

RESEARCH ARTICLE

Open Access



Computer simulation of human leukocyte antigen genes supports two main routes of colonization by human populations in East Asia

Da Di¹, Alicia Sanchez-Mazas^{1,2*} and Mathias Currat^{1*}

Abstract

Background: Recent genetic studies have suggested that the colonization of East Asia by modern humans was more complex than a single origin from the South, and that a genetic contribution via a Northern route was probably quite substantial.

Results: Here we use a spatially-explicit computer simulation approach to investigate the human migration hypotheses of this region based on one-route or two-route models. We test the likelihood of each scenario by using Human Leukocyte Antigen (HLA) – A, –B, and –DRB1 genetic data of East Asian populations, with both selective and demographic parameters considered. The posterior distribution of each parameter is estimated by an Approximate Bayesian Computation (ABC) approach.

Conclusions: Our results strongly support a model with two main routes of colonization of East Asia on both sides of the Himalayas, with distinct demographic histories in Northern and Southern populations, characterized by more isolation in the South. In East Asia, gene flow between populations originating from the two routes probably existed until a remote prehistoric period, explaining the continuous pattern of genetic variation currently observed along the latitude. A significant although dissimilar level of balancing selection acting on the three HLA loci is detected, but its effect on the local genetic patterns appears to be minor compared to those of past demographic events.

Keywords: East Asia, Human peopling history, HLA, Computer simulation, Approximate Bayesian computation, Balancing selection

Background

During their extensive colonization of the world since the emergence of *Homo sapiens* in East Africa around 200,000 years ago [1], modern human populations have evolved genetically through various mechanisms including random genetic drift, gene flow and natural selection [2]. While the effects of genetic drift and gene flow on the genetic variation within and among different populations are strongly related to population demography, those of natural selection depend on environmental pressures and may blur the signals of population demography. To reconstruct the migration history of anatomically modern humans throughout the world, an essential

issue is thus to estimate the respective contribution of these different evolutionary forces, in particular when the studied genes are known to evolve under substantial natural selection.

A topic which has been heavily debated over the last few years is the human peopling history of East Asia, mostly because Northern East Asian populations (NEAs), on one side, and Southern East Asian populations (SEAs), on the other side, have been found to differ greatly from each other from a genetic point of view [3]. Indeed, while population geneticists generally accept the idea of a recent African origin of East Asian populations (less than 100,000 years ago) despite possible, but minor contributions from other human species once inhabiting Eurasia, Neanderthals and/or Denisovans [4] for instance, two main alternative scenarios have been proposed to explain the first arrival of modern humans in this region. According to the Southern-origin model, modern humans

* Correspondence: alicia.sanchez-mazas@unige.ch; mathias.currat@unige.ch

¹Department of Genetics and Evolution - Anthropology Unit, Laboratory of Anthropology, Genetics and Peopling history (AGP lab), University of Geneva, 12 rue Gustave-Revilliod, Geneva CH-1211, Geneva 4, Switzerland
Full list of author information is available at the end of the article

migrated eastward along the southern edge of the Himalayan Mountains to reach the South of East Asia (SEA), and further differentiated by expanding northward to the North of East Asia (NEA) [5–7]. Some authors defending the Southern-origin hypothesis also admit a certain amount of gene flow from Central Asia at a recent period beginning 3,000 or 2,000 years ago [3]. An alternative scenario is known as the Pincer model [8]. It suggests that, besides the aforementioned Southern route, modern humans also followed a Northern route across Central Asia (along the northern edge of the Himalayas) and Southern Siberia to reach NEA [9].

Interestingly, the sharp genetic variation between NEAs and SEAs has not solely been observed for non-functional, supposedly neutral, genetic markers such as Short Tandem Repeats (STRs) and Single Nucleotide Polymorphisms (SNPs). Indeed, similar patterns have been found for the immune-system genes defining the human leukocyte antigen (HLA) polymorphisms [10–12], the evolution of which is partly driven by heterozygous advantage in pathogen-rich environments [13–15]. To account for the specific observations made for this system in East Asia (i.e. a continuous genetic differentiation along the latitude shaped by two groups of lineages and alleles showing clinal distributions, and a significant decrease of genetic diversity from north to south), a third scenario, named the Overlapping model, has recently been proposed (Fig. 1). This scenario is basically similar to the

Pincer model suggesting two routes of colonization into East Asia, the difference being that it emphasizes the role of large-scale and probably long-lasting genetic exchanges between NEAs and SEAs in a northern geographic region, which eventually created the observed north-south genetic continuity [10–12].

HLA genes provide an interesting complement to neutral variability for the study of human peopling history because they may reflect additional selective processes beyond the effect of past demography and migration. Their results may thus be contrasted to those obtained with neutral loci. HLA genes constitute very useful data to tackle the question of the settlement history of East Asia because allelic states (in the form of HLA lineages defined at the 1st-field level of resolution according to the official nomenclature found at <http://hla.alleles.org/nomenclature/naming.html>) are available at several loci for the same sets of individuals in a large database of East Asian populations [10–12], thus avoiding the difficulty to consider sampling variation or heterogeneity among loci [16]. To our knowledge, there are currently no other genetic datasets available for East Asian populations which are equivalent in terms of both sample numbers and sample sizes. A potential difficulty, however, is that HLA genes partly evolve under balancing selection, which may create confounding effects. Indeed, different evolutionary forces, which are hard to disentangle by classical population genetics approaches, sometimes

