42	0.51	---	5.65 (4.83)	0.037	0.51	0.71	0.34	0.19
rs389883	32055439	STK19	intron	0.23	0.10	5.49 (4.87)	---	0.050	0.31	0.07	0.17	0.33
rs2855453	33242370	COL11A2	intron	0.31	0.97	5.44 (4.51)	---	0.047	0.40	0.00	0.25	0.88
Block 8	30832409-30847883	---	1-1 haplotype	0.16	0.15	---	5.4 (4.15)	0.031	0.72	0.00	0.85	0.10
HLA-B*07/ C*0702 Haplotype	---	HLA-B/C	---	---	---	---	5.32 (5.32)	0.050	0.16	---	0.060	---
rs1800684	32259972	AGER, PBX2	Ala3Ala, 3' region	0.087	0.16	5.29 (5.07)	---	0.042	0.14	0.20	0.053	0.03
rs1810472	33191099	---	---	0.32	0.23	5.14 (4.74)	---	0.048	0.23	0.00	0.38	0.65
rs4713505	32212979	---	---	0.31	1.58	5.13 (4.01)	---	0.045	0.22	0.59	0.37	0.70
HLA-B*08/ C*0701 Haplotype	---	HLA-B/C	---	---	---	---	4.47 (4.82)	0.042	0.13	---	0.049	---
rs8512	30819336	IER3	3' UTR	0.19	4.39	2.34 (4.74)	---	0.017	0.14	0.00	0.22	**6.64**
rs211452	33438959	---	---	0.40	0.23	4.69 (4.11)	---	0.039	0.31	0.35	0.46	0.090
Block 17	32331236-32345991	---	2-1 haplotype	0.32	0.13	---	4.31 (4.64)	0.024	0.21	0.00	0.11	0.00

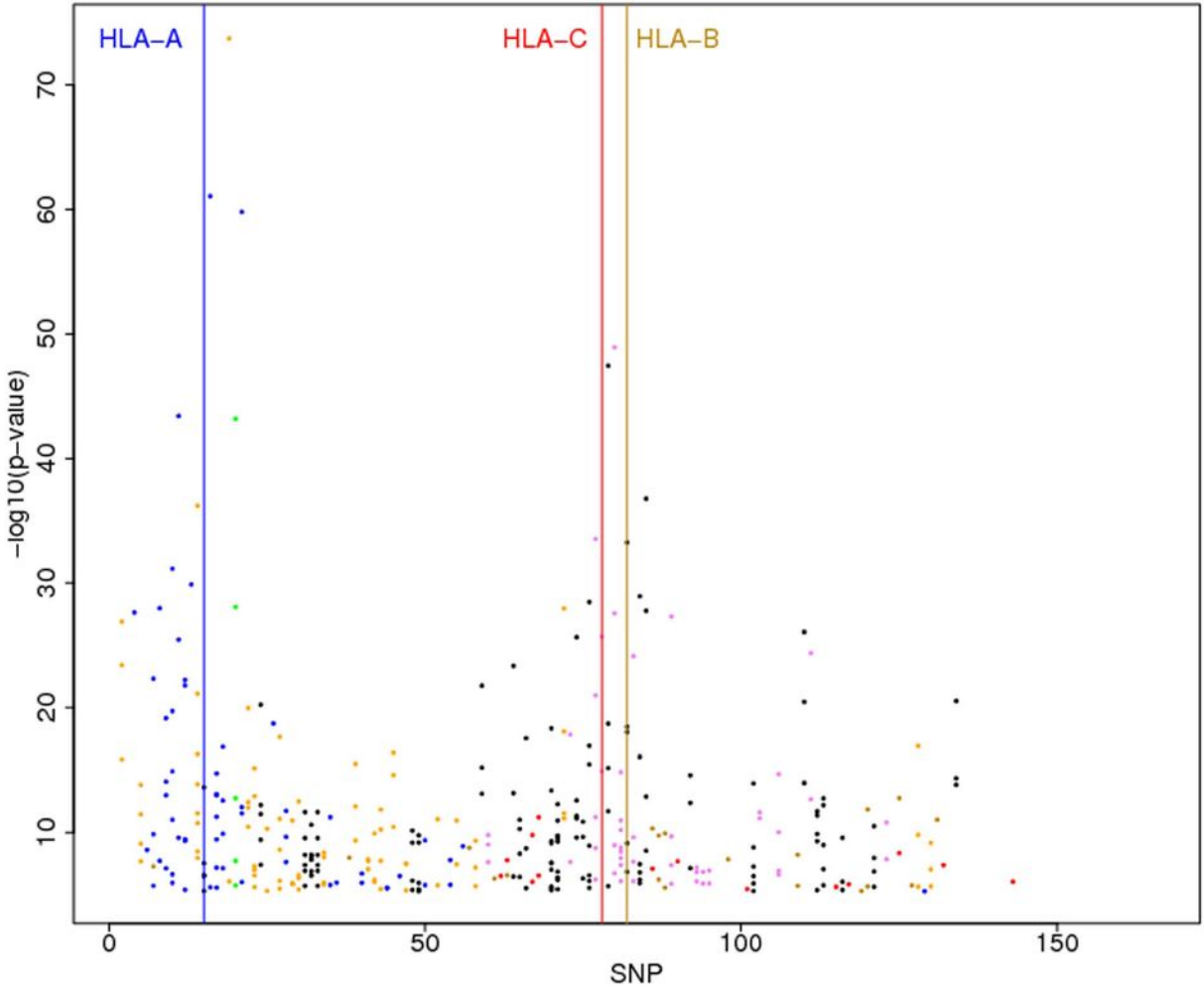
SNPs shown are associated with north-south status after Bonferroni correction for all allelic and genotypic tests. [1]p-values for association tests are shown as allele-based first, genotype-based in parentheses.

