Human Heredity

Hum Hered 2013;76:162–177
DOI: 10.1159/000360855

Published online: May 21, 2014

A New HLA Map of Europe: Regional Genetic Variation and Its Implication for Peopling History, Disease-Association Studies and Tissue Transplantation

Alicia Sanchez-Mazas Stéphane Buhler José Manuel Nunes

Laboratory of Anthropology, Genetics and Peopling History, Department of Genetics and Evolution – Anthropology Unit and Institute of Genetics and Genomics in Geneva (IGE3), University of Geneva, Geneva, Switzerland

Key Words

HLA · Population genetics · European · Caucasian · Demographic history · Genetic diversity · Human populations · Linkage disequilibrium · Multi-allelic loci · Natural selection · Population data · Phenotypic variation · HLA haplotypes · HLA-NET · EUROSTAM

Abstract

Objectives: HLA genes are highly polymorphic in human populations as a result of diversifying selection related to their immune function. However, HLA geographic variation worldwide suggests that demographic factors also shaped their evolution. We here analyzed in detail HLA genetic variation in Europe in order to identify signatures of migration history and/or natural selection. ***Methods:*** Relationships between HLA diversity and geography were analyzed at 7 loci through several approaches including linear regression on gene diversity and haplotype frequencies. Regional variation was also assessed on HLA multi-locus phenotypes through *structure* analysis. Deviation from neutrality was tested by resampling. ***Results:*** Geographic distance was a strong predictor of HLA variation at 5 loci (A, B, C, DRB1 and DPB1) in Europe, and latitude significantly shaped HLA gene diversity and haplotype frequencies. Whereas the main level of genetic diversity was found within populations, both HLA gene frequencies and phenotypic profiles revealed regional variation, Southeast Europe, Great Britain and Finland being the most distinctive. Effects of natural selection were suggested at the DQ loci. ***Conclusions:*** HLA regional variation was observed in Europe and can be related to population history, locus HLA-A providing by far the strongest signals. This new HLA map of Europe represents an invaluable reference for disease-association studies and tissue transplantation.

Introduction

The genetic diversity of European populations has been investigated through the analysis of many classical and DNA polymorphisms, most of them aiming at identifying the demographic processes by which this continent was peopled [1–8, among others]. As discussed in this volume [9], current debates among geneticists still focus, for example, on the real contributions of modern humans' settlements at either the Neolithic or Paleolithic periods, and recent results of ancient DNA studies are also examined in this context [see, e.g. 10]. Another key

KARGER

E-Mail karger@karger.com
www.karger.com/hhe

0001–5652/14/0764–0162$39.50/0

Alicia Sanchez-Mazas
Laboratory of Anthropology, Genetics and Peopling History
Department of Genetics and Evolution – Anthropology Unit, University of Geneva
12 rue Gustave-Revilliod, CH–1211 Geneva 4 (Switzerland)
E-Mail alicia.sanchez-mazas@unige.ch

issue to understand the patterns of genetic variation observed in Europe is the role of environmental factors. For instance, genes associated with visible phenotypic traits, like skin pigmentation differences in humans [11, 12], have been identified as likely targets of positive selection related to climatic factors. Another well-known example is lactase activity persistence; the high frequency of this trait in Europeans has been explained by different models including latitude-related biological advantage or gene-culture co-evolution, or both [13, 14].

Among all human genes showing some evidence of natural selection, those coding for HLA molecules involved in adaptive immune response [15] exhibit a particularly high level of polymorphism in human populations [16, 17]. This observation has been explained by many causes [for a review, see 18], the most realistic being perhaps an increased immune protection of heterozygous individuals in pathogenic environments (the model is known as the pathogen-driven balancing selection or PDBS). For that reason, HLA has often been discarded from population genetic studies seeking genetic signatures of past human migrations. As a revealing example related to the peopling history of Europe, Fix [19] sustained that the genetic clines presented as an evidence of demic diffusion (i.e. population migration) in the seminal paper written by Menozzi et al. [20] were principally driven by the inclusion of many (more than 50%) HLA alleles; and, for this author, the clinal variation patterns observed for HLA were the result of diversifying selection acting on Neolithic farmers, due to increased disease intensities related to plant and animal domestication. Actually, this interpretation is attractive as it supports the idea that the Neolithic period played a crucial role in the generation of gene frequency clines [21], although not in this case through a demic diffusion process. On the other hand, our knowledge on the HLA polymorphism has radically changed since the period of such debates thanks to the accumulation of large population data sets typed for these genes [22–27, www.allelefrequencies.net] and to the precise molecular characterization of HLA alleles achieved since then at multiple loci (www.ebi.ac.uk/ipd/imgt). The paper entitled 'An HLA Map of Europe', presented by Ryder et al. [28] in the same year as the publication by Menozzi et al. [20], thus needed to be completely revisited to better understand by which mechanisms, e.g. population migrations, PDBS or others, the HLA genetic patterns were shaped in this continent. Indeed, although many HLA studies have been published on specific European populations in the last decades [for a review, see 27], no updated meta-analysis has been performed so far at the level of this continent. In addition, such an update is fundamental in 2 other domains related to medicine: disease-association studies and tissue transplantation. Indeed, many HLA alleles and, more recently, numerous SNPs located in non-translated regions of the extended HLA region have revealed associations with a number of human diseases such as autoimmune disorders and severe infections like HIV or HBV among others [29–31]. In this context, a thorough knowledge of HLA variation among populations, e.g. in the case of Europe, is needed to control for possible effects of population stratification as a result of such associations. Also, donor-recipient HLA matching is a main issue in tissue transplantation, for patients needing either an organ graft or hematopoietic stem cell transplantation. Huge databases of potential donors are continuously interrogated in order to find HLA-matched individuals [see also 32, this volume]. Here again, a thorough knowledge of HLA diversity on different continents, like Europe, is most useful; it helps developing such databases and eventually optimizing the search for compatible donors.

In this study, we thus investigated the genetic variation at 7 HLA loci (A, B, C, DRB1, DQA1, DQB1 and DPB1) in Europe by analyzing a large set of 145 population samples located in different regions, including Southeast and Northeast Europe which were underrepresented in previous studies. Our aim was, first, to identify possible signals of past demographic events and/or of selective effects acting on the HLA loci to better understand the causes of the observed HLA genetic variation in Europe; and second, to provide a new and comprehensive picture of HLA genetic variation among European populations to be used in medical research and for clinical applications.

Materials and Methods

Population Data

We used a large database of 145 European (EUR), North African (NAFR) and West Asian (WASI) population samples tested for 1–7 HLA loci (HLA-A, -B, -C, -DRB1, -DQA1, -DQB1 and -DPB1) [33]. The total number of individuals per locus varied between 1,582 and 125,889 (also depending on the resolution level considered). This database includes data from the 12th, 13th, 15th and 16th International Histocompatibility workshops [22–25, 34], the 'HLA-NET' project [35] and other published sources [36–39] (online suppl. table S1 and suppl. fig. S1; for all online suppl. material, see www.karger.com/doi/10.1159/000360855). All these data correspond to samples analyzed with DNA-typing techniques (PCR-SSO, PCR-SSP or SBT). For the analyses, the data were reported at either a first-field level of resolution (to include larger sets of population samples which have not been fully described at high resolution) or a second-field level of resolution (to keep more

molecular information). We only considered populations fitting Hardy-Weinberg equilibrium (HWE) according to a likelihood-ratio test (LRT) implemented in Gene[rate] that compares the log-likelihoods of the frequency estimates under Hardy-Weinberg and under a generalized non-Hardy-Weinberg model [23, 40–42]. Updated frequencies and detailed summary statistics (HWE, neutrality and linkage disequilibrium) for each population can be found in Nunes et al. [33]. We only present here meta-analyses of these statistics.

Neutrality Test

As HLA loci are known to evolve under some selective pressure, the strength of which remains to be determined, we statistically tested selective neutrality in all population samples available at a second-field level of resolution to avoid a possible loss of information when alleles are grouped into first-field generic specificities. This was done by using a resampling procedure where samples drawn from the estimated population allele frequencies were tested for neutrality using Slatkin's version of the Ewens-Watterson test. The p values obtained from the resampling schema are taken as a proxy to assess the neutral status of the sample of interest. The method was improved by replacing overconservative results provided by Bonferroni's correction [43] with false-discovery-rate-adjusted p values [44], hence providing minimum and maximum adjusted p values [the procedure is described in detail in Nunes et al. 33]. The tests were performed without prior assumptions, thus two-tailed rejection at the 5% level occurred either below 2.5% for excess of heterozygosity or above 97.5% for excess of homozygosity. The summary of these analyses are presented graphically with box-and-whisker plots [graphs produced with R, version 2.15.3; 45].

Linkage Disequilibrium

The HLA region exhibits more linkage disequilibrium than the human genome average, although with high heterogeneity along the region [46]. To explore linkage disequilibrium in relation to population stratification in Europe, we assessed the global strength of association by means of 3 procedures:

- the use of a LRT based on the likelihood of allele and haplotype frequency estimates (under the null hypothesis of no linkage disequilibrium, the number of degrees of freedom is the number of potentially observed haplotypes);
- the generation of an empirical distribution for the LRT statistic defined above, through a resampling procedure: a number of samples (here 10,000) were drawn from the observed allele frequencies and used to re-estimate both allele and haplotype frequencies; the final result of the test is the percentile of the observed LRT statistic (PRS) in the empirical distribution;
- the contribution of each putative haplotype to a χ^2 test (standardized residuals); the difficulty of providing a global result for these latter statistics resides in the determination of the appropriate number of degrees of freedom to use (they depend on the number of classes observed, expected or collapsed, which vary hugely); therefore, we did not retain this procedure to report global results.