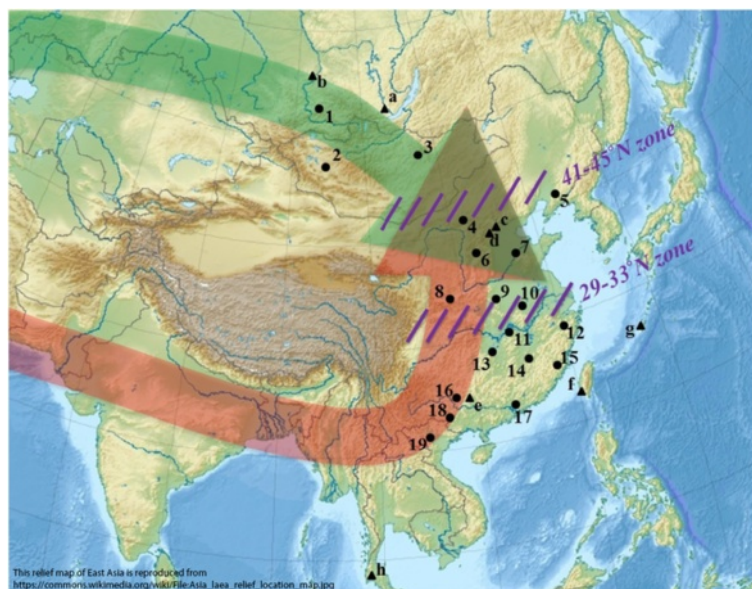


Fig. 1 Topographic map of East Asia showing hypothesized modern human migration routes and suggested barriers. Northern and Southern routes are represented by green and red arrows, respectively, while the two contact zones are marked in purple. Triangles indicate representative Upper Paleolithic archeological sites with human remains: a. Mal'ta; b. Afontova Gora-2; c. Upper Cave; d. Tianyuan Cave; e. Liujiang; f. Chochen; g. Minatogawa; h. Niah Cave and points indicate modern populations samples for HLA-A, -B, -DRB1 loci: 1. Tuvinians; 2. Oold; 3. Khalkha; 4. Mongolians; 5. Liaoning Han; 6. Shanxi Han; 7. Shandong Han; 8. Xi'an Han; 9. Henan Han; 10. Anhui Han; 11. Hubei Han; 12. Zhejiang Han; 13. Hunan Han; 14. Jiangxi Han; 15. Fujian Han; 16. Maonan; 17. Guangdong Han; 18. Zhuang; 19. Muong (see Additional file 1: Table S1 and Additional file 2: Table S2)

generate similar signals on patterns of genetic variation. For instance, genetic diversity within populations may be maintained through intensive gene flow but also through balancing selection; on the contrary, a loss of diversity may be the result of either rapid genetic drift or purifying selection [17]. Moreover, a continuous pattern of genetic variation among populations (often described by genetic clines) can be explained by demographic processes (e.g. demic diffusion with admixture between genetically distinct populations [18], isolation-by-distance with gene flow between neighboring populations [19, 20], population expansions [21]) or varying selective pressures in different environments [22]. However, such confounding effects are not expected to affect all HLA loci in the same way. Indeed, previous studies have shown that, whereas HLA-B and -DRB1 exhibit significant excess of heterozygotes in most populations tested so far, suggesting balancing selection, the genetic diversity observed at HLA-A is closer to neutral expectations [23, 24], and is thus more susceptible to reveal demographic signals. Therefore, we decided to tackle the problem mentioned above both by undertaking an original computer-simulation analysis allowing to disentangle the effects of distinct evolutionary forces and by applying this approach to three different HLA loci known to be submitted to distinct intensities of selection, namely HLA-A, -B and -DRB1, which were also tested on identical population samples available in our large database for East Asian populations. Our original computer-simulation algorithm allows a direct comparison between the three HLA genes investigated here in controlling for demography, hence a main interest of our study. When combined with Approximate Bayesian Computation (ABC) [25], it allows to estimate separately the influence of demography and the strength of selection on each locus, even in the absence of additional “control” neutral datasets. This approach has already proved its effectiveness in the field of population genetics, e.g. [16, 26, 27]. In this study, we have specifically adapted it to reproduce East Asian HLA-A, -B and -DRB1 genetic variation under the three hypothetical models described above taking into account various evolutionary parameters including population density, demographic growth, migration, and balancing selection.

