Immunogenetics (2014) 66:153–160
DOI 10.1007/s00251-014-0757-6

ORIGINAL PAPER

HLA variation reveals genetic continuity rather than population group structure in East Asia

Da Di · Alicia Sanchez-Mazas

Received: 14 November 2013 / Accepted: 7 January 2014 / Published online: 22 January 2014

Abstract Genetic differences between Northeast Asian (NEA) and Southeast Asian (SEA) populations have been observed in numerous studies. At the among-population level, despite a clear north–south differentiation observed for many genetic markers, debates were led between abrupt differences and a continuous pattern. At the within-population level, whether NEA or SEA populations have higher genetic diversity is also highly controversial. In this study, we analyzed a large set of HLA data from East Asia in order to map the genetic variation among and within populations in this continent and to clarify the distribution pattern of HLA lineages and alleles. We observed a genetic differentiation between NEA and SEA populations following a continuous pattern from north to south, and we show a significant and continuous decrease of HLA diversity by the same direction. This continuity is shaped by clinal distributions of many HLA lineages and alleles with increasing or decreasing frequencies along the latitude. These results bring new evidence in favor of the "overlapping model" proposed previously for East Asian peopling history, whereby modern humans migrated eastward from western Eurasia via two independent routes along each side of the Himalayas and, later, overlapped in East Asia across open land areas. Our study strongly suggests that intensive gene flow between NEA and SEA populations occurred and shaped the latitude-related continuous pattern of genetic variation and the peculiar HLA lineage and allele distributions observed in this continent. Probably for a very long period, the exact duration of these events remains to be estimated.

Keywords East Asia · Genetic continuity · HLA variation · Human peopling history · North–south differentiation

Introduction

East Asia is one of the most populated geographic regions of the world, being the homeland of more than 1.5 billion people in China, Korea, Japan, and Mongolia. Genetic variation among East Asian populations has raised a great interest among scholars of different disciplines, notably physical anthropologists and geneticists (Cavalli-Sforza et al. 1994), as a north–south differentiation was suggested from the analysis of both physical traits (Turner 1987; Zhang 1988) and genetic markers (Zhang et al. 2007). These studies led to consider two population groups, Northeast Asian (NEA) and Southeast Asian (SEA), with the Yangtze River generally taken as a geographic boundary between the two (Zhang et al. 2007). Nevertheless, this question remained controversial, as some studies proposed abrupt differences across a highly significant genetic boundary (Xue et al. 2005; Xue et al. 2008; Zhao and Lee 1989) or the existence of two differentiated neighboring groups (Chu et al. 1998; Du et al. 1997), while others defended the existence of a continuous pattern along the latitude (Chen et al. 2009; Karafet et al. 2001; Poloni et al. 2005; Sanchez-Mazas et al. 2005).

Differences between NEA and SEA populations were also found in the amount of genetic variation within populations. Using Y-chromosome and mtDNA markers, some authors claimed that SEA populations were more diverse because the haplogroups found in NEA populations were a subset of those

Electronic supplementary material The online version of this article (doi:10.1007/s00251-014-0757-6) contains supplementary material, which is available to authorized users.

D. Di (✉) · A. Sanchez-Mazas
Laboratory of Anthropology, Genetics and Peopling History (AGP lab), Anthropology Unit, Department of Genetics and Evolution, University of Geneva, 12 rue Gustave-Revilliod, CH-1211 Geneva, Switzerland
e-mail: da.di@unige.ch

A. Sanchez-Mazas
Institude of Genetics and Genomics in Geneva (IGE3), University of Geneva Medical Center (CMU), 1 rue Michel-Servet,
CH-1211 Geneva, Switzerland

observed in SEA populations, and because more ancestral mutations were found in the South (Shi et al. 2005; Shi et al. 2008; Su et al. 1999; Yao et al. 2002). However, a main concern of these studies is to focus on "Asian-specific" haplogroups, when these are merely a part of the total set of haplogroups observed in East Asia. Actually, some haplogroups observed in East Asia are more widely distributed in NEA populations (Karafet et al. 2001). On the basis of genome-wide SNPs, the HUGO project also sustained the view of SEA being more diverse than NEA populations (Abdulla et al. 2009), but this study was also criticized because of a lack of NEA representative samples (Sanchez-Mazas et al. 2011a).

To understand better the peopling history of East Asia, which was probably the first main region to be colonized by anatomically modern humans after their departure from East Africa or the Middle East between 100,000 and 60,000 years ago (Jin and Su 2000), a more thorough analysis of the genetic diversity among and within the NEA and SEA populations was thus necessary. In this context, we recently investigated East Asian genetic differentiations by collecting and analyzing a large dataset of populations tested for human leukocyte antigen (HLA) complex (Di and Sanchez-Mazas 2011a), a system whose highly polymorphic genes are very informative for anthropological studies (Sanchez-Mazas et al. 2011b). This work revealed a north–south genetic differentiation and a significant, although weak genetic boundary between Northern and Southern Han Chinese populations, but not between SEA and NEA populations as a whole (Di and Sanchez-Mazas 2011a). We also observed a higher level of genetic diversity within NEA as compared to SEA populations, together with an uneven distribution of some HLA lineages and alleles between them, which we tentatively classified in two groups: group-1 lineages and alleles are commonly observed in the North but not or rarely found in the South, while group-2 lineages and alleles are commonly observed in the South but not or less frequently found in the North (see Table 2 in Di and Sanchez-Mazas 2011a and Table 2 in Di and Sanchez-Mazas 2011b). This led us to propose a new peopling scenario named "the overlapping model", suggesting that East Asia was colonized from West Asia through two migration routes along each side of the Himalayas, with subsequent gene flow between northern and southern populations (Di and Sanchez-Mazas 2011a).