The assessment was made for the following pairs of loci: A-B, B-DRB1 and A-DRB1, i.e. the most widely used loci pairs in studies reporting haplotype and linkage disequilibrium results, as well as B-C, DRB1-DQB1 and DQA1-DQB1 corresponding to the most proximal loci from a physical point of view. Linkage disequilibrium was studied only at the first-field level both because of the greater number of samples available and the larger sample sizes allowing a better estimation, and because second-field level estimates of haplotype frequencies are often very low and less reliable.

Genetic Distance Analyses

To get a general view on pan-European HLA genetic variation, we first used Prevosti et al.'s [47] genetic distances to plot non-metric multidimensional-scaling analyses (NMDS) [48] for each locus and each resolution level, and we assessed the correlations between genetic and geographic distances through Mantel's tests [49]. These analyses were performed with R [45] extended with packages vegan version 2.0 [50] and ade4 version 1.5 [51]. Wright's F_{ST}, F_{SC} and F_{CT} indexes and their significance, tested through an analysis of molecular variance, were computed with Arlequin version 3.0 [52].

Linear Regression Models

To better understand the relationship between genetics and geography in Europe, we used linear models to assess how gene diversity, linkage disequilibrium and selected haplotype frequencies were related to the geographic location of the populations. Gene diversity (estimated by the expected heterozygosity under HWE based on allele frequencies), global linkage disequilibrium (as described above) and selected AB haplotype frequencies were chosen as dependent variables, and latitude and longitude as independent variables. The models were considered both globally (i.e. at all loci or loci pairs together) and separately (i.e. at each locus, loci pair or AB haplotype) allowing for interactions. A model was considered better than alternative models on the basis of partial F test results for nested models and if it contained significant coefficients (slopes) other than the base intercepts. The coefficients were considered significantly different from 0 if the p value of the associated t test was smaller than 5%, the same level as that used for the F statistic. The diagnostics of the retained models included inspection of residuals' distribution for independence and homoscedasticity and for normality and existence of influential points. No outliers presented a Cook's distance larger than 0.5, thus no points were excluded from the analysis. These models were used with HLA data reported at both first-field and second-field levels of resolution for gene diversity, but only with first-field data for linkage disequilibrium and AB haplotype frequencies (the number of samples available for a second-field analysis was considered too small). Model estimation, comparisons by means of partial F statistics, diagnostics and graphs were done using the standard functionalities of R [45].

Structure Analysis

Finally, to take into account the genetic information provided by several HLA loci simultaneously, we used the *structure* clustering method proposed by Pritchard et al. [53] on a subset of 25 population samples (8,170 individuals) typed for HLA-A, -B and -DRB1 (first-field level of resolution). By using *structure*, our aim was to identify a number of putatively stable multi-locus allelic combinations (hereafter named 'phenotypic profiles') that would be the most distinctive for the populations in different geographic regions of Europe. This approach has the advantage of avoiding comparing populations on the basis of 3-locus haplotype frequencies, the latter being generally too low to be estimated accurately and to be used with confidence in genetic distance analyses. The

admixture model was of course the most appropriate to account for the huge diversity of phenotypic profiles among individuals, and we set the default values proposed by the program for most parameters (in particular, the burn-in and MCMC replications were set to 1,000 and 10,000, respectively). We determined the best number of phenotypic profiles K by estimating the log-likelihood ln Pr(X|K) for K = 2, 3, 4, 5, 6. As ln Pr(X|K) may vary significantly between different runs, we carried out 100 runs for each K and reported the distributions of ln Pr(X|K) in box-and-whisker plots and density graphs drawn with R [45]. To display the results for K = 2, 3, 4, 5, 6, the *distruct* software [54] was used with the run results nearest to the mode of each density distribution. To keep the maximum of 5,000 individuals imposed by *distruct* to represent the results graphically, we truncated the largest samples (1,000 British Welsh, 1,000 Irish and 2,390 North Italians) to 350 individuals each to display a final graph representing 4,980 individuals (this had no consequences on the numerical results and their interpretation, which were always based on the total set of 8,170 individuals).

Results

Neutrality Tests

The proportions of minimum and maximum adjusted p values at each locus (second-field level of resolution) are given in table 1 and shown as box-and-whisker plots in online supplementary figure S2. No excess of heterozygotes was ever observed for HLA-DPB1, while the other loci exhibited between 18% (HLA-C) and 45% (HLA-DQA1) significant rejections, with close values for A, DQB1 and DRB1 (34–35%). However, besides the remarkable result found for HLA-DPB1, the minimum adjusted p values exhibited a high median and a very large variance at locus HLA-A, while the median was much lower and the values much more concentrated at HLA-B, -C, -DQA1, -DQB1 and -DRB1 (online suppl. fig. S2). This indicates a weaker and more heterogeneous signal of balancing selection at locus A. In addition, 47% of the populations exhibited a significant excess of homozygotes at locus DPB1, against 0–8% at the other loci (table 1), revealing directional or purifying selection for the former.

HLA Genetic Variation among Populations

Mantel's statistics *r* between genetic and geographic distances are reported in table 2. We considered both all populations (EUR, NAFR and WASI) taken together and only European (EUR) populations, at both levels of reso-

Table 1. Selective neutrality tests at the second-field level of resolution

Locus	n	Heterozygous excess		Homozygous excess	
		min. sign.	proportion	max. sign.	proportion
A	26	9	34.6%	2	7.7%
B	30	12	40.0%	1	3.3%
C	22	4	18.2%	0	–
DPB1	17	0	–	8	47.1%
DQA1	22	10	45.5%	0	–
DQB1	49	17	34.7%	0	–
DRB1	44	15	34.1%	1	2.3%

n = Number of populations tested; min. sign. = number of populations with a significant minimum adjusted p value; max. sign. = number of populations with a significant maximum adjusted p value.

Table 2. Correlation between the genetic and geographic distances

Locus	First-field level of resolution						Second-field level of resolution					
	all			EUR			all			EUR		
	n	*r*	p	n	*r*	p	n	*r*	p	n	*r*	p
A	75	0.583	**1e-4**	58	0.517	**1e-4**	26	0.491	**1e-4**	13	0.363	**0.0087**
B	79	0.603	**1e-4**	58	0.625	**1e-4**	30	0.409	**1e-4**	13	0.518	**3e-4**
C	44	0.574	**1e-4**	27	0.638	**1e-4**	22	0.304	**8e-4**	10	0.329	0.0435
DPB1	–	–	–	–	–	–	17	0.419	**0.0024**	13	0.374	**0.0018**
DQA1	26	0.116	0.056	19	0.173	0.0167	22	0.220	0.0403	15	0.217	0.0188
DQB1	67	0.195	**6e-4**	45	0.253	**4e-4**	49	0.150	0.039	30	0.231	**0.0076**
DRB1	113	0.395	**1e-4**	83	0.430	**1e-4**	44	0.397	**1e-4**	25	0.351	**3e-4**

Prevosti et al.'s [47] genetic distances and natural logarithm of geographic distances are used.

all = EUR, NAFR and WASI; n = number of populations tested; *r* = correlation coefficient; p = Mantel's test p value (9,999 permutations). $p < 0.01$ are shown in bold.

Table 3. Analysis of molecular variance showing, at each HLA locus, the proportion of the total genetic variation that is due to differences between Northeastern, Central-Western and Southeastern Europe (F_{CT}) and between populations within all 3 regions (F_{SC}), and their significance

Locus	First-field level of resolution				Second-field level of resolution			
	F_{CT}	p	F_{SC}	p	F_{CT}	p	F_{SC}	p
A	0.0041	<1e-5	0.0018	<1e-5	0.0087	0.0020	0.0041	<1e-5
B	0.0056	<1e-5	0.0029	<1e-5	0.0201	0.0001	0.0067	<1e-5
C	0.0108	<1e-5	0.0040	<1e-5	0.0134	0.0040	0.0102	<1e-5
DPB1	–	–	–	–	0.0035	0.0318	0.0075	<1e-5
DQA1	0.0150	0.0002	0.0095	<1e-5	0.0163	0.0005	0.0064	<1e-5
DQB1	0.0148	<1e-5	0.0080	<1e-5	0.0160	0.0003	0.0189	<1e-5
DRB1	0.0081	<1e-5	0.0038	<1e-5	0.0155	0.0002	0.0085	<1e-5

There are no Northeastern EUR populations characterized at the second-field of resolution at HLA-DQA1, thus a structure of only 2 groups was tested.

lution. Except in one case (HLA-C, second-field, EUR only), the *r* value was both high (above 30%) and highly significant ($p < 0.01$) for HLA-A, -B, -C, -DRB1 and -DPB1, with extreme values for loci A, B and C (50–60%) at the first-field level. By contrast, DQA1 and DQB1 showed low (below 26%) and most often poorly ($0.01 < p < 0.05$) or nonsignificant ($p > 0.05$) *r* values, except for DQB1 at the first-field level and second-field level in EUR. Interestingly, these two loci exhibit greater F_{ST} values among populations than the others (online suppl. table S2) at the first-field level of resolution, and at the second-field level for DQB1 (however, as many more population samples were tested for DQB1 at this resolution level, a sampling effect is not excluded). We also noted that in general the *r* value was lower when data were considered at the second-field rather than at the first-field level of resolution, but this was probably due to smaller numbers of populations tested at the second-field level (e.g. only 10 for HLA-C in EUR, which may explain the nonsignificant p value).