Methods

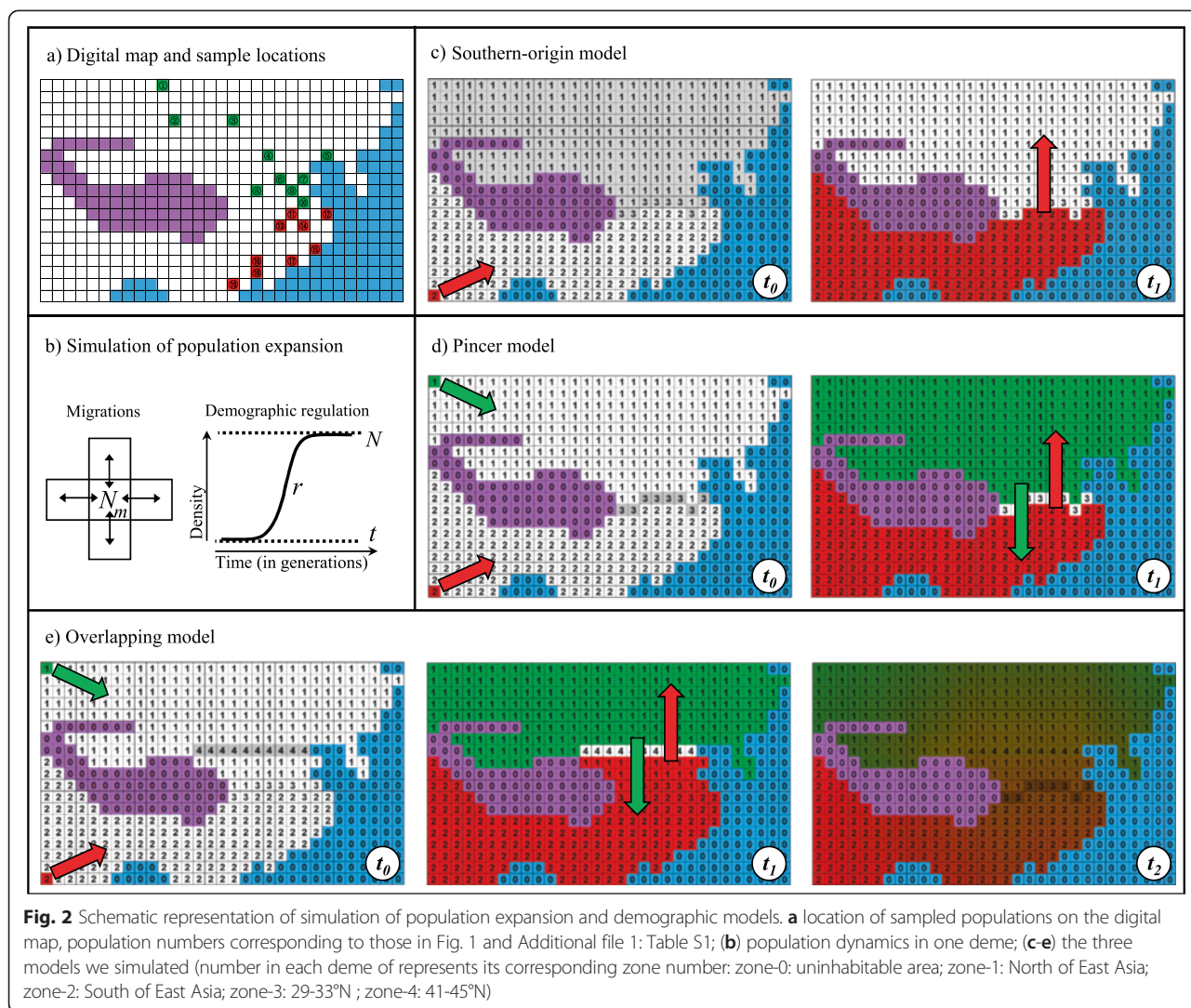
Simulations

A simulation program called SELECTOR was developed by our laboratory in recent years, written in C++ and compiled on Linux environment. The main algorithms of the program have been described and applied in a previous study focusing on human genetic differentiations across the Strait of Gibraltar [16]. The program simulates populations of diploid individuals from generation to generation, forward in time, within a “stepping-stone” framework [28], using a

set of predefined parameters with prior distributions. The geographical region of interest is represented by a digital map subdivided into demes with equal size of 40,000 (200×200) km² (Fig. 2a). Generations are discrete, which means that individuals from one generation, in each deme, are completely replaced by their descendants in the next generation. The population density (N) increases logistically (with growth rate r) until reaching the maximum carrying capacity of the deme (Fig. 2b). A proportion of individuals belonging to each deme (the migration rate m) migrates to neighboring demes (spread equally among neighbors, Fig. 2b), allowing a population expansion into all demes from one or two source demes. The genotype of each diploid individual (homozygote or heterozygote) is composed by lineages and is created randomly from a common ancestral pool of lineages (Number of initial lineages A) at the first generation, then from the parental pool of the same deme for all subsequent generations.

Geographic barriers to ancient human migrations were defined as uninhabitable demes, either permanently, like seas and Himalayan Mountains, or temporarily, like great rivers, climate limits, and/or cultural frontiers (Fig. 2). SELECTOR was adapted to the current study by defining two different zones, NEA and SEA (see next section), accounting for the well documented north-south genetic variation in East Asia. NEA and SEA are separated by the 29-33°N zone, which represents the area of Qinling-Huaihe line or Yangtze River.

In each deme, SELECTOR simulates the evolution of lineage frequencies of a gene by considering as simulation parameters various demographic factors (N , m , r). In addition, SELECTOR may introduce the effect of balancing selection on a gene by giving advantage (depending on a selection coefficient s) to heterozygotes compared to homozygous individuals when creating a new generation, which corresponds to the Symmetric Overdominant Selection model (SOS [16]). For evaluating the robustness of our results regarding the choice of the selection model, we tested two alternative balancing selection models with a relatively smaller number of simulations: heterozygote advantage varying along latitude (LOS for Latitudinal Overdominant Selection) and Negative Frequency-Dependent Selection (NFDS). The total number of generations corresponds to about 60,000-50,000 years of the overall modern human peopling history in East Asia, according to archaeological and genetic studies [5]. The modified version of SELECTOR used here allows division of the total simulation duration into several periods with different demographic characteristics. It allows to take into consideration critical scenarios such as the emergence or disappearance of geographic barriers (by changing m), or the Paleolithic-Neolithic transition (by changing N and r). At the end of a simulation, SELECTOR samples individuals in each cell representing a geographic location where real genetic



samples are available, with sample sizes being identical to real data (Fig. 2a). This approach allows a direct comparison between simulated and observed data.

Different mechanisms such as point mutations, recombinations and gene conversions frequently introduce new HLA variants in human populations [29, 30], the generation of new alleles (i.e. defined at 2nd-field level of resolution or above, according to the official nomenclature) being thus difficult to model. In contrast, HLA lineages are more conservative and considered to be ancient [31, 32]. As we do not use HLA allelic but lineage data in this study (both to take advantage of the large amount of lineage data available at several HLA loci for the same population datasets and to avoid the challenging issue of modelling complex mechanisms for the generation of new HLA alleles), SELECTOR does not simulate the appearance of new alleles, but simulates changes in lineage frequencies due to genetic drift and possible balancing selection.

Demographic models

Because complex models often lead to biased estimations when using ABC, it is recommended to simulate a very large set of data under a simplified model with a restricted number of parameters [33]. For each of the three scenarios that we have simulated, we tested several alternative versions in a qualitative way in order to find the most representative but simplified models presented below, and we performed 100,000 simulations according to each final scenario for further quantitative estimations. NEA and SEA were always defined as zone-1 and zone-2.

- 1) Southern-origin model (one-route model, Fig. 2c). One expansion initiates from the southwestern-most deme at generation $t_0 = 0$ (between 60,000 and 50,000 years ago), the duration of a simulation being 2,500 generations. A complete barrier to migration corresponding to the 29-33°N zone (zone-3) exists, which prevents the passage of migrants. At generation t_1

- ($0 < t_1 < 2,000$), the barrier either disappears or becomes a partial barrier to gene flow according to the specific parameters of simulation. Maximum t_1 was set to 2,000 generations to allow the NEA to be entirely colonized at the end of the 2,500 generations.
- 2) Pincer model (two-route model, Fig. 2d). Two expansions initiate from the southwestern-most and northwestern-most demes, respectively, at generation $t_0 = 0$ (between 60,000 and 50,000 years ago). Individuals in these two demes are derived from a common ancestral source population. Like for the Southern-origin model, a complete barrier to migration corresponding to the 29-33°N zone (zone-3) exists until its disappearance, at least partially, at generation t_1 ($0 < t_1 < 2,500$).
 - 3) Overlapping model (two-route model, Fig. 2e). Two expansions initiate from the southeastern-most and the northeastern-most demes, respectively, at generation $t_0 = 0$ (between 60,000 and 50,000 years ago). A complete barrier corresponding to the 41-45° N zone (zone-4) exists until its disappearance, at least partially, at generation t_1 ($0 < t_1 < 2,500$). This barrier to gene flow is located further north than the 29-33°N zone (zone-3). The latter also constitutes a partial barrier emerging later and reducing gene flow considerably with migration rate of 0.1 m at generation t_2 ($2,500 < t_2 < 3,000$).