However, the detailed pattern of genetic differentiation between NEA and SEA populations remained unclear, as our previous results did not show whether present East Asian populations were subdivided into two main groups or whether they were differentiated according to a geographic continuity, e.g., with clinal distributions of HLA frequencies between northern and southern regions. Moreover, the above-mentioned two groups of HLA lineages and alleles were previously defined through very rough comparisons between populations, without any statistical support. Some of them were conservatively removed because they were susceptible to be classified into either group-1 or group-2 due to a lack of objective criteria to decide. This is why we undertook new analyses, presented in this study. By using both a wider HLA dataset including many East Asian populations recently typed at the 2nd-field level of resolution and appropriate statistics to test allelic frequency clines and genetic continuity, we here present a much more robust characterization of the HLA genetic pattern in East Asia and a map of the HLA genetic diversity within populations in this continent.

Materials and methods

Samples

We worked on a large database of HLA frequency data for East Asian populations tested at one to five loci including A, B, and C of class I and DRB1 and DPB1 of class II. A part of these data were collected from the literature during our previous study (Di and Sanchez-Mazas 2011a). They include numerous Chinese populations studied for HLA between 1980 and 2012 but whose data were rarely used in international research due to their publication in Chinese. A total of 90 East Asian populations representing about 150,000 individuals were finally used (Fig. 1 and Supplementary Table 1). HLA data were defined at two resolution levels: 1st-field level data (lineages, e.g., HLA-A*01) and 2nd-field level data (alleles, e.g., HLA-A*01:01). Some populations were reported to deviate from Hardy–Weinberg equilibrium. Because the precise reasons for these deviations (e.g., heterogeneity of the samples, genotyping ambiguities, or, alternatively reasons linked to the specific demographic history of the tested populations, such as admixture) could not be assessed due to unavailable genotypic information for these samples, these data were excluded from the dataset to avoid spurious interpretation of the observed genetic patterns.

Statistical analyses

To better understand the geographic pattern of HLA lineage and allele distribution in East Asia, we studied the relationship between latitude or longitude and HLA lineage and allele frequencies. We used Spearman's nonparametric correlation coefficient (Deheuvels 1980) to account for deviations from normality (assessed by Kolmogorov–Smirnov's tests) of the frequency distributions of several lineages and alleles (in general the rarest ones). The calculations were performed with the statistical package R, version 2.10.1 (R Development Core Team 2011). Spearman's correlation coefficient *Rho* is defined as Pearson's correlation coefficient *R* between ranked variables. All HLA lineages and alleles observed in more than five populations were tested. Holm's correction (Holm 1979), an