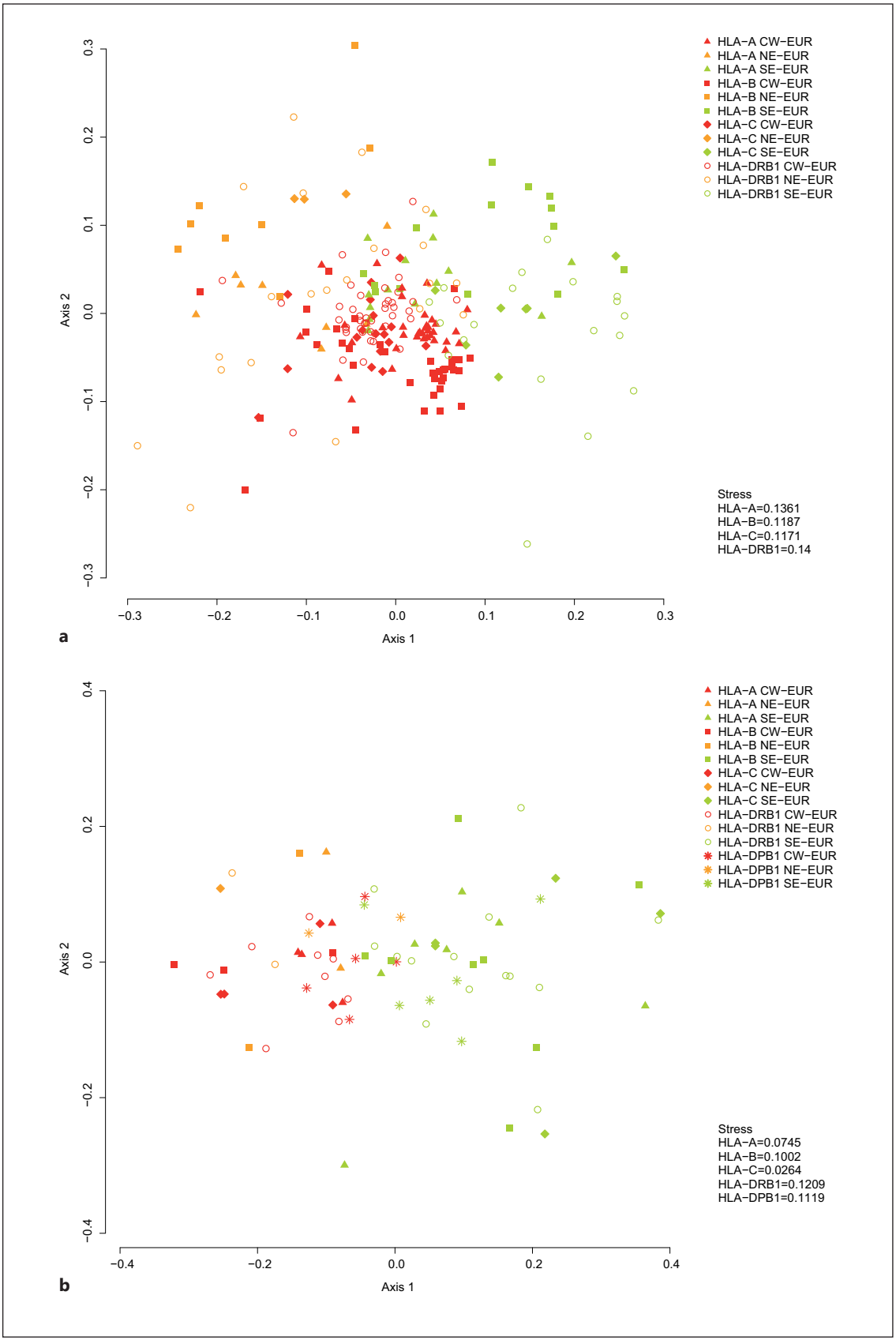
Based on the results described above (contrasting Mantel's *r* and F_{ST} values), we set apart the DQA1 and DQB1 loci, and we superimposed the NMDS obtained for A, B, C, DRB1 and DPB1 on identical scales to illustrate HLA genetic variation in relation to geography in Europe, with 3 colors showing Southeastern, Northeastern and Central-Western subregions (fig. 1a, b at the first-field and second-field levels of resolution, respectively) and for all populations to show the relationships between Europe, West Asia and North Africa (online suppl. fig. S3a, b, respectively). The 5 loci show very similar pictures, both in the relative position of each subregion (highlighted by colors) and in the extension of the population clouds (as expected from close F_{ST} values). European populations from Northeastern (NE-EUR, in orange), Central-Western (CW-EUR, in red) and Southeastern (SE-EUR, in green) regions differentiate from each other despite some overlap mostly between Northeastern and Central-Western subregions at the second-field level of resolution (fig. 1a, b). This differentiation is confirmed by the results of an analysis of molecular variance presented in table 3: we detect highly significant F_{CT} values among the 3 European regions, with values up to 2-fold the ones of the F_{SC} at all loci except DPB1 and DQB1 (second-field level of resolution). In addition, when compared to neighboring regions, Southeastern European largely overlap with West-Asian (in blue) populations (online suppl. fig. S3a, b). Overall, these composite NMDS (in particular those corresponding to first-field level data) resemble a geographic map of Europe and its surrounding regions.

Distinctive HLA Phenotypic Profiles in European Populations

By running the *structure* program, we explored pan-European HLA variation in more details at the pheno-

Fig. 1. Composite NMDS based on Prevosti et al.'s [47] genetic distances between European populations for HLA-A, -B, -C and -DRB1 at the first-field level of resolution (**a**) and for HLA-A, -B, -C, -DRB1 and -DPB1 at the second-field level of resolution (**b**). Symbols are used to differentiate the loci. Central-Western (CW-EUR), Northeastern (NE-EUR) and Southeastern (SE-EUR) European populations are colored in red, orange and green, respectively.

(For figure see next page.)



1

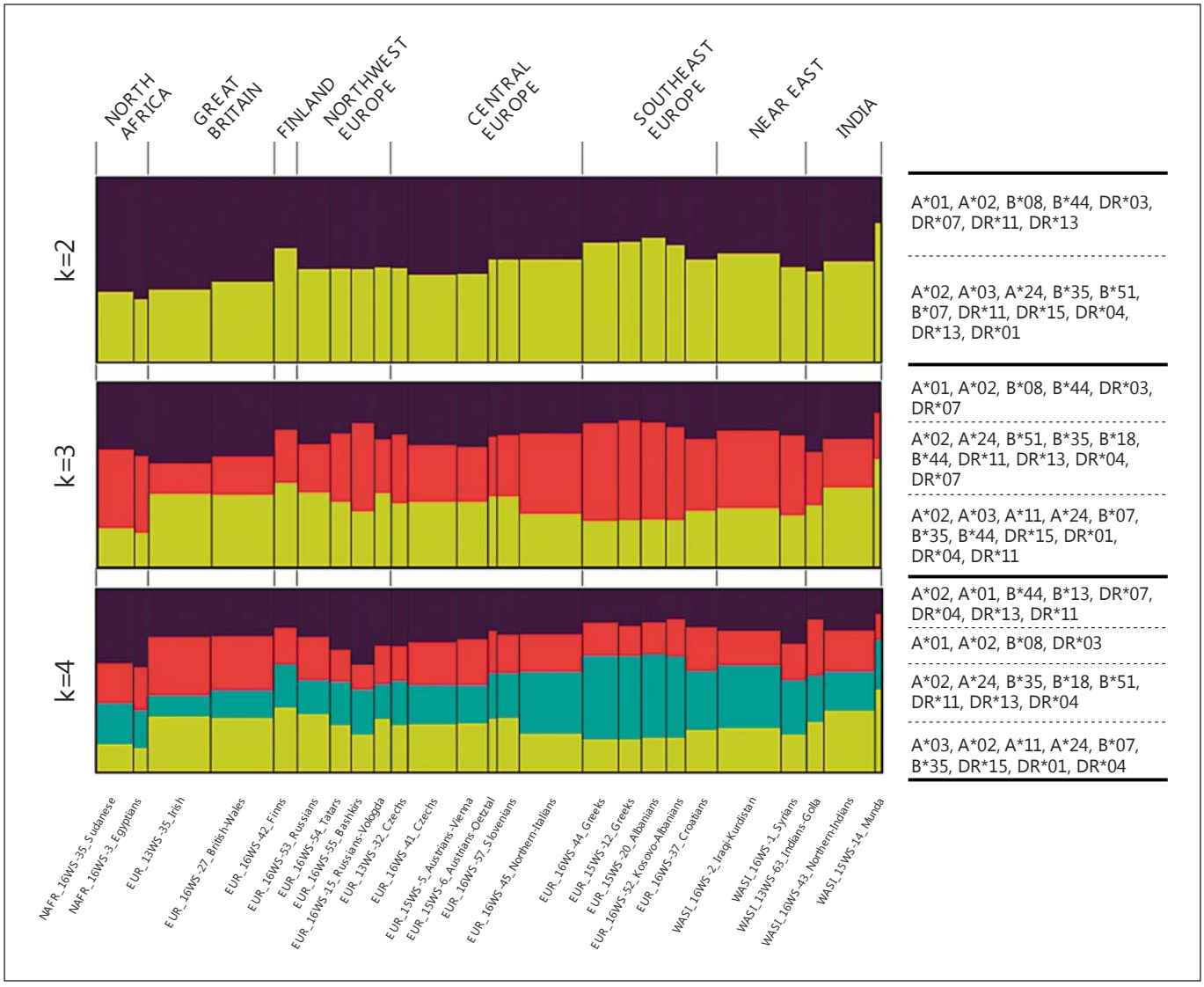


Fig. 2. Estimated proportions of HLA phenotypic profiles (in different colors) inferred for 25 European, North African and West Asian populations for K = 2, K = 3 and K = 4 predefined clusters. Best clustering: K = 4 (see text for explanations). Each population sample is represented by a vertical bar whose width is proportional to the sample size. Alleles with estimated frequencies greater than 10% in each phenotypic profile (details given in online suppl. table S3) are listed on the right by decreasing frequency. Nomenclature used for each sample: Region(here EUR, NAFR or WASI)_Reference-Number_Population name (see online suppl. table S1). Individual profiles are shown in online supplementary figure S6.

typic level. Using a subset of 8,170 individuals belonging to 25 populations tested simultaneously for HLA-A, -B and -DRB1, we retained K = 4 as the best number of phenotypic profiles (see above) inferred from the data. Indeed, with K = 4 the median of ln Pr(X|K) was the highest, the variance on 100 runs was much lower than for K = 5 and K = 6 (online suppl. fig. S4) and the density of this statistic was more regular (online suppl. fig. S5). Figure 2 shows how the 25 populations differ from or resemble each other when K = 2, 3, 4 phenotypic profiles are considered, online supplementary table S3 gives the estimated allele frequencies of each phenotypic profile for K = 2, 3, 4, and online supplementary figure S6 shows the details for all individuals.