In summary, the three models have 12 parameters in common: number of initial lineages in the source population A , selection coefficient s , north-south separation time t_1 , as well as independent N , m and r for the NEA, the SEA and the barrier zone. For the two-route models, we were particularly interested by the admixture time T ($T = 2,500 - t_1$ for the Pincer models and $T = t_2 - t_1$ for the Overlapping model) between NEAs and SEAs. Note that the Southern-origin and Pincer models run for 2,500 generation in total while the Overlapping model runs for 3,000 generations. The product of N and m (Nm) was estimated as a whole because it represents the absolute number of migrants leaving a deme. The prior

distributions of parameters were defined based on literature (Table 1).

Observed data

The major objective of the simulation approach is to reproduce the general genetic pattern of East Asian populations and to estimate parameters at a large geographic scale. In order to avoid potential bias due to local demographic events specific to certain populations, such as bottlenecks, directional selection or recent gene flow, we excluded some populations known for their small size and/or isolation, such as the Nu, Wa, Jinuo [34], or for their recent interbreeding history, such as the Uyghurs, Kinh and Thai [35]. We also excluded populations from peninsula and island areas, such as Koreans, Taiwanese aborigines, Japanese and Ryukyans because these populations may have been subject to strong founder effect or admixed origins which are more difficult to model in our study [35–38].

In total, 21 lineages were considered for HLA-A (A*01, *02, *03, *23, *24, *25, *26, *34, *66, *11, *29, *30, *31, *32, *33, *74, *68, *69, *36, *43, *80), 34 for HLA-B (B*07, *08, *13, *14, *15, *18, *27, *35, *37, *38, *39, *40, *41, *42, *44, *45, *46, *47, *48, *49, *50, *51, *52, *53, *54, *55, *56, *57, *58, *59, *67, *73, *78, *81) and 14 for HLA-DRB1 (DRB1*01, *02, *03, *04, *07, *08, *09, *10, *11, *12, *13, *14, *15, *16). Due to the relatively abundant population data tested for these HLA lineages, we managed to keep 19 populations (average sample size 2,662, ranging from 52 to 9,678 individuals), each of them being typed on the same samples for the three loci in the same study (Figs. 1, 2a and Additional file 1: Table S1). This enabled us to make a direct comparison among the loci by excluding possible sampling bias for the same population. All HLA frequency data used in this study were taken from publicly available publications (see references in Additional file 1: Table S1).

Model comparison

A total of 12 summary statistics were carefully chosen to capture a maximum of information from the simulated

Table 1 Prior distributions and description of the parameters used in the simulations

Parameters	Meaning	Prior distribution	References and/or explanation
T	Admixture time between NEA and SEA under two-route models (in generations)	1 to 2,500	Between 0 and 65,000 years (estimation of the spread of modern humans in East Asia [5]), each generation representing 25 years
A	Number of initial lineages in the source population	10 to 50	Numbers of HLA-A, -B and -DRB1 lineages vary between 10 and 40 in East Asian samples (from data listed in Additional file 1: Table S1)
N	Maximum population density of one deme	0 to 5,000	According to the estimation of hunter-gatherer densities at the end of Paleolithic (0-0.4 individuals per km^2 [18])
m	Population migration rate of one deme	0.01 to 0.20	Between 1 % and 20 % of emigration per generation [18]
r	Population growth rate of one deme	0.01 to 0.20	According to the estimation of growth rate for Paleolithic populations [18]
s	Coefficient of balancing selection	0 to 0.025	Selection rate higher than 2.5 % was tested by preliminary simulations but it never lead to the reproduction of observed data [40–42]

and observed data, at various levels of genetic diversity (locus variability, intra-population diversity overall but also separately in NEAs and SEAs, gradient of diversity along latitude, inter-population differentiation overall, between and within NEAs and SEAs groups, Table 2). The statistics were estimated by arlsumstat program [39]. The ABCtoolbox package was used to perform model comparisons and later parameter estimations [33]. The likelihood of a model was evaluated by computing Euclidean distances between the simulated summary statistics and those of the observed data [33, 40, 41]. To evaluate whether the observed data were in agreement with the simulated data, ABCtoolbox reports a *p-value* varying between 0 (no fit) and 1 (good fit) [33].

The relative probabilities of the three scenarios were computed using the model choice acceptance method [42] consisting in the proportion δ of data sets, among the N simulated, that are the closest to the observed data set. Here, N is equal to 300,000 simulations (100,000 per scenario), and we used three different values of δ to ensure the robustness of the comparison: $\delta = 0.25\%$ (750 simulations), 0.5% (1,500 simulations) and 1% (3,000 simulations). In addition, ABCtoolbox allows performing model selection through the computation of the Bayes factors. The Bayes factor B_{AB} in favor of model M_A over model M_B is $B_{AB} = fM_A(s_{obs})/fM_B(s_{obs})$ where fM_A and fM_B are the marginal densities of model A and B, respectively.

Estimation of parameters and validation

In order to optimize the amount of information on the model parameters extracted from the data, we further calculated Partial Least Squares (PLS) on the original statistics [43], a method available in ABCtoolbox [33] and similar to the Principal Component Analysis in extracting the main independent axes of variance. A

Root Mean Squared Error Prediction (RMSEP) chart is reported for each parameter, reflecting how its value influences the simulated results and allowing to decide the best number of PLS to use for parameter estimation.

We estimated the parameters under the most likely model using the ABC-GLM method implemented in ABCtoolbox [44]. This method initiates a process of rejection by retaining a small proportion of simulations based on the minimized Euclidean distance calculated between the simulated and observed statistics (with or without PLS). The posterior distribution for each parameter is computed from these retained simulations [33]. We set the proportion of retained simulations as 0.5% but also tested other fractions (0.25% , 1%) to ensure the robustness of the estimation.

The results obtained were validated by studying the distribution of “posterior quantiles” for each parameter, in order to test whether the posterior distribution of the parameters is biased compared to the prior distribution. A total of 100 sets of statistics were generated under the best model with parameters drawn from the posterior distributions and considered as if they had been observed in reality (“pseudo-observations”). It has been proved mathematically [45] that the positions of these true parameters are distributed uniformly in the marginal cumulative posterior distribution. Deviation from the uniform distribution was detected by the Kolmogorov-Smirnov test implemented in ABCtoolbox [33]. Moreover, the simulated and retained values for each pair of statistics were plotted to verify whether the observed values fell within the range of the retained ones (Additional file 3: Table S3).

Computation platform

Our computation tasks were sent to a platform called EZ-Grid composed of hundreds of nodes in several locations in Switzerland and France [46].