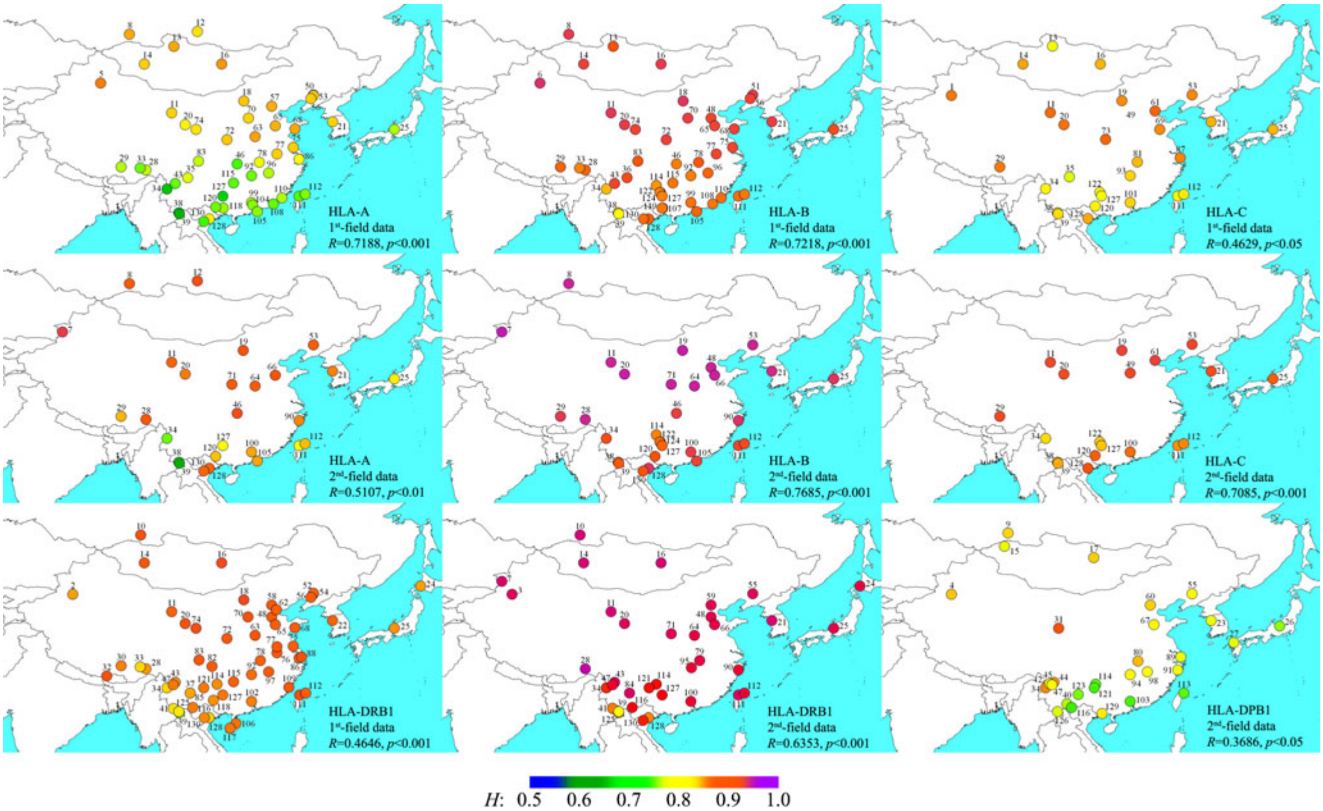

Fig. 1 Geographical distribution of the 90 sampled populations analyzed in this study. The population points were *colored* by using QGIS (see text) according to heterozygosity index (H). The correlation coefficient R between H and latitude is indicated (**: p<0.01; *: 0.01<p<0.05). Note that a single point may represent several different populations from the same location, while a same population sampled in different studies is indicated by different numbers: *1–4*, Uyghur from Xinjiang; *5–6*, Uyghur from Urumqi; *7*, Uyghur from Yining; *8–9*, Tuvinian from Tuva Republic; *10*, Tuvinian from Kyzyl; *11*, Yugur from Sunan; *12*, Buriat from Angarsk; *13*, Tsaatan from Hovsgol; *14*, Oold from Hovd; *15*, Khoton from Uvs Aïmag; *16*, Khalkha from Central and Eastern Mongolia; *17*, Khalkha from Ulaanbaatar; *18–19*, Mongolian from Inner Mongolia; *20*, Tu from Qinghai; *21–23*, Korean from South Korea; *24*, Japanese from Hokkaido; *25*, Japanese from Central Japan; *26*, Japanese from Nagano; *27*, Japanese from Fukuoka; *28*, Monba from Tibet; *29–30*, Tibetan from Tibet; *31*, Tibetan from Qinghai; *32*, Lachung from Sikkim; *33*, Luoba from Linzhi; *34*, Lisu from Nujiang; *35–36*, Yi from Liangshan; *37*, Yi from Kunming; *38*, Jinuo from Jinghong; *39*, Hani from Jinghong; *40*, Hani from Honghe; *41*, Lahu from Lancang; *42*, Nu from Nujiang; *43–44*, Naxi from Lijiang; *45*, Pumi from Yunnan; *46*, Tujia from Wufeng; *47*, Bai from Jianchuan; *48*, Han from Tianjin, Beijing, Shijiazhuang; *49*, Han from Beijing and Xi'an; *50–52*, Han from Liaoning; *53–55*, Han from Shenyang; *56*, Han from Liaonan; *57–60*, Han from Beijing; *61–62*, Han from Tianjin; *63–64*, Han from Henan; *65–67*, Han from Shandong; *68*, Han from Qingdao; *69*, Han from Linqiu; *70*, Han from Shanxi; *71*, Han from Shaanxi; *72–73*, Han from Xi'an; *74*, Han from Lanzhou; *75*, Han from Jiangsu; *76*, Han from Jianghuai; *77*, Han from Anhui; *78–80*, Han from Hubei; *81*, Han from Wuhan; *82*, Han from Chongqing; *83*, Han from Sichuan; *84*, Han from Yunnan; *85*, Han from Kunming; *86*, Han from Jiangsu, Zhejiang, Shanghai; *87–90*, Han from Shanghai; *90–91*, Han from Zhejiang; *92–94*, Han from Hunan; *95*, Han from Yueyang; *96–98*, Han from Jiangxi; *99–103*, Han from Guangdong; *104*, Han from Guangzhou; *105*, Han from Hong Kong; *106*, Han from Hainan; *107*, Han from Guangxi; *108*, Han from Chaoshan; *109*, Han from Fujian; *110*, Han from Xiamen; *111*, Han from Hsinchu; *112–113*, Han from Taipei; *114*, Miao from Guiding; *115*, Miao from Hunan; *116*, Yao from Jinping; *117*, Li from Baisha; *118–119*, Zhuang from Guangxi; *120*, Zhuang from Tiandeng; *121*, Buyi from Guizhou; *122*, Buyi from Libo; *123*, Buyi from Luoping; *124*, Shui from Libo; *125–126*, Dai from Xishuangbanna; *127*, Maonan from Huanjiang; *128*, Kinh from Hanoi; *129*, Jing from Fangcheng; *130*, Muong from Hoa Binh (see also Supplementary Table 1)

improvement of Bonferroni's correction (Cupples et al. 1984), was used to avoid false positive results due to multiple tests.

We also applied spatial autocorrelation analysis (Sokal and Wartenberg 1983) with PASSaGE software (Rosenberg 2001). This method reveals geographic patterns of genetic diversity by comparing genetic frequencies within each of several predefined geographic distance classes and by analyzing the degree of genetic similarity at various geographic distances. We used Moran's I (1950) as autocorrelation coefficients calculated by:

$$I = n\frac{\sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij}\left(y_i-\overline{y}\right)\left(y_j-\overline{y}\right)}{W\sum_{i=1}^{n}\left(y_i-\overline{y}\right)^2}$$

where y_i is the value of the variable at the ith location (for $i \neq j$), n is the number of populations, w_{ij} is a weight indicating the spatial relationship of populations i and j, and $W = \sum_{ij} w_{ij}$ represents the sum of the values in the weight matrix. Moran's I is similar to a correlation coefficient and usually ranges from 1 to −1. While positive values of I indicate positive spatial autocorrelation, negative values reveal negative spatial autocorrelation (Rosenberg 2008). Prevosti's genetic distances (Prevosti et al. 1975) between each pair of populations were computed from gene frequencies. Between pairs of populations *P* and *Q* with *k* alleles observed, the genetic distance can be calculated as:

$$D_{P,Q} = \frac{1}{2}\sum_{i=1}^{k} |p_i - q_i|$$

where p_i and q_i represent the frequency of allele *i* in populations *P* and *Q*, respectively. The distance matrices were plotted by principal coordinate analysis (PCoA) using R, version 2.10.1.

Heterozygosity index (*H*) measuring genetic diversity was estimated from gene frequencies for each population, and we used QGis software (Quantum GIS Development Team 2013) to infer gradual color changes to the genetic diversity of different populations.

Pearson's correlation test was also performed between latitude and the first coordinate of PCoA and between latitude and *H*, respectively, after having checked the normality of distributions.

Results

We created PCoA charts with the plots for several loci superimposed, and we colored the plots according to the latitude of the corresponding populations (Fig. 2). Whereas a clear genetic differentiation between NEA and SEA populations is observed, the transition between them appears to be continuous. The genetic differentiation along the latitude is mainly represented by the first principal coordinate, whose variance falls between 31.5 % (HLA-B 1st-field data) and 49.2 % (HLA-A 2nd-field data) (See Supplementary Fig. 1). Its significance is confirmed by the significant linear correlation between latitude and the first PCoA coordinate, for all loci and at both resolution levels (Fig. 2 and Supplement Fig. 1).