A first observation was that most populations exhibit even proportions of all phenotypic profiles for each K (fig. 2), suggesting that the major part of the genetic diversity lies within and not among populations, as repeatedly

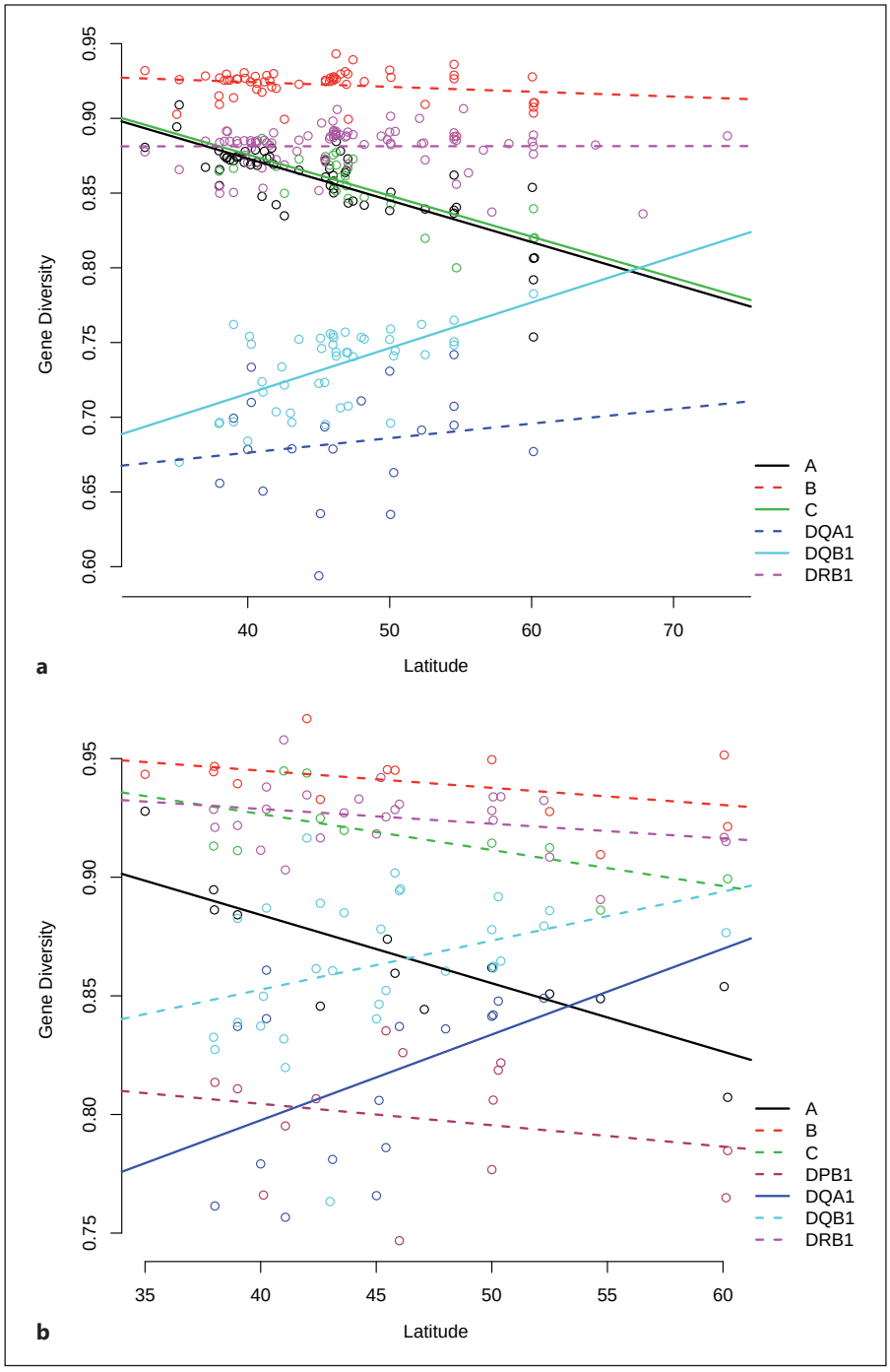


Fig. 3. Linear regression of gene diversity on latitude at the first-field level of resolution (**a**) and at the second-field level of resolution (**b**).

shown for most genetic markers [55–57]. This diversity is also reflected by the fact that within populations, individuals exhibit very diverse phenotypes always matching several phenotypic profiles (among the 4 identified) although with different probabilities (online suppl. fig. S6). However, groups of populations (possibly single populations) can be identified by either greater or smaller proportions of one or several phenotypic profile(s), as described below.

K = 2. The most contrasting proportions (above 60% vs. below 40%) of each phenotypic profile were found between the Indian Munda, several populations from Southeast Europe (2 Albanian and 2 Greek) and the Finns, on

the one side (profile 1 >60%), and populations from North Africa (Egyptian and Sudanese) and Great Britain (Irish and Welsh), on the other side (profile 2 >60%). Frequent co-occurrences of alleles (here listed by decreasing frequencies) A*02, *03, *24, B*35, *51, *07, DRB1*11, *15, *04 and *01 were found in profile 1, and of A*01, *02, B*08, *44, DRB1*03, *07, *13 and *11 in profile 2 (fig. 2).

K = 3. Contrasting proportions (above 45% for one of the phenotypic profiles) were found between the Indian Munda (profile 1 >50%) and the Finns (profile 1 >45%), on the one side, and the same 4 populations from Southeast Europe as for K = 2 (2 Albanian and 2 Greek, profile 2 >50%) and the Bashkir (profile 2 >45%), on the other side. The 2 populations from Great Britain (Irish and Welsh) also exhibited around 40% of profiles 1 and 3, and the populations from North Africa (Egyptian and Sudanese) around 40% of profiles 2 and 3. Profile 1 revealed frequent co-occurrences of A*02, *03, *11, *24, B*07, *35, *44, DRB1*15, *01, *04 and *11, profile 2 of alleles A*02, *24, B*51, *35, *18, *44, DRB1*11, *13, *04 and *07, and profile 3 of A*01, *02, B*08, *44, DRB1*03 and *07 (fig. 2).

K = 4 (Best Clustering). Contrasting proportions (above 45% for one of the phenotypic profiles) were found between the Indian Munda (profile 1 >45%), again the same 4 populations from Southeast Europe (2 Albanian and 2 Greek, profile 2 >45%), the populations from North Africa (Egyptian and Sudanese) and the Bashkir (profile 4 >45%). The Finns also exhibited a large proportion of profile 1 (>35%), the 2 populations from Great Britain (Irish and Welsh) and the 2 populations from North Africa (Egyptian and Sudanese) a very small proportion (<15%) of profile 2 and of profile 1, respectively. The main differences between the 4 profiles were frequent co-occurrences of alleles A*03, *02, *11, *24, B*07, *35, DRB1*15, *01 and *04 for profile 1, of A*02, *24, B*51, *35, *18, DRB1*11, *13 and *04 for profile 2, of A*01, *02, B*08, DRB1*03 for profile 3, and of A*02, *30, B*44, *13, DRB1*07, *04, *13 and *11 for profile 4 (fig. 2). Note that the Bashkir are peculiar for HLA and resemble different populations depending on the number of profiles considered.

Overall, the HLA phenotypic information at the 3 loci A, B and DRB1 revealed that in Europe the 4 profiles defined above are the most distinctive for the populations from Southeastern Europe, Great Britain and Finland. The allele frequencies estimated in each profile for each K are given in online supplementary table S3.