Table 2 Statistics computed from observed data for the three HLA loci

Statistic	Meaning	A	B	DRB1
M_A	Mean of lineage number for all populations	14.47	27.26	12.84
SD_A	Standard deviation of lineage number for all populations	3.96	6.64	1.01
M_H	Average of heterozygosity index H for all populations	0.79	0.91	0.89
SD_H	Standard deviation of H for all populations	0.054	0.021	0.019
$M_{H-North}$	Mean of H for NEAs	0.84	0.92	0.90
$SD_{H-North}$	Standard deviation of H for NEAs	0.012	0.0056	0.0070
$M_{H-South}$	Mean of H for SEAs	0.75	0.89	0.88
$SD_{H-South}$	Standard deviation of H for SEAs	0.044	0.010	0.021
R_{H-lat}	Correlation coefficient between H and latitude	0.78	0.84	0.65
F_{ST}	F statistic measuring within- and inter-group differentiation	0.017	0.011	0.0075
F_{CT}	F statistic measuring inter-group differentiation	0.015	0.0092	0.0055
F_{SC}	F statistic measuring within-group differentiation	0.0023	0.0016	0.0020

Results

Likelihood of the different models tested

At among-population level, the north-south differentiation is more frequently observed under the two-route models, with higher values of correlation coefficient between the first coordinate of Principal Coordinate Analysis and latitude of populations ($|R_{Coord1-lat}|$, Fig. 3a). At within-population level, the simulated genetic diversity (H) decreases systematically from south to north under the Southern-origin model, leading to negative values of R_{H-lat} in the majority (>95 %) of simulations (Fig. 3b). Under the two-route models, R_{H-lat} values thus vary between -1 and +1 (Fig. 3b), depending on several parameters including the product of population density by migration rate Nm and the selection coefficient s .

When the results of the 300,000 simulations are mixed, the proportion found for each model among the 1,500 best simulations retained ($\delta = 0.5$ %) indicates that the

Pincer and the Overlapping models are much more likely (>96 % combined, Table 3) than the Southern-origin model, the latter being almost not supported (<4 %). Similar results are obtained with either 750 ($\delta = 0.25$ %) or 3,000 ($\delta = 1$ %) best simulations retained.

In addition, the Bayes factors are extremely in favor of the Pincer and Overlapping (both two-route) models because of the low marginal density computed for the Southern-origin model (Table 4). The Bayes factor computed between models also “decisively” [47] supports better the Overlapping than the Pincer model (between 5,000 and 40,000 times depending on the locus, Table 4).

The simulations of the Southern-origin model are never able to reproduce the observed HLA-A, -B and -DRB1 data, as indicated by the p -values, equal to 0 for the three loci. For both two-route models, the p -values are higher, reaching 1-8 % for the Pincer model and of 7-24 % for the Overlapping model (Table 4). Plots of simulated, retained

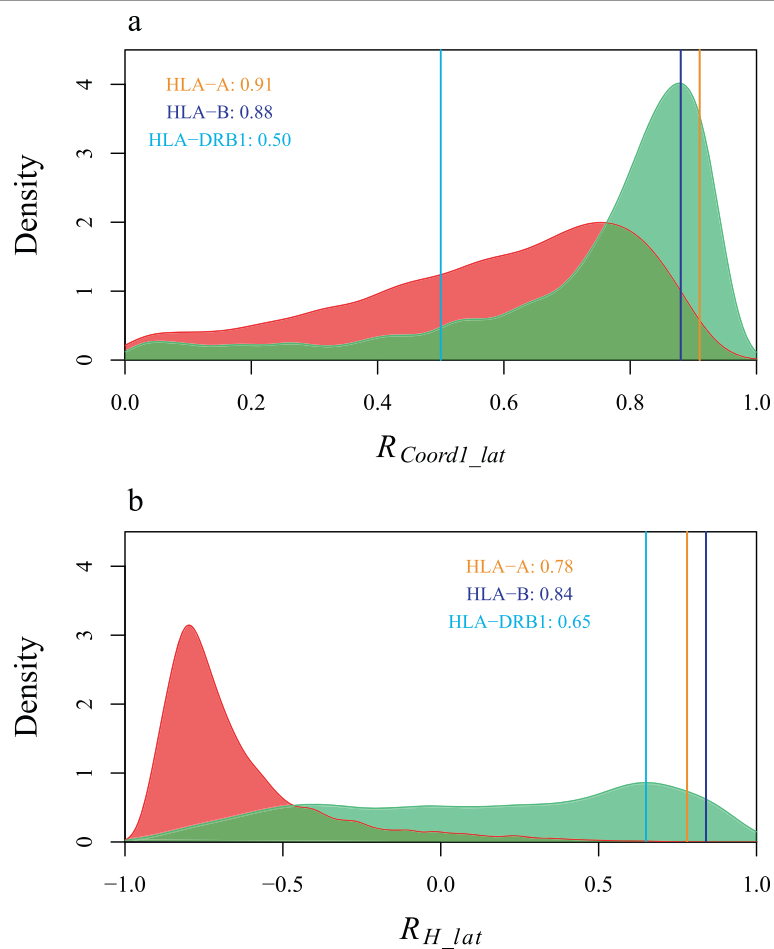


Fig. 3 Density distribution of simulated correlation coefficient between latitude and among- or within-population variation indices. Density distribution of: (a) absolute value of correlation coefficient between the first Principal Coordinate and latitude of populations and (b) value of correlation coefficient between heterozygosity index H and latitude of populations for one-route model and two-route model, respectively (red curve: one-route model; green curve: two-route model; observed values for each HLA locus are listed and displayed by vertical lines: orange line: HLA-A, blue line: HLA-B, light blue line: HLA-DRB1)

Table 3 Model comparison using retained simulations. Proportions of simulations (%) under each of the three models among 750, 1,500 and 3,000 best simulations retained from 300,000 simulations (100,000 for each model)

Number of retained simulations	Locus	Southern-origin model	Pincer model	Overlapping model
750	A	2.4	31.2	66.4
	B	0.5	26.3	73.2
	DRB1	0.2	37.5	62.3
1,500	A	3.8	33.1	63.1
	B	0.7	27.3	71.9
	DRB1	0.3	48.1	51.6
3,000	A	5.4	47.0	47.6
	B	1.4	40.4	58.2
	DRB1	1.0	48.8	50.2

and observed data (Additional file 3: Table S3) for each pair of statistics further show the goodness of the Overlapping model to reproduce simultaneously different statistics, and the model fits better the HLA-B and -DRB1 loci than the HLA-A locus. In conclusion, the results allow us to reject the Southern-origin model as a relevant hypothesis to explain the East Asian HLA genetic variation. They also “decisively” [47] favor the Overlapping model over the Pincer model.