The heterozygosity (*H*) of each population at each locus is shown with colors on geographic maps of East Asia (Fig. 1). These maps illustrate very well the significant positive correlation between latitude and *H* assessed through linear regression for all loci at both resolution levels.

Figure 3 synthesizes *Rho* values and significance levels of Spearman's test by different symbols for each lineage and each allele considered in this study (see "Materials and methods" and Supplementary Table 3). These results reveal significant correlations between latitude and gene frequencies for several lineages and alleles (listed in Fig. 3) at each locus,

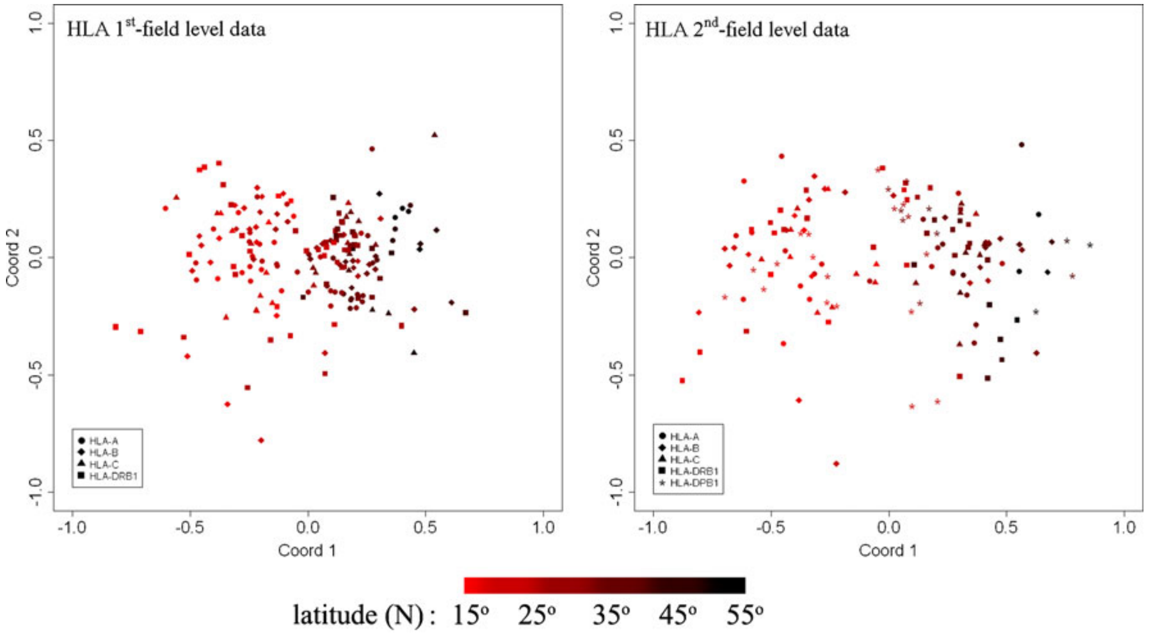

Fig. 2 Principal coordinate analysis (PCoA) for HLA-A, -B, -C, and -DRB1 1st-field level data (left) and HLA-A, -B, -C, -DRB1, and -DPB1 2nd-field level data (right) represented by the plots for different loci superimposed at the same scale. The plots were *colored* according to the latitude of the corresponding populations. Pearson's tests show significant correlation between latitude and the first coordinate: 1st-field level data: $|R_{\text{lat-coord1}}|=0.7693$; $p_{\text{lat-coord1}}<0.001$; 2nd-field level data: $|R_{\text{lat-coord1}}|=0.8266$; $p_{\text{lat-coord1}}<0.001$. Genetic variance represented by the first coordinate: 1st-field level data: HLA-A: 43.7 %, HLA-B: 31.5 %; HLA-C: 35.6 %; HLA-DRB1: 32.5 %; 2nd-field level data: HLA-A: 49.0 %; HLA-B: 39.7 %; HLA-C: 45.4 %; HLA-DRB1: 33.0 %; HLA-DPB1: 45.2 %. Genetic variance represented by the second coordinate: 1st-field level data: HLA-A: 16.4 %; HLA-B: 16.6 %; HLA-C: 24.3 %; HLA-DRB1: 18.2 %; 2nd-field level data: HLA-A: 13.0 %; HLA-B: 11.4 %; HLA-C: 13.6 %; HLA-DRB1: 14.5 %; HLA-DPB1: 17.2 %