Main Predictors of HLA Genetic Variation in Europe

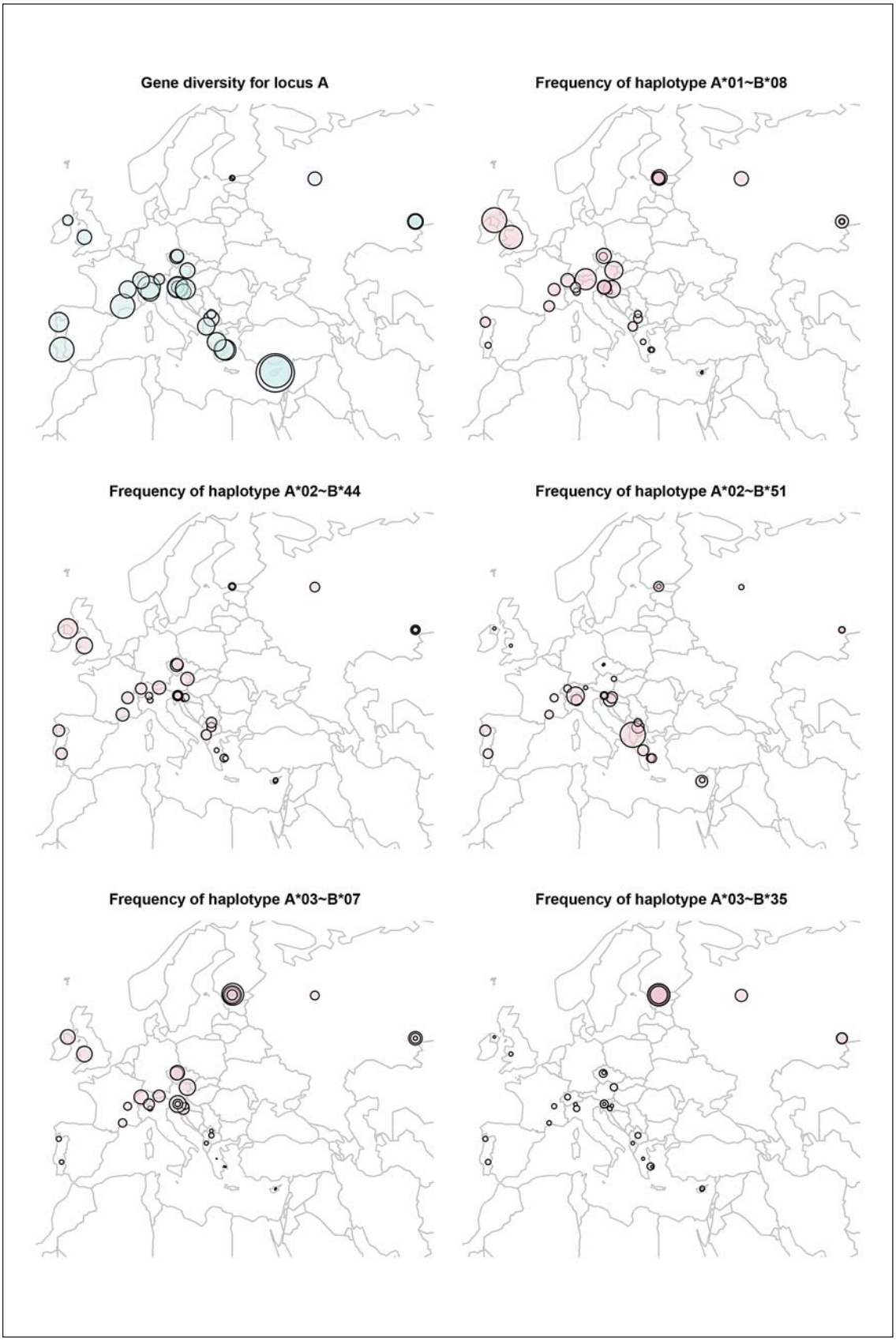
The linear model (regression) approach allowed us to identify 1 main geographic variable – latitude – explaining gene diversity at the HLA-A, -C and -DQB1 loci, at the first-field level, as well as at HLA-A, -DQA1 and -DQB1 loci, at the second-field level of resolution (fig. 3a, b and online suppl. table S4). At the first-field level of resolution (fig. 3a and suppl. table S4a), both locus A and locus C showed a significant negative relationship between gene diversity and latitude (p = 2.87e-15 and 5.43e-6, respectively), and DQB1 showed a significant positive relationship (p = 7.03e-9). Models also including longitude were not significantly better. At the second-field level of resolution (fig. 3b and online suppl. table S4b), only HLA-A showed a highly significant negative correlation between gene diversity and latitude (p = 0.000595), and HLA-DQA1 and -DQB1 showed significant positive relationships (p = 0.00775 and 0.0154, respectively). Longitude again did not significantly improve the model when included (except for DQA1, with a negative relationship, p = 0.02155, results not shown). The only highly significant ($p < 0.01$) associations found at both levels of resolution are thus for locus A with very similar coefficients (–0.00279 and –0.00288, and online suppl. table S4a, b). The heterogeneous associations observed for DQA1 and DQB1 are probably due to a sampling effect and cannot be confirmed as characteristic of each locus.

The map of Europe presented in figure 4 (with geographic coordinates for all populations given in online suppl. table S5) illustrates the latitudinal variation of gene diversity for locus HLA-A (in blue). This figure also shows the frequency variation of the 5 HLA-A-B haplotypes observed in more than 90% of the European populations (in pink), as explained below.

Box-and-whisker plots of LRT p values and PRS quantiles are shown in online supplementary figure S7, with details in online supplementary table S6, for all population samples (i.e. also including NAFR and WASI data). The pairs B-C, DQA1-DQB1 and DRB1-DQB1 are in tight linkage disequilibrium in most populations (in all of them for the 2 last pairs), as expected from their close physical distance. More heterogeneity was observed for the pair B-

Fig. 4. Genetic diversity for locus A and haplotype frequencies for the 5 most common AB haplotypes in Europe. To avoid cluttering the graphs, only 2 samples for Switzerland and Portugal have been included. The 4 Finnish samples were kept because, being located at the same coordinates and being, generally, homogeneous, they did not clutter the graph. The bubbles' sizes are proportional to the haplotype frequencies in the respective populations. Due to the small range of genetic diversity at locus A, these values have been squared to obtain magnitudes in the range from 1 to 10; the bubbles' sizes are proportional to these scaled quantities.

(For figure see next page.)



4

DRB1, and much more for A-B and A-DRB1, which contain the physically most distant loci. The 2 independent approaches used here (LRT and PRS) revealed very concordant results, LRT being more conservative, however.

Overall, global linkage disequilibrium, as measured by LRT p values or PRS quantiles, was too strong to allow mapping linkage disequilibrium in relation to geography. Therefore, as a proxy of a measure of association between loci that would provide significant variation among populations at the European scale, we considered individual linkage disequilibrium measures for the 5 HLA-A-B haplotypes observed in more than 90% of the European samples considered, i.e. for A*01~B*08, A*03~B*07, A*02~B*44, A*02~B*51 and A*03~B*35. Their frequencies are shown in figure 4, their regression with latitude and longitude in online supplementary table S4c, and the results of the linkage disequilibrium analysis in online supplementary figure S8. Haplotype frequencies revealed marked and relevant regional variation towards higher frequencies of A*01~B*08 and A*03~B*07 in Central and Northwestern Europe, slightly higher frequencies of A*02~B*44 in Central and Western Europe, higher frequencies of A*02~B*51 in Southeastern Europe as well as high frequencies of A*03~B*35 in Finland and in Russia. A regression analysis confirmed the significant correlations with latitude for all these haplotypes (negative only for A*02~B*51) except A*02~B*44 and revealed a significant correlation with longitude (with negative interaction between latitude and longitude) for A*01~B*08 (online suppl. table S4c).

The conclusions are less straightforward for linkage disequilibrium assessed by standardized residuals (online suppl. fig. S8). While the histograms and densities of these residuals suggest significant linkage disequilibrium in many populations for the 5 haplotypes, only A*01~B*08 always showed significant deviations, and no specific geographic patterns could be deduced from this analysis (results not shown). This is also true when exploring standardized D' values; actually, D' values appeared to be strongly associated with haplotype frequencies (results not shown) which suggests that they are not suitable to explore linkage disequilibrium among populations.

Discussion

This study presents the first detailed analysis of HLA molecular variation in European populations based on information provided by a total of 7 loci (A, B, C, DRB1, DQA1, DQB1 and DPB1). This large meta-analysis has been possible thanks to a compilation of data collected within the frame of several research programs, i.e. 4 International Histocompatibility Workshops and the European network HLA-NET, for which detailed allele and haplotype frequencies and summary statistics are provided in a companion paper [33]. European populations were not only considered alone in this study but also in comparison with some populations from 2 neighboring areas, North Africa and West Asia (as far as India), to better understand European variation in a broader geographic context. Also, despite unequal sets of population samples available, the comparison of the results obtained for the different HLA loci was expected to clarify the role of specific evolutionary mechanisms, i.e. different kinds of natural selection, acting on them.