Both alternative models of selection (LOS and NFDS) give similar results for scenario comparison (Additional file 4: Table S4 and Additional file 5: Table S5). For all loci, the Overlapping scenario is the most likely (>52 % relative probability under LOS and 38-85 % under NFDS, respectively) and the Southern route scenario significantly less likely (<1 % under both LOS and NFDS). The only exception is HLA-A under NFDS which reaches a support up to 28 % while still being the worst scenario. The Pincer model is always intermediate, with a relative probability between 30-47 % under LOS and 14-44 % under NFDS, respectively.

Posterior distribution of parameters

For the most likely scenario (Overlapping) we have computed the Bayes factors between the selection models SOS and LOS and we found a decisive support to SOS over LOS for all loci (Bayes factors > 2,700). Consequently, we estimated the parameters under the Overlapping scenario and SOS selection model. According to the RMSEP charts (Additional file 6: Table S6), two locus-related parameters, i.e. the initial number of lineages (A) and the selection coefficient (s), affect most statistics under the Overlapping model. Regarding demography, the duration of admixture (T) appears as an essential factor, and the product of population density by migration rate (Nm) also has considerable influence. When N and m are treated separately, we notice that m , rather than N , is the decisive factor. By contrast, the growth rate (r) has very slight influence on the results, so do the other parameters. The posterior distribution of these parameters was estimated with 6 PLS under the Overlapping model, which resulted in relatively high p -values (>29 %).

Table 4 Model comparison using marginal density and Bayes' factor. Marginal density and p -values were output by ABCtoolbox for each of the three models, while the Bayes' factors were computed between each pair of models. (3,000 best simulations retained (1 %) from 300,000 simulations, 100,000 for each model)

Model	Locus	HLA-A	HLA-B	HLA-DRB1
Southern-origin model	<i>p-value</i>	0	0	0
	<i>Marginal density</i>	3.58E-27	1.93E-90	198
Pincer model	<i>p-value</i>	0.01	0.05	0.08
	<i>Marginal density</i>	0.79	1.17E4	1.22E6
Overlapping model	<i>p-value</i>	0.07	0.08	0.24
	<i>Marginal density</i>	8.20E3	6.00E7	4.99E10
Bayes' factor in favor of Pincer model to the Southern-origin model		2.20E26	6.06E93	6.16E3
Bayes' factor in favor of Overlapping model to the Southern-origin model		2.29E30	3.12E97	2.52E8
Bayes' factor in favor of Overlapping model to the Pincer model		1.04E4	5.12E3	4.09E4

The graphical representations of 95 % posterior distributions of the estimated parameters of interest are reported in Fig. 4. The estimation of the demographic parameters is consistent among the different loci: Nm in NEA is considerably higher compared to Nm in SEA (Fig. 4d, f), while the most probable T values lie between 1,193 and 1,363 generations, representing at least 24,000 years (Fig. 4c). By contrast, the estimated values of A and s clearly differ among the loci. While A is related to the number of lineages of the corresponding locus (Fig. 4a), s reveals that balancing selection shaped the three HLA loci differently, the effect of which is greater for HLA-B and HLA-DRB1 compared to HLA-A (Fig. 4b).

Validation of results

The validation procedure using the Kolmogorov-Smirnov test (Additional file 7: Table S7) confirmed that the prior distribution was not biased for any parameter except the number of lineages A , the latter revealing a much larger prior distribution compared to the posterior distribution, because an identical prior distribution was defined for the three different HLA loci (Table 1). Plots of simulated, retained and observed data for each pair of statistics further show a good coverage of posterior distribution by the prior distribution of parameters, which is true for each locus (Additional file 3: Table S3).

Discussion

Effectiveness of our simulation approach and reliability of results

The peopling scenarios we simulated have been designed to capture the main processes as simply as possible in order to limit considerably the parameters and thus increase the likelihood of ABC estimation. The prior distributions of parameters were carefully designed in order to create a set of non-biased simulated data. Comparison of models has been performed both by the model choice acceptance method and by using Bayes factors between each pair of models, two different ways leading to identical conclusions. One advantage of the Bayesian approach used here (ABC) is that models with different numbers of parameters can be directly compared thanks to the prior distributions. Finally, the parameters were estimated under the most likely model, which has been shown to be reliable according to independent validation procedures.

A significant contribution of the Northern route

Our results show that the Southern-origin model can be rejected as it is almost never able to reproduce the observed data in East Asia, whatever the initial conditions are. A formal comparison of the three models using the ABC approach indicates that both two-route models (Pincer and Overlapping) are significantly more likely

than the Southern-origin model (>96 % among 1,500 retained simulations) for the three HLA loci (Table 3). The result of this direct comparison is confirmed by the “decisive” [47] Bayes factor (>6,000 for all loci, Table 4). These simulation results thus very strongly support the crucial role of the Northern route in the peopling history of East Asia: human expansion(s) through a unique Southern route is not sufficient to explain the observed HLA genetic patterns.

When comparing our results to those obtained with putatively neutral variability, we find that, interestingly, a recent study on abundant Y chromosome data from China reached a similar conclusion to ours, supporting a main Southern route of migration for the peopling of East Asia but also an ancient Northern route, following the LGM period about 18,000 years ago, for the ancestors of current East Asian males [48]. In addition, a recent paper by Zhao et al. [49] combining new mtDNA data from 89 to 3,000 years old human remains of northern China with other published genetic data showed that the genetic structure of East Asian populations into a northern and a southern group was already shaped 3,000 years ago, suggesting a substantial mtDNA genetic contribution of Northern East Asian populations to the East Asian genetic pool in the past. Another study investigating genome-wide SNP diversity in Central and East-Asia [50] found (like us) that haplotype diversity is strongly correlated with latitude as a result of prehistoric population divergence, but suggested that the geographic sources contributing to East Asian populations were mainly from South East Asia with a minor contribution from Central Asia. Nevertheless, the dataset used by these authors is hardly comparable to ours because it is based on the analysis of 10 “combined” populations, among which only 4 correspond to Chinese groups and North-East Asian are largely underrepresented, as emphasized in a former paper [22]. Therefore, our results tend to reflect those of related studies carried out on putatively neutral markers by supporting a two-route model of migration in East Asia, with an emphasis on the role of the second route north to the Himalaya. This suggests that the influence of demography on HLA genes is stronger than selection when studying evolutionary processes at this geographical and temporal scales, as further discussed below.