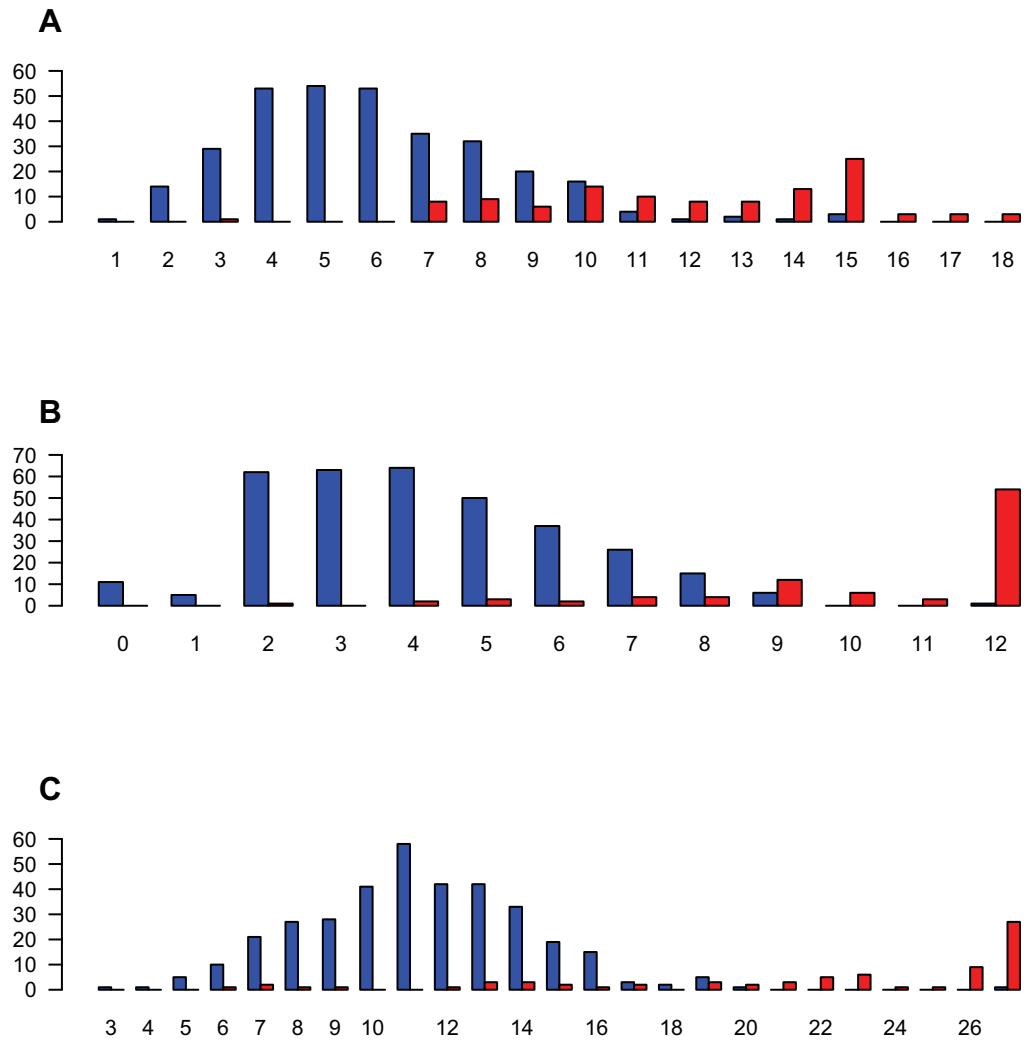
Table 4. North-south population membership prediction error for logistic regression models using the most associated SNPs, blocks and HLA-A/B and HLA-B/C haplotypes.