How HLA Varies across Europe

Our results first indicate that for 5 of the 7 loci studied, i.e. HLA-A, -B, -C, -DRB1 and -DPB1, a very strong predictor of HLA genetic variation in Europe is the geographic distance between populations (table 1). Interestingly, the greatest correlation coefficients were found at the first-field level for HLA-A, -B and -C, suggesting that HLA low-resolution data may be as informative as high-resolution data to investigate population relationships (at least at the current state of research where data fully typed at high resolution are still limited). We also showed for the first time that these 5 HLA loci exhibited very similar diversity patterns (close F_{ST} values and NMDS configurations), allowing us to plot composite NMDS (fig. 1a, b). These plots, the details of which are given in [33], finely illustrate the regional genetic diversity of HLA in Europe. They show, in agreement with the results obtained by means of analysis of variance, that populations located in the Southeast widely differ, genetically, from populations living in Northern and, to a lesser extent, Central-Western regions and that Southeastern European resemble West Asian populations with which they are in geographic proximity (online suppl. fig. S3a, b). Therefore, rather than following a pure model of isolation-by-distance (as we first deduced from the significant correlations between genetic and geographic distances), HLA frequencies appear to vary along a general south-southeast to north-northwest axis, the latter fitting quite well the clinal models retained for many other markers in Europe [20, 21]. Geographic clines were indeed confirmed for many HLA alleles on the basis of spatial autocorrelation analyses, the resulting correlograms being interpreted according to Barbujani [58]. A few examples are given in online supplementary figure S9. Note, however, that some alleles follow a 'depression' rather than a clinal pattern (a few

examples are also given in online suppl. fig. S9), suggesting that their highest (or lowest) frequency is centered in some place of Europe and decreases (or increases) all around; a further exploration of these results will indicate whether such irregularities are due to outlier populations (e.g. having undergone rapid genetic drift due to isolation) or to environmental factors (e.g. disease prevalence or pathogen richness) affecting specific HLA frequencies. Moreover, both gene diversity at locus A and the frequencies of several HLA-A-B haplotypes, among the most widely represented in Europe [like A*01~B*08 and A*03~B*07, initially remarked by Degos and Dausset 59], also follow this general trend (fig. 4). Again, autocorrelation analyses performed, this time, on heterozygosity values confirmed the observed cline for HLA-A, and, to a lesser extent, for HLA-B and -C (online suppl. fig. S10), which is in agreement with the results of our regression analyses. Finally, Great Britain, Finland and Southeastern Europe, which are located at the margins of this geographic area, exhibited to most distinctive proportions of the 4 main HLA phenotypic profiles observed in Europe (fig. 2). All these results, thus, revealed a directional pattern which was not obscured by other processes, as previously observed for classical polymorphisms [60], and which may have been caused by population migrations.

The Signatures of Demography

As mentioned in the Introduction section, one main question arises of course from the results presented above: has the observed HLA variation been shaped by natural selection in Europe? While all specialists agree that natural selection drives the evolution of HLA genes, it is still not clear to what extent it has affected the patterns of HLA differentiations among populations. For example, Meyer et al. [61] did not find any significant difference between the patterns of interpopulation differentiations at HLA compared to putatively neutral loci; Currat et al. [62] explained the peculiar HLA-DRB1 differentiation patterns observed in the Gibraltar Strait region by a significant but relatively weak effect of natural selection (s = 2.2%) compared to other well-known selected alleles [e.g. lactase persistence 63], and this effect was no more significant in Southwestern Europe when this region was considered separately from Northwestern Africa. In the present study, the only locus showing a significant decrease of gene diversity along the latitude was HLA-A, a locus which is less affected by diversifying selection compared to B and DRB1 [current results and 27, 64, 65]. This pattern is, thus, better explained by past populations moving northwards with an origin in the Southeast (gene diversity is extreme in Cyprus, close to the Near East) than by unequal intensities of diversifying selection, because in the case of selection, we would expect a similar or even more pronounced pattern following the latitude for loci B and DRB1. Note, however, that we were not in a position here to determine when (e.g. in Neolithic, Paleolithic or other periods) such putative migrations occurred.

Actually, these results are congruent with the principal conclusions drawn on the basis of genome-wide studies. Lao et al. [4] analyzed 309,790 SNPs in 23 European populations and found that although the amount of differentiation among them was small, the existing differences correlated well with geographic distances. This is similar to what we observed for HLA, and more particularly so at loci HLA-A, -B, -C and -DRB1: the F_{ST} values were small (between 0.004 and 0.019, online suppl. table S2), but the correlation coefficients between the genetic and geographic distances were relatively high (between 0.329 and 0.638, table 2) and significant (p values between 0.044 and 1e-4). Lao et al. [4] also observed a larger mean heterozygosity in the south than in the north of Europe; this is exactly what we found for locus HLA-A, which showed a highly significant negative correlation between gene diversity and latitude (fig. 3a, b and online suppl. tables S4a, b). Novembre et al. [7] also analyzed 197,146 SNPs in 1,387 individuals from 37 European countries and showed a notable resemblance of their principal component analysis to the geographic map of Europe. In addition, they concluded that the direction of the first principal component axis, which aligns south-southeast to north-northwest (with a correlation r^2 vs. latitude of 0.71) and accounts for approximately twice the amount of variation of the second axis (the latter correlating with longitude), reflects a special role in the demographic history of Europeans. In the present study, HLA-A, which provided the strongest geographic signal, also revealed a directional pattern along a south-southeast-to-north-northwest axis, with a very similar correlation between the first axis of the NMDS and latitude ($r^2 = 0.74$, $p < 2.2e\text{-}16$), the second axis being also correlated, but to a lesser extent, with longitude ($r^2 = 0.22$, $p = 2e\text{-}4$). Finally, Novembre et al. [7] also observed a decrease of haplotype diversity from south to north. The overall results found for HLA at the European level, and in particular at locus HLA-A, are thus strikingly similar to those based on genome-wide studies at the same geographic scale. This suggests that the amount of information provided by this single HLA locus, the allelic diversity of which is very high, is sufficient to describe the same general patterns as those re-

vealed by hundreds of thousands of biallelic SNPs of the entire genome, such patterns being interpreted as signatures of human migration history. Of course the quality of our data is probably of major importance here, as we were working (for HLA-A data tested at the first-field level of resolution) with 105,873 individuals representing 58 European populations, which represents a huge amount of information compared to previous studies.

Other specific demographic effects may also satisfactorily explain the regional differences observed in Europe: the marked differentiation between Southeastern Europeans, on the one side, and Northern and Central-Western Europeans, on the other side (fig. 1), may have resulted from reduced gene flow across the Alps. Indeed, a significant genetic boundary has previously been detected in this region through HLA comparisons, both at the continental scale [36] and at the country level [i.e. Switzerland, 66]. Interestingly, a study based on the Y-chromosome polymorphism is congruent with this hypothesis, central Europe (including the Alps) being detected as a zone of sharp genetic contrasts within the continent [67]. The more pronounced genetic differentiation of several Northern populations, i.e. from Great Britain and Finland, may also be due to a greater isolation of these regions which are separated from the continent by the sea (the English Channel and the Baltic Sea, respectively); in the case of Finland, we may also explain it by the specific origin of the Finno-Ugric-speaking populations (Uralic linguistic family) between the Baltic Sea and the Ural Mountains [68], compared to Indo-European-speaking populations which probably originated in Anatolia [69]. We thus think that the HLA-A locus (including haplotypes strongly associated with it) is particularly informative to reconstruct human peopling history in Europe despite possible effects of natural selection.

The Marks of Natural Selection

By contrast, the HLA-B, -C, -DRB1 and -DPB1 loci probably lost some of the genetic signatures related to demographic events. For B, C and DRB1, stronger heterozygous advantages would have homogenized gene frequencies or prevented differentiation among populations, as has been observed in other species [70]. Similarly, the persistence time of the HLA polymorphism, which is at least partially shared among populations as a consequence of the long-lasting effect of balancing selection, could influence the values of Wright's F statistics and thus affect the inferences made on migrations (of interest for this issue, some authors have also proposed that, while many HLA lineages are old and have been inherited in a trans-specific fashion, the alleles within these lineages are the result of a recent diversification [71, 72]). However, to investigate such a putative effect would require complete HLA sequence data and a detailed comparison of different regions of these genes in many populations, for instance the exons coding for the peptide-binding region where evidence of selection has been observed versus other exons and intronic regions suggested to evolve under (or close to) neutral conditions. Unfortunately, looking for sequence, the currently available population data are at best very incomplete and do not permit such an analysis. In the present study, it is of note, however, that the slope of the regression line of gene diversity on latitude is also negative for the 3 loci mentioned above (fig. 3), as a possible relic of northward migrations; it is even significant for HLA-C at the first-field level of resolution, which could be expected from the smaller proportion of populations rejecting neutrality at this locus (table 1).

Besides the 4 loci mentioned above, HLA-DPB1, -DQA1 and -DQB1 are particular. DPB1 deviates from selective neutrality, but towards an excess of homozygotes (table 1). The large proportion of such deviations found in the present study (almost 50% of the populations tested) is a rather unexpected result compared to previous estimations [27, 65]. Actually, new associations of specific markers located in the HLA-DP region to some severe diseases (like chronic hepatitis B) are being discovered [73], and hitchhiking effects on HLA-DPB1 alleles may have occurred. Interestingly, however, the correlation with geography is very high (table 2) for this locus and the slope of the regression line of gene diversity on latitude is again negative (fig. 3b) despite a very different kind of natural selection acting on this locus; this again favors demography rather than selection being the cause of the latitudinal variation.

By contrast, opposite patterns are found for HLA-DQA1 and -DQB1 (fig. 3). While these 2 loci exhibit an excess of heterozygotes (table 1), they probably evolved according to more complex mechanisms, discussed elsewhere, in relation to the PDBS hypothesis [74], which would explain the higher F_{ST} values observed (online suppl. table S2). Moreover, HLA-DQA1 and -DQB1 alleles, which are in strong linkage disequilibrium (online suppl. table S6 and online suppl. fig. S7), are associated with specific diseases in a concerted way, as illustrated by the well-known example of coeliac disease resulting from the presentation of the gluten-derived gliadin peptide by HLA-DQ2 heterodimers (DQA1*0501- and DQB1*0201-

associated chains) [75]. Together with the low (and often nonsignificant) correlations between the genetic and geographic distances (table 2), the results found for the 2 DQ loci suggest a predominant effect of selective forces over demography, in agreement with other studies [27, 65]. Therefore, the lower DQA1 and DQB1 gene diversity observed in Southern Europe (fig. 3) is probably the result of some strong directional or purifying selection, its precise causes remaining to be explored.