Further comparisons between the two two-route models indicate a greater likelihood of the Overlapping model compared to the Pincer model, according both to the acceptance method (Table 3) and to the Bayes factor (>5,000 thus “decisive” [47], Table 4). This is in agreement with the previous conclusions based on a detailed analysis of HLA lineage and allele frequency variation [10] and on archaeological data [51, 52]. In addition, under the Overlapping model, the p -values computed on

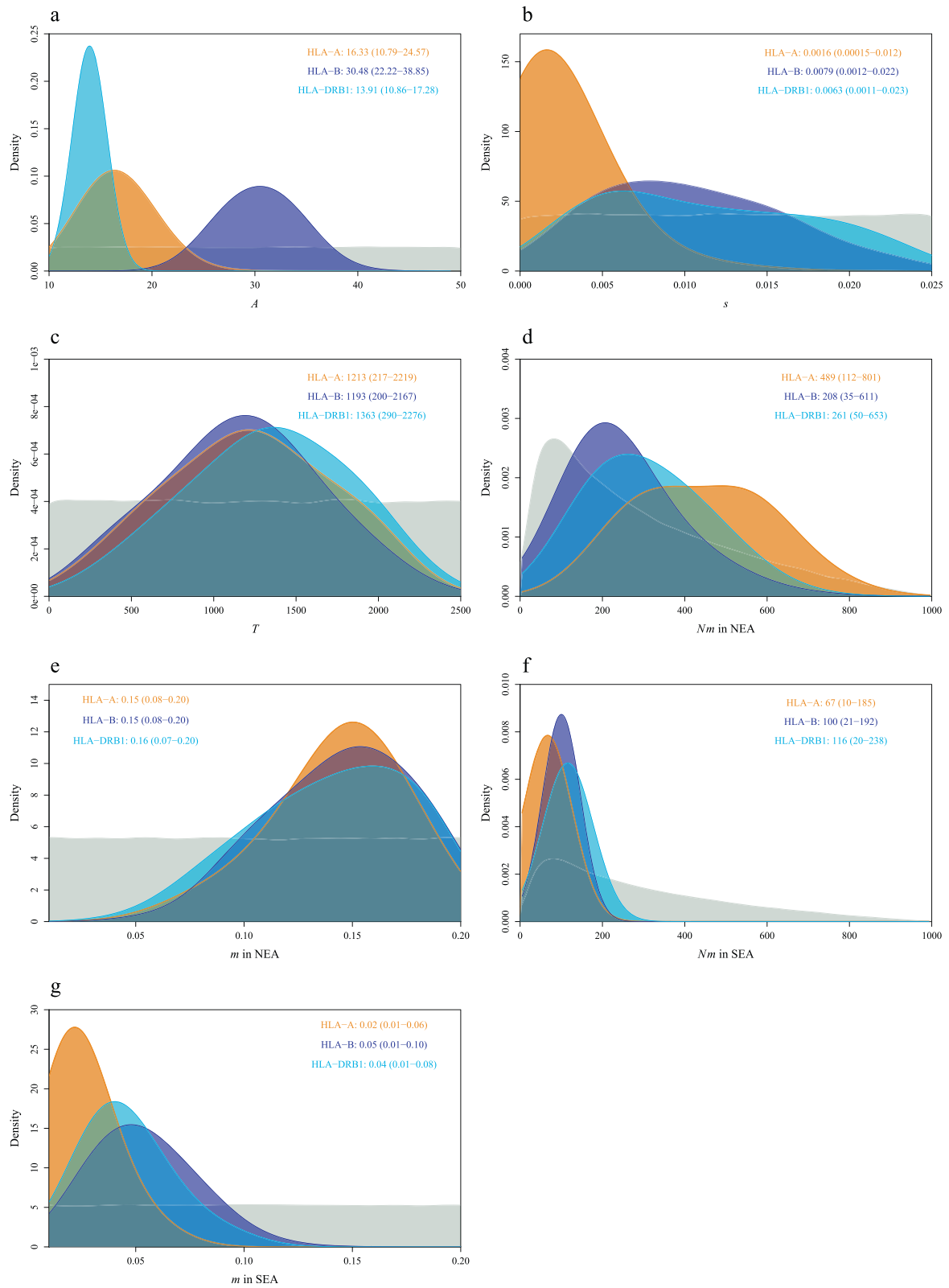


Fig. 4 (See legend on next page.)

(See figure on previous page.)

Fig. 4 Posterior distribution of parameters estimated with 6 PLS under the Overlapping model for HLA-A, -B, -DRB1 loci. Mode of the distribution and 95 % HPD Interval are given for **(a)** allele number A , **(b)** selection rate s , **(c)** admixture time T , as well as **(d,f)** product of population density and migration rate Nm and **(e,g)** migration rate m in NEAs and in SEAs (orange curve: HLA-A, blue curve: HLA-B, light blue curve: HLA-DRB1; grey curve: the prior distribution)

the raw statistics are always above 5 %, which is not the case under the Pincer model (1-8 %).

These results are robust to the choice of balancing selection model, thus accounting for the uncertainty of the mode of selection on HLA [53]. Indeed, a model of Symmetric Overdominant Selection varying in space (LOS) or Negative Frequency-Dependent Selection (NFDS) gives similar scenario comparisons to those obtained with the model of uniform heterozygote advantage (SOS) presented here. Indeed, under both alternative selection models and for all loci, the Overlapping scenario is the most likely and the Southern route scenario is (almost always significantly) the less supported (see Additional file 4: Table S4 and Additional file 5: Table S5).

North-south genetic variation

The genetic differences between NEAs and SEAs observed in numerous studies are evidenced by both among- and within-population variation. Among populations, significant north-south differentiation can be reproduced under any of the three models but is much more frequently observed under the two-route models (Fig. 3a). By contrast, greater genetic diversity (H) within populations in SEA is systematically produced under the Southern-origin model, which is incompatible with the observed data. At this level, differences between NEAs and SEAs can only be explained by the two-route models (Fig. 3b).