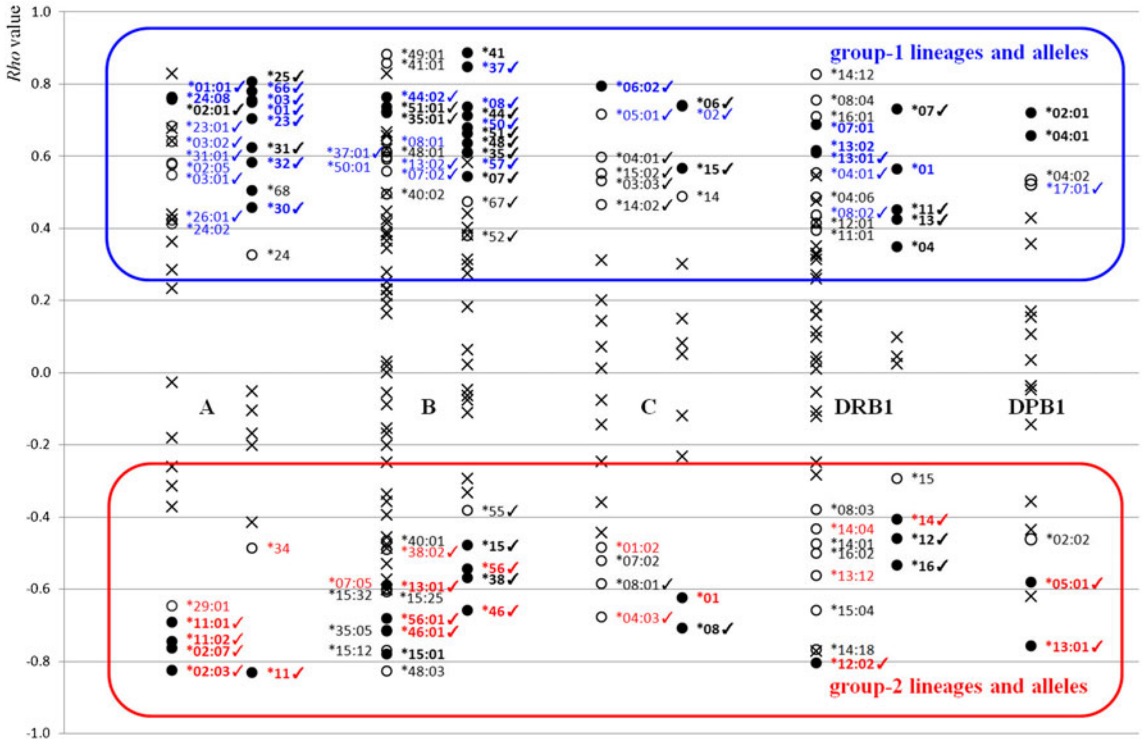

Fig. 3 Spearman's correlation coefficient *Rho* values (either positive or negative) with statistical significance of the correlation between gene frequency and latitude for each HLA lineage (1st-field level data, HLA-A, -B, -C, and -DRB1 loci) and allele (2nd-field level data, HLA-A, -B, -C, -DRB1, and -DPB1 loci) observed in more than five studied populations. Lineages and alleles with significant correlation between frequency and latitude at the 5 % level after correction by Holm's method for multiple tests (see text) are represented by *solid points*; those with significant correlation at the 5 % level but becoming nonsignificant after Holm's correction for multiple tests are represented by *open circles*; those with nonsignificant correlation at the 5 % level are represented by *crosses*. The name of group-1 and group-2 lineages and alleles are shown in *blue* and *red*, respectively; other lineages and alleles are shown in *black*. Lineages and alleles with significant clinal patterns are *checkmarked*

even after Holm's correction for multiple tests. This is not the case with longitude, however, where most significant correlations at 5 % level became nonsignificant after correction, especially for HLA lineages, the data of which are more abundant compared to HLA alleles (Supplementary Table 3).

Taking the list of "group-1" and "group-2" lineages and alleles that we defined in our previous study, we find a significant correlation for most of them, except those with sporadic distribution. According to the global distribution of HLA alleles described by Solberg et al. (2008), we renewed our list of group-1 and group-2 lineages and alleles (groups-1 and -2 lineage and allele names in blue and red, respectively, in Fig. 3 and Supplementary Table 2). While the gene frequencies of group-1 lineages and alleles correlated significantly with latitude, those of group-2 lineages and alleles correlated negatively. At the global level, most group-1 lineages and alleles are widely distributed in Africa and Europe, and in some cases in America, while most group-2 lineages and alleles are mainly observed in Asia and rarely elsewhere, except in Oceania. Interestingly, we detect more lineages and alleles with significant latitude–frequency correlation than expected as they include those lineages and alleles that we observed globally (lineage and allele names shown in black in Fig. 3).

The significant results of spatial autocorrelation analyses (Supplementary Table 3) most often reveal clinal patterns (Barbujani 2000), and most of the lineages and alleles exhibiting this kind of pattern coincide with those that were significantly correlated with latitude according to Spearman's tests (lineages and alleles with clinal patterns were checkmarked in Fig. 3).

Discussion

The new results presented in this study converge toward a robust conclusion on the HLA genetic structure of East Asia. Indeed, based on multiple statistical evidence, we show that the genetic transition from NEA to SEA populations (or vice versa) follows a latitudinal cline and a continuous pattern, thus challenging the idea of a genetic subdivision into two population groups in East Asia (Figs. 1, 2, and 3, Supplementary Fig. 1 and Supplementary Table 2). This continuous pattern is shaped by the distribution of many HLA lineages and alleles, the frequencies of which reveal significant clinal patterns with increasing or decreasing frequencies from north to south (Fig. 3). By contrast, when analyzing the data with a clustering procedure such as SAMOVA (Dupanloup et al. 2002), no significant boundary between NEA and SEA populations was observed in this continental area (results not shown). In our previous work, a significant (but weak) genetic boundary was only detected between NEA and SEA Chinese Han populations (Di and Sanchez Mazas 2011a), fitting with the linguistic subdivision of Chinese (Sinitic) languages which arouse during the last 2,000 years, a relatively recent event compared to early human migrations. More generally, the analysis of HLA genetic relationships between populations in relation to their linguistic families reveals a complex pattern in East Asia. Indeed, we previously showed that families (such as Sinitic and Tibeto-Burman) which are widely distributed from a geographic point of view are characterized by a high internal genetic diversity, while others (like Hmong–Mien, Tai–Kadai, and Austro-Asiatic) which are geographically more concentrated and close to each other are both genetically more homogeneous and more similar to each other (Di and Sanchez-Mazas 2011a). This confirms the fact that geography, compared to linguistics, is a better predictor of HLA genetic relationships in East Asia.