Implications of European Regional Genetic Diversity for Human Health

Whereas the patterns of multi-locus HLA haplotype frequencies are generally difficult to describe and analyze (and also to interpret, as mentioned in the Results section for A-B haplotypes), the present study proposes an original approach (never applied to HLA before) to analyze HLA regional variation in Europe: by identifying several common phenotypic profiles at 3 HLA loci, i.e. A, B and DRB1, in this continent. Four main profiles have been identified, with varying proportions in different geographic regions (fig. 2). They not only confirmed previously detected patterns of allelic variation, like high A*01, *03, B*07 and *08 frequencies in Northern and Central-Western Europe and A*24, B*18, *35, *51 and DRB1*11 in Southeastern Europe [23]; they also indicated which allelic combinations at the 3 loci A, B and DRB1 are the most common in different geographic regions, thus providing an invaluable information. This knowledge is indeed crucial for donor search in tissue transplantation, in particular because new efforts are currently undertaken to optimize graft compatibility (e.g. for kidney transplantation) through HLA matching at the phenotypic rather than haplotypic level, as proposed by the EUROSTAM program (www.eurostam.eu). Once applied to more data defined at high resolution, this analysis will provide key references for the search of compatible allelic combinations, something which other approaches focusing on haplotype frequencies only partially supply. This will also stimulate the development of specific algorithms using the phenotypic profiles observed in Europe for the optimization of donor-recipient HLA matching.

Finally, in another domain also related to health, i.e. disease-association studies, the regional HLA diversity that we have observed in Europe suggests that more attention has to be devoted to the populations taken as reference in such studies. Population stratification may indeed lead to spurious associations, as already stressed for variable DNA sites [7], although this remains to be formally demonstrated. Very often, 'Caucasian' samples are used as controls; but as European populations are genetically diverse, the term 'Caucasian' – used to describe indifferently any population from Europe and its surrounding regions – should be abandoned as it does not correspond to any scientific reality, and the use of data defined in this way should be banned [35, 76]. This would certainly improve the robustness of the associations found, as the majority of them are currently observed in specific studies and not confirmed by others.

Conclusions

Thirty-five years after the publication of the paper entitled 'An HLA Map of Europe' in this journal [28], the present study provides a new and original analysis of molecular variation at 7 HLA loci in European populations, in relation to their geographic location. While some seminal results remain valid (the high frequency and peculiar geographic variation of the famous HLA-A*01~B*08 and A*03~B*07 haplotypes), HLA diversity, here explored in more details, reveals both regional population differentiations (more pronounced between Southeastern and the rest of Europe) and a common trend of decreasing diversity with increasing latitude for 5 loci, the strongest signal being observed for HLA-A. Actually, the main information provided by this locus is equivalent to that revealed by genome-wide studies at a comparable geographic scale. We also provide for the first time estimated proportions, in each population, of the main HLA-A, B, DRB1 phenotypic profiles observed in Europe. Overall, this new HLA map of Europe provides key references for studies related to European genetic diversity, not only for research in population genetics, but also for important domains in medicine and essential clinical applications.

Acknowledgments

This work was supported by the Swiss National Science Foundation (SNSF grants No. 31003A_144180 and 10CO11_145293), the European COST Action BM0803 'HLA-NET' and the FP-7-Health-2012 program (grant No. 305385 EUROSTAM). We thank all participants of HLA-NET and of the AHPD project of the 15th and 16th International Histocompatibility and Immunogenetics Workshops for their collaboration, as well as Stephan Weber for his technical help. We also thank two anonymous reviewers for their insightful suggestions on a previous version of this paper.

References

1 Balaresque P, Bowden GR, Adams SM, Leung HY, King TE, Rosser ZH, Goodwin J, Moisan JP, Richard C, Millward A, Demaine AG, Barbujani G, Previdere C, Wilson IJ, Tyler-Smith C, Jobling MA: A predominantly neolithic origin for European paternal lineages. PLoS Biol 2010;8:e1000285.

2 Barbujani G, Bertorelle G: Genetics and the population history of Europe. Proc Natl Acad Sci USA 2001;98:22–25.

3 Chikhi L, Nichols RA, Barbujani G, Beaumont MA: Y genetic data support the neolithic demic diffusion model. Proc Natl Acad Sci USA 2002;99:11008–11013.

4 Lao O, Lu TT, Nothnagel M, et al: Correlation between genetic and geographic structure in Europe. Curr Biol 2008;18:1241–1248.

5 Rosser ZH, Zerjal T, Hurles ME, et al: Y-chromosomal diversity in Europe is clinal and influenced primarily by geography, rather than by language. Am J Hum Genet 2000;67:1526–1543.

6 Simoni L, Calafell F, Pettener D, Bertranpetit J, Barbujani G: Geographic patterns of mtDNA diversity in Europe. Am J Hum Genet 2000;66:262–278.

7 Novembre J, Johnson T, Bryc K, Kutalik Z, Boyko AR, Auton A, Indap A, King KS, Bergmann S, Nelson MR, Stephens M, Bustamante CD: Genes mirror geography within Europe. Nature 2008;456:98–101.

8 Richards M, Macaulay V, Torroni A, Bandelt HJ: In search of geographical patterns in European mitochondrial DNA. Am J Hum Genet 2002;71:1168–1174.

9 Barbujani G: Genetic evidence for prehistoric demographic changes in Europe. Hum Hered 2013;76:133–141.

10 Deguilloux MF: Ancient DNA: a window to the past of Europe. Hum Hered 2013;76:121–132.

11 Sabeti PC, Varilly P, Fry B, et al: Genome-wide detection and characterization of positive selection in human populations. Nature 2007;449:913–918.

12 Lao O, de Gruijter JM, van Duijn K, Navarro A, Kayser M: Signatures of positive selection in genes associated with human skin pigmentation as revealed from analyses of single nucleotide polymorphisms. Ann Hum Genet 2007;71:354–369.

13 Gerbault P, Liebert A, Itan Y, Powell A, Currat M, Burger J, Swallow DM, Thomas MG: Evolution of lactase persistence: an example of human niche construction. Philos Trans R Soc Lond B Biol Sci 2011;366:863–877.

14 Gerbault P, Moret C, Currat M, Sanchez-Mazas A: Impact of selection and demography on the diffusion of lactase persistence. PLoS One 2009;4:e6369.

15 Parham P: The Immune System, ed 3. London and New York, Garland Science, 2009.

16 Mack SJ, Cano P, Hollenbach JA, et al: Common and well-documented HLA alleles: 2012 update to the CWD catalogue. Tissue Antigens 2013;81:194–203.

17 Robinson J, Halliwell JA, McWilliam H, Lopez R, Parham P, Marsh SG: The IMGT/HLA database. Nucleic Acids Res 2013;41:D1222–1227.

18 Meyer D, Thomson G: How selection shapes variation of the human major histocompatibility complex: a review. Ann Hum Genet 2001;65:1–26.

19 Fix AG: Gene frequency clines in Europe: demic diffusion or natural selection? J R Anthropol Inst 1996;2:625–643.

20 Menozzi P, Piazza A, Cavalli-Sforza L: Synthetic maps of human gene frequencies in Europeans. Science 1978;201:786–792.

21 Cavalli-Sforza L, Menozzi P, Piazza A: The History and Geography of Human Genes. Princeton, Princeton University Press, 1994.

22 Riccio ME, Buhler S, Nunes JM, et al: 16th IHIW: analysis of HLA population data, with updated results for 1996 to 2012 workshop data (AHPD project report). Int J Immunogenet 2013;40:21–30.

23 Nunes JM, Riccio ME, Buhler S, et al: Analysis of the HLA population data (AHPD) submitted to the 15th International Histocompatibility/Immunogenetics Workshop by using the Gene[rate] computer tools accommodating ambiguous data (AHPD project report). Tissue Antigens 2010;76:18–30.

24 Mack SJ, Sanchez-Mazas A, Meyer D, Single RM, Tsai Y, Erlich HA: 13th International Histocompatibility Workshop Anthropology/Human Genetic Diversity Joint Report. Chapter 2: Methods used in the generation and preparation of data for analysis in the 13th International Histocompatibility Workshop; in Hansen JA (ed): Immunobiology of the Human MHC: Proceedings of the 13th International Histocompatibility Workshop and Conference. Seattle, IHWG Press, 2007, vol 1, pp 564–579.

25 Clayton J, Lonjou C, Whittle D: Allele and haplotype frequencies for HLA loci in various ethnic groups; in Charron D (ed): Genetic Diversity of HLA: Functional and Medical Implication. Proceedings of the 12th International Histocompatibility Workshop and Conference (Paris, June 1996). Paris, EDK, 1997, vol 1, pp 665–820.

26 Imanishi T, Akaza T, Kimura A, Tokunaga K, Gojobori T: Allele and haplotype frequencies for HLA and complement loci in various ethnic groups; in Tsuji K, Aizawa M, Sasazuki T (eds): HLA 1991. Oxford, Oxford University Press, 1992, vol 1, pp 1065–1220.

27 Solberg OD, Mack SJ, Lancaster AK, Single RM, Tsai Y, Sanchez-Mazas A, Thomson G: Balancing selection and heterogeneity across the classical human leukocyte antigen loci: a meta-analytic review of 497 population studies. Hum Immunol 2008;69:443–464.

28 Ryder LP, Andersen E, Svejgaard A: An HLA map of Europe. Hum Hered 1978;28:171–200.

29 Pereyra F, Jia X, McLaren PJ, et al: The major genetic determinants of HIV-1 control affect HLA class I peptide presentation. Science 2010;330:1551–1557.