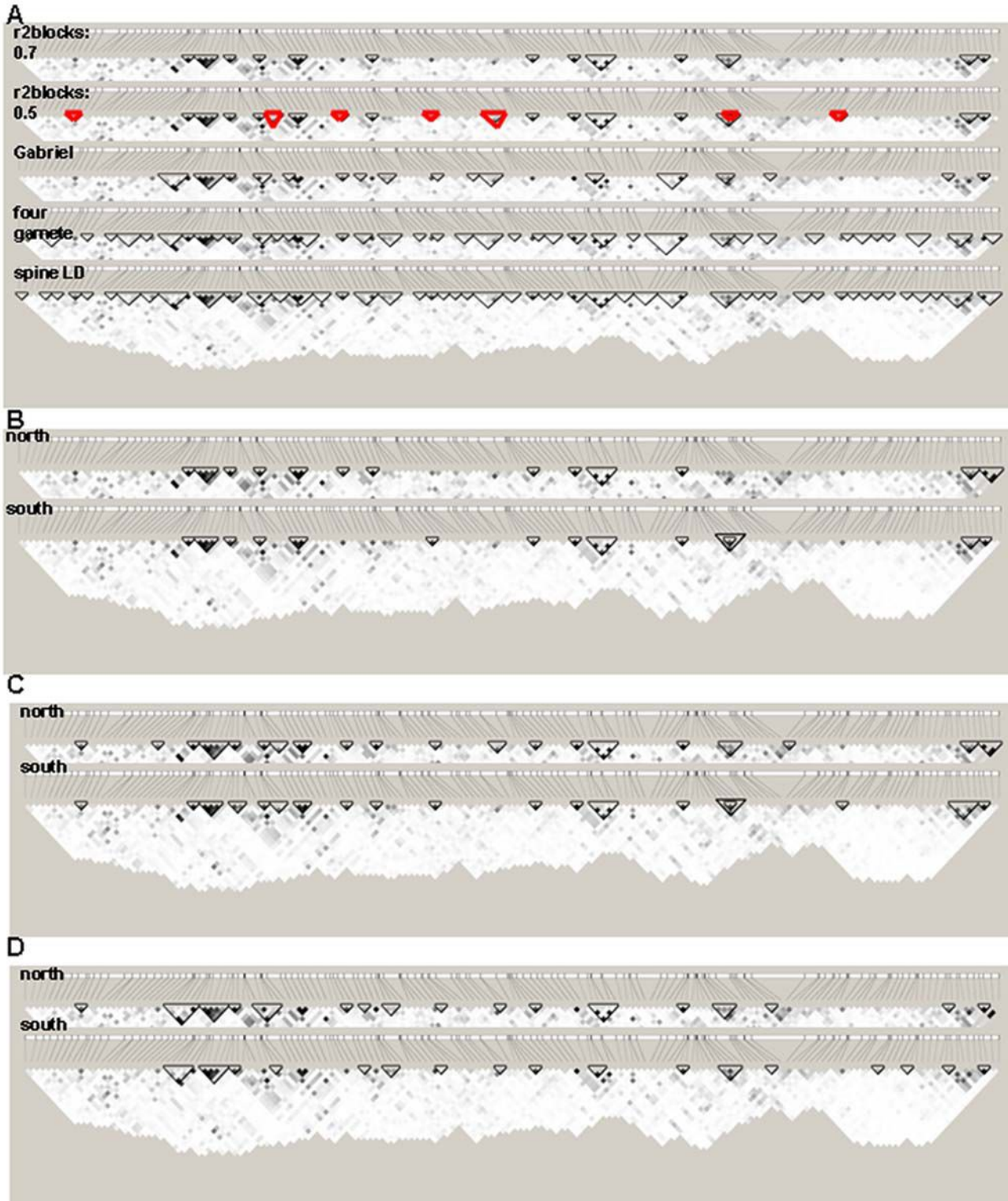
Population removed	Population Predicted	Misclassification rate
Catalan	Catalan	0.18
Catalan	CEPH	0.23
Bergamo	Bergamo	0.13
Bergamo	CEPH	0.22
Piedmont	Piedmont	0.00
Piedmont	CEPH	0.22
Sicily	Sicily	0.085
Sicily	CEPH	0.22


A
1 2 3 4 5 6 7
12 10 22 22 71 19
82 0 27 26
18 10 71 8
82 8 26
58 25
18


B
1 3 5 6
82 71
82 71
58


HLA–A
HLA–C
HLA–B
−log10(p−value)
10
20
30
40
50
60
70
0
50
100
150
SNP


A
60
50
40
30
20
10
0
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18
B
70
60
50
40
30
20
10
0
0 1 2 3 4 5 6 7 8 9 10 11 12
C
60
50
40
30
20
10
0
3 4 5 6 7 8 9 10 12 14 16 18 20 22 24 26


A
r2blocks:
0.7
r2blocks:
0.5
Gabriel
four
gamete
spine LD
B
north
south
C
north
south
D
north
south