Genetic diversity is commonly used to trace the direction of ancient human migrations, because a higher diversity may indicate a more ancient population differentiation (Manica et al. 2007). Genetic diversity is thus expected to decrease along a migration route, as long as no other demographic factor (such as gene flow) is involved. Based on our results showing that the expected heterozygosity *H* decreases with decreasing latitudes, we thus conclude that populations migrated from north to south. However, genetic clines can also be generated by other mechanisms such as admixture between genetically distinct populations (Ammerman and Cavalli-Sforza 1984), serial founder effects (Deshpande et al. 2009), isolation by distance (Novembre and Stephens 2008; Reich et al. 2008), or natural selection (Suo et al. 2012 and see Fig. 2 in Sanchez-Mazas et al. 2011a). In the case of HLA genes, which play a crucial role in immunity, natural selection is often considered as having a significant evolutionary impact (Meyer and Thomson 2001; Satta et al. 1994). In particular, balancing selection would interfere with demographic factors in maintaining high levels of genetic diversity protecting populations against pathogens. A higher genetic diversity would then be expected in regions with higher levels of pathogens. At the global scale, a significant correlation between genetic diversity and pathogen richness appears to stand for some HLA class I (Prugnolle et al. 2005; Qutob et al. 2012; Sanchez-Mazas et al. 2012) and class II (Sanchez-Mazas et al. 2012) genes, although this relationship has been shown to be weak and not necessarily positive (Sanchez-Mazas et al. 2012). On the other hand, available information provided by the Global Infectious Diseases Database (GIDEON; http://www.gideononline.com) for each country of the world indicates higher pathogen richness in China compared to more northern areas of East Asia like Mongolia and Siberia (see Fig. 1 or 2 of Sanchez-Mazas et al. 2012). Despite a lack of information at regional scales, this pathogen distribution is not compatible with the above-mentioned expectation of a positive correlation between HLA genetic diversity and pathogen richness in East Asia because a greater genetic diversity is found in the North. Note also that for the HLA-A, -B, -C, and -DRB1 loci, which generally exhibit an excess of heterozygotes compared to neutral expectations (Buhler and Sanchez-Mazas 2011; Sanchez-Mazas 2007; Solberg et al. 2008), only seldom East Asian populations exhibit significant deviation from neutrality according to Ewens–Watterson's test (results not shown). Decreasing genetic diversity along the latitude is also observed in East Asia for the HLA-DPB1 gene whose allelic distributions markedly differ from those of the other HLA loci, suggesting another kind of evolution not governed by balancing selection (Solberg et al. 2008). Concerning individual HLA lineages and alleles, although some of them, like HLA-A*02 and HLA-B*27:05, reveal frequency clines at the global scale (De Petris et al. 2004; Mathieu et al. 2009), this is not the case for most lineages and alleles outside East Asia (Supplementary Table 2) where selective forces may also be at work (e.g., in Africa and Europe). Natural selection is thus not a satisfactory explanation of the HLA variation observed in East Asia. We thus assume that demographic factors, e.g., southward population migrations, better explain the observed genetic patterns.

Three main theories have been proposed in regards to the peopling history of East Asia. According to the first one, the genetic differences observed between NEA and SEA populations would simply be due to them having different origins. Two groups of modern humans, a few thousands of decades after their departure from East Africa or the Middle East, expanded eastward, following distinct migration routes along both sides of the Himalayas. Human populations would thus have reached East Asia both from the North and from the South (Cavalli-Sforza and Feldman 2003; Ding et al. 2000; Karafet et al. 2001; Xiao et al. 2000). Other researchers challenge this thesis by proposing a single southern origin of all East Asian populations: modern human populations would have entered East Asia through a unique southern route along the coasts, dated to about 60,000 years ago (Abdulla et al. 2009; Shi et al. 2005; Shi et al. 2008; Su et al. 1999). These authors interpret the current north–south genetic differentiation of East Asian populations either by an effect of isolation by distance during later northward migrations from Southeast Asia or by recent (<3,000 years) gene flow from Central Asian populations (Shi et al. 2005; Zhang et al. 2007). However, neither of these two scenarios can explain why a greater genetic diversity is observed in NEA populations. Moreover, the second scenario (i.e., a single southern route) does not explain why many alleles and lineages (included in the "group-1") are frequent in NEA populations and widely distributed in other continents like Africa, Europe, and in some cases America, but are not or rarely observed in SEA populations. We thus proposed a third scenario, the "overlapping model" of East Asian peopling history (Di and Sanchez-Mazas 2011a). This model first suggests that human populations followed two migration routes along both sides of the Himalayas (like in the first theory); then, populations which arrived in East Asia via the southern route migrated northward, and populations which arrived in East Asia via the northern route migrated southward (possibly at a later period), overlapping the northward migration. The most likely geographic area where populations came across and underwent gene flow would have been in the North where genetic diversity is the highest.

The genetic differences between NEA and SEA populations that we observe in the present study, and more particularly the new inventory of HLA lineages and alleles showing uneven distributions in East Asia, are in accordance with our previous assumptions (Di and Sanchez-Mazas 2011a; Sanchez-Mazas et al. 2005). Indeed, we show here that the main distribution pattern of these lineages and alleles is clinal rather than shaped by isolation by distance, which supports population admixture from two geographic origins of genetically differentiated populations (Barbujani 2000), rather than geographic distance-related variation from a single source (Supplementary Table 2). In East Asia, due to the presence of the huge mountain range of the Himalayas in the west, the genetic structure of East Asian populations shows a clear signature of a past north–south geographic subdivision, just like the Saharan desert shaped a discontinuous genetic pattern between Northern and sub-Saharan Africa (Sanchez-Mazas 2001). The continuous pattern of the north–south differentiations observed in our study indicates more intensive degree of admixture of NEA and SEA populations than previously expected. The strength of gene flow during subsequent migrations from the North and the South across open areas is then certainly the reason of the observed latitude-related pattern of genetic diversity and HLA lineage and allele distributions. Along with other arguments including the distribution of group-1 lineages and alleles in Africa, Europe, and America (Di and Sanchez-Mazas 2011a) as well as more evident north–south morphological differentiation observed among early Neolithic populations compared to modern populations (Chen and Zhang 1998; Han and Pan 1984; Wu et al. 2012), these events probably did not occur only in recent times (<3,000 years) as proposed in some studies (Shi et al. 2005; Zhang et al. 2007), but more likely during a much longer period, the exact time of which remains to be estimated.

Acknowledgments This study was supported by the Swiss National Science Foundation (FNS) projects Nos. 31003A_112651, 31003A_127465, and 31003A_144180 to Prof. Alicia Sanchez-Mazas. We also thank Mr. Nuno Miguel Da Silva, Mr. Stephan Weber, and Dr. José Manuel Nunes for their technical help and useful recommendations. We would like to thank the anonymous reviewers.