30 Blackwell JM, Jamieson SE, Burgner D: HLA and infectious diseases. Clin Microbiol Rev 2009;22:370–385.

31 Trowsdale J, Knight JC: Major histocompatibility complex genomics and human disease. Annu Rev Genomics Hum Genet 2013;14:301–323.

32 Tiercy JM, Claas FHJ: Impact of HLA diversity on donor selection in organ and stem cell transplantation. Hum Hered 2013;76:178–186.

33 Nunes JM, Buhler S, Roessli D, et al: The HLA-net GENE[RATE] pipeline for effective HLA data analysis and its application to 145 population samples from Europe and neighbouring areas. Tissue Antigens 2014;83:307–323.

34 Bodmer J, Cambon-Thomsen A, Hors J, Piazza A, Sanchez-Mazas A: Anthropology report. Introduction; in Charron D (ed): Genetic Diversity of HLA: Functional and Medical Implication. Proceedings of the 12th International Histocompatibility Workshop and Conference (Paris, June 1996). Paris, EDK, 1997, vol 1, pp 269–284.

35 Sanchez-Mazas A, Vidan-Jeras B, Nunes JM, et al: Strategies to work with HLA data in human populations for histocompatibility, clinical transplantation, epidemiology and population genetics: HLA-net methodological recommendations. Int J Immunogenet 2012;39:459–476.

36 Buhler S, Megarbane A, Lefranc G, Tiercy JM, Sanchez-Mazas A: HLA-C molecular characterization of a Lebanese population and genetic structure of 39 populations from Europe to India-Pakistan. Tissue Antigens 2006;68:44–57.

37 Hors J, El Chenawi F, Djoulah S, Hafez M, Abbas F, El Borai MH, Kamel M, Abbal M, Cambon-Thomsen A, Mercier P, Reviron D, Magzoub MA, Rosner G, Delgado JC, Yunis E, Raffoux C, Tamouza R, Izaabel H, Hmida S, Benhamamouch S, Bessaoud K, Langaney A, Sanchez-Mazas A: HLA in North African populations: 12th international Histocompatibility Workshop: NAFR report; in Charron D (ed): Genetic Diversity of HLA: Functional and Medical Implication. Proceedings of the 12th International Histocompatibility Workshop and Conference (Paris, June 1996). Paris, EDK, 1997, vol 1, pp 328–334.

38 Abdennaji Guenounou B, Loueslati BY, Buhler S, Hmida S, Ennafaa H, Khodjet-Elkhil H, Moojat N, Dridi A, Boukef K, Ben Ammar Elgaaied A, Sanchez-Mazas A: HLA class II genetic diversity in Southern Tunisia and the Mediterranean area. Int J Immunogenet 2006;33:93–103.

39 Buhler S, Sanchez-Mazas A, Zanone R, Djavad N, Tiercy JM: PCR-SSOP molecular typing of HLA-C alleles in an Iranian population. Tissue Antigens 2002;59:525–530.

40 Nunes JM: Tools for analysing ambiguous HLA data. Tissue Antigens 2007;69:203–205.
41 Nunes JM: Generate: Tools for Analysis and Handling of Data with Ambiguities. Geneva, Switzerland, Laboratory of Anthropology, Genetics and Peopling History, University of Geneva, 2006.
42 Nunes JM, Riccio ME, Tiercy JM, Sanchez-Mazas A: Allele frequency estimation from ambiguous data: using resampling schema in validating frequency estimates and in selective neutrality testing. Hum Biol 2011;83:437–447.
43 Bonferroni C: Teoria statistica delle classi e calcolo delle probabilità. Pubblicazioni del R Istituto Superiore di Scienze Economiche e Commerciali di Firenze 1936;8:3–62.
44 Benjamini Y, Hochberg Y: Controlling the false discovery rate: a practical and powerful approach to multiple testing. J R Stat Soc Series B Stat Methodol 1995;57:289–300.
45 R-core-team: R: A Language and Environment for Statistical Computing. Vienna, Austria, R Foundation for Statistical Computing, 2013.
46 Cullen M, Perfetto S, Klitz W, Nelson G, Carrington M: High-resolution patterns of meiotic recombination across the human major histocompatibility complex. Am J Hum Genet 2002;71:759–776.
47 Prevosti A, Ocaña J, Alonso G: Distances between populations of *Drosophila subobscura*, based on chromosome arrangement frequencies. Theor Appl Genet 1975;45:231–241.
48 Kruskal J: Nonmetric multidimensional scaling: a numerical method. Psychometrika 1964;29:115–129.
49 Mantel N: The detection of disease clustering and a generalized regression approach. Cancer Res 1967;27:209–220.
50 Oksanen J, Blanchet FG, Kindt R, Legendre P, Minchin PR, O'Hara RB, Simpson GL, Solymos P, Stevens MHM, Wagner H: Package 'vegan': community ecology package, 2012. http://cran.r-project.org, http://vegan.r-forge.r-project.org.
51 Dray S, Dufour AB: The ade4 package: implementing the duality diagram for ecologists. J Stat Softw 2007;22:1–20.
52 Excoffier L, Laval G, Schneider S: Arlequin (version 3.0): an integrated software package for population genetics data analysis. Evol Bioinform Online 2005;1:47–50.
53 Pritchard JK, Stephens M, Donnelly P: Inference of population structure using multilocus genotype data. Genetics 2000;155:945–959.
54 Rosenberg NA: Distruct: a program for the graphical display of population structure, 2007. Molec Ecol Notes 2004;4:137–138.
55 Lewontin RC: The apportionment of human diversity; in Dobzhansky T, Hecht MK, Steere WC (eds): Evolutionary Biology 6. New York, Appleton-Century-Crofts, 1972, pp 381–398.
56 Barbujani G, Magagni A, Minch E, Cavalli-Sforza LL: An apportionment of human DNA diversity. Proc Natl Acad Sci USA 1997;94:4516–4519.
57 Sanchez-Mazas A: An apportionment of human HLA diversity. Tissue Antigens 2007;69(suppl 1):198–202.
58 Barbujani G: Geographic patterns: how to identify them and why. Hum Biol 2000;72:133–153.
59 Degos L, Dausset J: Human migrations and linkage disequilibrium of HLA system. Immunogenetics 1974;3:195–210.
60 Sokal RR, Menozzi P: Spatial autocorrelations of HLA frequencies in Europe support demic diffusion of early farmers. Am Nat 1982;119:1–17.
61 Meyer D, Single RM, Mack SJ, Erlich HA, Thomson G: Signatures of demographic history and natural selection in the human major histocompatibility complex loci. Genetics 2006;173:2121–2142.
62 Currat M, Poloni ES, Sanchez-Mazas A: Human genetic differentiation across the strait of Gibraltar. BMC Evol Biol 2010;10:237.
63 Gerbault P: The onset of lactase persistence in Europe. Hum Hered 2013;76:154–161.
64 Satta Y, O'Huigin C, Takahata N, Klein J: Intensity of natural selection at the major histocompatibility complex loci. Proc Natl Acad Sci USA 1994;91:7184–7188.
65 Buhler S, Sanchez-Mazas A: HLA DNA sequence variation among human populations: molecular signatures of demographic and selective events. PLoS One 2011;6:e14643.
66 Buhler S, Nunes JM, Nicoloso G, Tiercy JM, Sanchez-Mazas A: The heterogeneous HLA genetic makeup of the Swiss population. PLoS One 2012;7:e41400.
67 Rosser ZH, Zerjal T, Hurles ME, et al: Y-chromosomal diversity in Europe is clinal and influenced primarily by geography, rather than by language. Am J Hum Genet 2000:67:1526–1543.
68 Campbell L: Historical Linguistics: An Introduction. Edinburgh, Edinburgh University Press, 2004.
69 Bouckaert R, Lemey P, Dunn M, Greenhill SJ, Alekseyenko AV, Drummond AJ, Gray RD, Suchard MA, Atkinson QD: Mapping the origins and expansion of the Indo-European language family. Science 2012;337:957–960.
70 Mona S, Crestanello B, Bankhead-Dronnet S, Pecchioli E, Ingrosso S, D'Amelio S, Rossi L, Meneguz PG, Bertorelle G: Disentangling the effects of recombination, selection, and demography on the genetic variation at a major histocompatibility complex class II gene in the alpine chamois. Mol Ecol 2008;17:4053–4067.
71 Bergström TF, Josefsson A, Erlich HA, Gyllensten U: Recent origin of HLA-DRB1 alleles and implications for human evolution. Nat Genet 1998;18:237–242.
72 von Salome J, Gyllensten U, Bergström TF: Full-length sequence analysis of the HLA-DRB1 locus suggests a recent origin of alleles. Immunogenetics 2007;59:261–271.
73 Wong DK, Watanabe T, Tanaka Y, Seto WK, Lee CK, Fung J, Lin CK, Huang FY, Lai CL, Yuen MF: Role of HLA-DP polymorphisms on chronicity and disease activity of hepatitis B infection in Southern Chinese. PLoS One 2013;8:e66920.
74 Sanchez-Mazas A, Lemaitre JF, Currat M: Distinct evolutionary strategies of human leucocyte antigen loci in pathogen-rich environments. Philos Trans R Soc Lond B Biol Sci 2012;367:830–839.
75 Kim CY, Quarsten H, Bergseng E, Khosla C, Sollid LM: Structural basis for HLA-DQ2-mediated presentation of gluten epitopes in celiac disease. Proc Natl Acad Sci USA 2004;101:4175–4179.
76 Nunes JM, Buhler S, Sanchez-Mazas A: No to obsolete definitions, yes to blanks. Tissue Antigens 2014;83:119–120.