References

Abdulla MA, Ahmed I, Assawamakin A, Bhak J, Brahmachari SK, Calacal GC, Chaurasia A, Chen CH, Chen J, Chen YT, Chu J, Cutiongco-de la Paz EM, De Ungria MC, Delfin FC, Edo J, Fuchareon S, Ghang H, Gojobori T, Han J, Ho SF, Hoh BP, Huang W, Inoko H, Jha P, Jinam TA, Jin L, Jung J, Kangwanpong D, Kampuansai J, Kennedy GC, Khurana P, Kim HL, Kim K, Kim S, Kim WY, Kimm K, Kimura R, Koike T, Kulawonganunchai S, Kumar V, Lai PS, Lee JY, Lee S, Liu ET, Majumder PP, Mandapati KK, Marzuki S, Mitchell W, Mukerji M, Naritomi K, Ngamphiw C, Niikawa N, Nishida N, Oh B, Oh S, Ohashi J, Oka A, Ong R, Padilla CD, Palittapongarnpim P, Perdigon HB, Phipps ME, Png E, Sakaki Y, Salvador JM, Sandraling Y, Scaria V, Seielstad M, Sidek MR, Sinha A, Srikummool M, Sudoyo H, Sugano S, Suryadi H, Suzuki Y, Tabbada KA, Tan A, Tokunaga K, Tongsima S, Villamor LP, Wang E, Wang Y, Wang H, Wu JY, Xiao H, Xu S, Yang JO, Shugart YY, Yoo HS, Yuan W, Zhao G, Zilfalil BA (2009) Mapping human genetic diversity in Asia. Science 326(5959):1541–1545

Ammerman A, Cavalli-Sforza LL (1984) The Neolithic transition and the genetics of populations in Europe. Princeton University Press, Princeton

Barbujani G (2000) Geographic patterns: how to identify them and why. Hum Biol 72(1):133–153

Buhler S, Sanchez-Mazas A (2011) HLA DNA sequence variation among human populations: molecular signatures of demographic and selective events. PLoS One 6(2):e14643

Cavalli-Sforza LL, Feldman MW (2003) The application of molecular genetic approaches to the study of human evolution. Nat Genet 33(Suppl):266–275

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

Chen DZ, Zhang JZ (1998) The physical characteristics of the early Neolithic human in Jiahu site and the comparison with other Neolithic humans and modern man. Acta Anthropol Sin 17:191–211

Chen J, Zheng H, Bei JX, Sun L, Jia WH, Li T, Zhang F, Seielstad M, Zeng YX, Zhang X, Liu J (2009) Genetic structure of the Han Chinese population revealed by genome-wide SNP variation. Am J Hum Genet 85(6):775–785

Chu JY, Huang W, Kuang SQ, Wang JM, Xu JJ, Chu ZT, Yang ZQ, Lin KQ, Li P, Wu M, Geng ZC, Tan CC, Du RF, Jin L (1998) Genetic relationship of populations in China. Proc Natl Acad Sci U S A 95(20):11763–11768

Cupples A, Heeren T, Schatzkin A, Colton T (1984) Multiple testing of hypotheses in comparing two groups. Ann Intern Med 100:22–129

De Petris L, Bergfeldt K, Hising C, Lundqvist A, Tholander B, Pisa P, van der Zanden HG, Masucci G (2004) Correlation between HLA-A2 gene frequency, latitude, ovarian and prostate cancer mortality rates. Med Oncol 21(1):49–52

Deheuvels P (1980) Non parametric tests of independence. In: Raoult J-P (ed) Statistique non Paramétrique Asymptotique, Springer, pp 95–107

Deshpande O, Batzoglou S, Feldman MW, Cavalli-Sforza LL (2009) A serial founder effect model for human settlement out of Africa. Proc Biol Sci 276(1655):291–300

Di D, Sanchez-Mazas A (2011a) Challenging views on the peopling history of East Asia: the story according to HLA markers. Am J Phys Anthropol 145(1):81–96

Di D, Sanchez-Mazas A (2011b) The peopling history of continental East Asia revealed by the HLA system. Commun Contemp Anthropol 5(1):273–280

Ding YC, Wooding S, Harpending HC, Chi HC, Li HP, Fu YX, Pang JF, Yao YG, Yu JG, Moyzis R, Zhang Y (2000) Population structure and history in East Asia. Proc Natl Acad Sci U S A 97(25):14003–14006

Du RF, Xiao CJ, Cavalli-Sforza LL (1997) Genetic distances between Chinese populations calculated on gene frequencies of 38 loci. Sci China C Life Sci 40(6):613–621

Dupanloup I, Schneider S, Excoffier L (2002) A simulated annealing approach to define the genetic structure of populations. Mol Ecol 11(12):2571–2581

Han KX, Pan QF (1984) Ethnological structure of ancient populations of China. Acta Archaeol Sin 2:245–263

Holm S (1979) A simple sequentially rejective multiple test procedure. Scand J Stat 6(2):65–70

Jin L, Su B (2000) Natives or immigrants: modern human origin in East Asia. Nat Rev Genet 1(2):126–133

Karafet T, Xu L, Du R, Wang W, Feng S, Wells RS, Redd AJ, Zegura SL, Hammer MF (2001) Paternal population history of East Asia: sources, patterns, and microevolutionary processes. Am J Hum Genet 69(3):615–628

Manica A, Amos W, Balloux F, Hanihara T (2007) The effect of ancient population bottlenecks on human phenotypic variation. Nature 448(7151):346–348

Mathieu A, Paladini F, Vacca A, Cauli A, Fiorillo MT, Sorrentino R (2009) The interplay between the geographic distribution of HLA-B27 alleles and their role in infectious and autoimmune diseases: a unifying hypothesis. Autoimmun Rev 8(5):420–425

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

Novembre J, Stephens M (2008) Interpreting principal component analyses of spatial population genetic variation. Nat Genet 40(5):646–649

Poloni ES, Sanchez-Mazas A, Jacques G, Sagart L (2005) Comparing linguistic and genetic relationships among East Asian populations: a study of the RH and GM polymorphisms. In: Sagart L, Blench R, Sanchez-Mazas A (eds) The peopling of East Asia: putting together archaeology. Linguistics and Genetics, Routledge Curzon (Taylor and Francis Group), London and New York

Prevosti A, Ocaña J, Alonso G (1975) Distances between populations of *Drosophila subobscura* based on chromosome arrangement frequencies. Theoret Appl Genet 45(6):231–241

Prugnolle F, Manica A, Charpentier M, Guegan JF, Guernier V, Balloux F (2005) Pathogen-driven selection and worldwide HLA class I diversity. Curr Biol 15(11):1022–1027

Quantum GIS Development Team (2013) Quantum GIS Geographic Information System. Open Source Geospatial Foundation Project. http://qgis.osgeo.org. Accessed 29 July 2013.

Qutob N, Balloux F, Raj T, Liu H, Marion de Proce S, Trowsdale J, Manica A (2012) Signatures of historical demography and pathogen richness on MHC class I genes. Immunogenetics 64(3):165–175

R Development Core Team (2011) R: a language and environment for statistical computing. R Foundation for Statistical Computing, Vienna

Reich D, Price AL, Patterson N (2008) Principal component analysis of genetic data. Nat Genet 40(5):491–492

Rosenberg MS (2001) PASSaGE: pattern analysis, spatial statistics, and geographic exegesis 1.0. Arizona State University, Tempe, AZ

Rosenberg MS (2008) PASSaGE manual. Arizona State University, Tempe, AZ

Sanchez-Mazas A (2001) African diversity from the HLA point of view: influence of genetic drift, geography, linguistics, and natural selection. Hum Immunol 62(9):937–948

Sanchez-Mazas A (2007) An apportionment of human HLA diversity. Tissue Antigens 69(Suppl 1):198–202

Sanchez-Mazas A, Poloni ES, Jacques G, Sagart L (2005) HLA genetic diversity and linguistic variation in East Asia. In: Sagart L, Blench R, Sanchez-Mazas A (eds) The peopling of East Asia: putting together archaeology, linguistics and genetics. Routledge Curzon (Taylor and Francis Group), London and New York

Sanchez-Mazas A, Di D, Riccio ME (2011a) A genetic focus on the peopling history of East Asia: critical views. Rice 4(3–4):159–169

Sanchez-Mazas A, Fernandez-Vina M, Middleton D, Hollenbach JA, Buhler S, Di D, Rajalingam R, Dugoujon JM, Mack SJ, Thorsby E (2011b) Immunogenetics as a tool in anthropological studies. Immunology 133(2):143–164

Sanchez-Mazas A, Lemaitre JF, Currat M (2012) Distinct evolutionary strategies of human leucocyte antigen loci in pathogen-rich environments. Philos Trans R Soc Lond B Biol Sci 367(1590):830–839

Satta Y, O'HUigin C, Takahata N, Klein J (1994) Intensity of natural selection at the major histocompatibility complex loci. Proc Natl Acad Sci U S A 91(15):7184–7188

Shi H, Dong YL, Wen B, Xiao CJ, Underhill PA, Shen PD, Chakraborty R, Jin L, Su B (2005) Y-chromosome evidence of southern origin of the East Asian-specific haplogroup O3-M122. Am J Hum Genet 77(3):408–419

Shi H, Zhong H, Peng Y, Dong YL, Qi XB, Zhang F, Liu LF, Tan SJ, Ma RZ, Xiao CJ, Wells RS, Jin L, Su B (2008) Y chromosome evidence of earliest modern human settlement in East Asia and multiple origins of Tibetan and Japanese populations. BMC Biol 6:45

Sokal RR, Wartenberg DE (1983) A test of spatial autocorrelation analysis using an isolation-by-distance model. Genetics 105(1):219–237

Solberg OD, Mack SJ, Lancaster AK, Single RM, Tsai Y, Sanchez-Mazas A, Thomson G (2008) Balancing selection and heterogeneity across the classical human leukocyte antigen loci: a meta-analytic review of 497 population studies. Hum Immunol 69(7):443–464

Su B, Xiao J, Underhill P, Deka R, Zhang W, Akey J, Huang W, Shen D, Lu D, Luo J, Chu J, Tan J, Shen P, Davis R, Cavalli-Sforza L, Chakraborty R, Xiong M, Du R, Oefner P, Chen Z, Jin L (1999) Y-Chromosome evidence for a northward migration of modern humans into Eastern Asia during the last Ice Age. Am J Hum Genet 65(6):1718–1724

Suo C, Xu H, Khor CC, Ong RT, Sim X, Chen J, Tay WT, Sim KS, Zeng YX, Zhang X, Liu J, Tai ES, Wong TY, Chia KS, Teo YY (2012) Natural positive selection and north–south genetic diversity in East Asia. Eur J Hum Genet 20(1):102–110

Turner CG 2nd (1987) Late Pleistocene and Holocene population history of East Asia based on dental variation. Am J Phys Anthropol 73(3): 305–321

Wu X, Liu W, Bae C (2012) Craniofacial variation between southern and northern Neolithic and modern Chinese. Int J Osteoarchaeol 22:98–109

Xiao CJ, Cavalli-Sforza LL, Minch E, Du RF (2000) Geographic distribution maps of human genes in China. Yi Chuan Xue Bao 27(1):1–6

Xue F, Wang J, Hu P, Ma D, Liu J, Li G, Zhang L, Wu M, Sun G, Hou H (2005) Identification of spatial genetic boundaries using a multifractal model in human population genetics. Hum Biol 77(5): 577–617

Xue F, Wang Y, Xu S, Zhang F, Wen B, Wu X, Lu M, Deka R, Qian J, Jin L (2008) A spatial analysis of genetic structure of human populations in China reveals distinct difference between maternal and paternal lineages. Eur J Hum Genet 16(6):705–717

Yao YG, Kong QP, Bandelt HJ, Kivisild T, Zhang YP (2002) Phylogeographic differentiation of mitochondrial DNA in Han Chinese. Am J Hum Genet 70(3):635–651

Zhang ZB (1988) An analysis of the physical characteristics of modern Chinese. Acta Anthropol Sin 7(4):314–323

Zhang F, Su B, Zhang YP, Jin L (2007) Genetic studies of human diversity in East Asia. Philos Trans R Soc Lond B Biol Sci 362(1482):987–995

Zhao TM, Lee TD (1989) Gm and Km allotypes in 74 Chinese populations: a hypothesis of the origin of the Chinese nation. Hum Genet 83(2):101–110