Evolutionary and demographic factors

Under the best of the three models considered, the Overlapping model, we estimated the effects of several evolutionary factors in shaping the genetic patterns. Based on the RMSEP charts, the balancing selection coefficient s has a significant impact on the simulated genetic structure (Additional file 6: Table S6). The estimated s values are close to each other for HLA-B and HLA-DRB1 under the Overlapping model, with point estimates of 0.79 % (0.12 %-2.2 %) and 0.63 % (0.11 %-2.3 %), respectively (Fig. 4b), while it is only of 0.16 % (0.015 %-1.2 %) for HLA-A. These results are in agreement with the ranking of HLA loci based on balancing selection reported previously [24, 29, 54]. On the other hand, the absolute values that we estimated for selection rates are much lower compared to these studies, where estimated s were of 1.5-2.2 %, for HLA-A, 4.2-4.4 %, for HLA-B, and 1.9-2.2 %, for HLA-DRB1, respectively [16, 54-55]. This discrepancy could be due to the fact that we used HLA lineages rather than alleles (used by the other

authors), suggesting that natural selection would have left a more visible signature on the latter despite their much more recent origin [32] compared to the long-term evolution of lineages [53, 56]. Alternatively, the observed discrepancy could result from differences in the methods used to estimate selective coefficients, our approach comprising a larger set of parameters including demography, but, on the other hand, no molecular variables. In any case, the selective coefficients estimated for the different HLA loci (regardless of the level of resolution considered to define lineages or alleles) are lower than for other genes known to undergo pathogen-mediated balancing selection, like G6PD/A- (s of about 10-20 %) [57] or HbC (4-9 %) [58], or positive directional selection, like Fy*O (6.6 %) [59] or lactase persistence (LP, up to 15.9 %, depending on the studies [60]). The present results suggest that (at least for HLA lineages) the impact of demographic factors on the East Asian HLA genetic variation overcame to a large extent the impact of balancing selection, as also revealed at the global scale by significant correlations between HLA population differentiations and geography resulting from population migrations [29]. Also, the demographic effect was particularly stronger on HLA-A, as supported by lower compatibility (p -value) of the Overlapping model at this locus compared to HLA-B and -DRB1 with original statistics (Table 4) and 6 PLS. The compatibility at HLA-A can only be improved when the effect of balancing selection is largely excluded by using less than 5 PLS (Additional file 6: Table S6). This hypothesis is congruent with previous results indicating a nearly neutral mode of evolution for HLA-A [23, 24], with a lower genetic diversity of this gene compared to HLA-B and -DRB1 in different continents [29]. In East Asian populations, for example, some HLA-A lineages are very common, such as A*02 (>20 %), A*11 (>30 % in SEA) and A*24 (>20 % in NEA). By contrast, a higher genetic diversity is maintained at HLA-B and -DRB1 loci [10].

Of course, the selection coefficients estimated here are appropriate only for the selection model with which they have been estimated and could be quite different if HLA genes were affected by selective pressures varying in time and space, in link with environmental changes. The implementation of such complex selective pressures would involve many more assumptions and parameters (range and variance of selection coefficients, spatial patterns of selection, choice of association between selection values and specific HLA alleles) and would constitute an exciting but difficult further development

of this study. However, we believe that the three selection models investigated here constitute good approximations for representing the average effect of balancing selection acting on HLA and we showed that the model comparison is robust to the selection model.

For the three HLA loci under study, Nm is considerably higher in NEA (between 208 and 489) than in SEA (between 67 and 116), even if the highest probability density (HPD) intervals are wide and partially overlapping (Fig. 4d,f). By treating separately N and m , we notice that m is the decisive factor explaining this different Nm (Additional file 6: Table S6). The estimated m values for SEA are much lower compared to NEA (Fig. 4e,g), supporting the hypothesis, proposed earlier, that modern humans colonizing East Asia via the Southern route underwent significant founder effects, or passed through periods of isolation with very rapid genetic drift [10]. Human migrations may have been more limited due to the mountainous reliefs of SEA, while long-range migrations in NEA would have been favored by the open landscapes of vast plateaus and plains. As a consequence, a greater differentiation among populations is now observed in SEA: the populations speak languages belonging to many distinct language families (Sino-Tibetan, Miao-Yao, Tai-Kadai and Austro-Asiatic), and are highly differentiated genetically within each of these families [10]; by contrast, in NEA the Altaic- and Mandarin-speaking populations are fairly homogeneous from a genetic point of view [10].

To summarize, despite significant but different effects of balancing selection acting on HLA-A, -B and -DRB1, similar conclusions are drawn for these three HLA loci concerning the likelihood of the different scenarios tested, suggesting that demography shaped the observed genetic patterns to a greater extent than did selection. These results support the hypothesis that the intensity of natural selection which acted on these HLA loci was too weak to blur the main signals of past demographic events of East Asia. Similar observations have also been found in Europe [23].

Genetic boundaries

We tested the existence and location of two geographic areas, hereafter named “genetic boundaries”, across which populations underwent low gene flow in the past. They could have represented areas where populations originating from different routes came into contact and exchanged few migrants, or where a significant genetic divergence appeared among related populations due to some specific reasons. We first considered, as a potential boundary, the 29-33°N zone, which represents an area between the 0 °C isotherm in winter (Qinling-Huaihe line) and an important geographic barrier (Yangtze River). Indeed, according to several studies using different genetic markers, the current genetic boundary between NEAs and SEAs was located in

this zone [10, 61-62]. Actually, the Overlapping model suggests that this boundary, regardless of its precise location, had been formed more recently by the creation of political subdivisions, while a more ancient boundary was located more to the north [10]. The high likelihood obtained in the present study for this model strongly supports the existence of an ancient boundary, in NEA (41-45°N zone), between the populations originating from the two migration routes, and the emergence, at a more recent time, of a genetic boundary around Qinling-Huaihe line or Yangtze River (29-33°N zone). The latter may have been due to the political subdivisions coinciding with the diversification of Chinese languages [37]. It is possible that, during a long period in the last Ice Age, the Mongolian Plateau and Siberia were not conducive to Southern populations (coming from the Southern route) practicing hunting and gathering in tropical and subtropical environments, whereas populations from the Northern route were better adapted to these harsh environments [63]. The separation between these populations would have thus been related to different degrees of cultural adaptation to their environment rather than to geographical barriers such as mountain ranges.

Long-term genetic exchanges among NEAs and SEAs

The admixture time T measures the duration of genetic exchange between NEAs and SEAs since they met in East Asia. Under the Overlapping model, T appears as a critical parameter (Additional file 6: Table S6), and the most likely values estimated for T are very consistent among the three HLA loci (in generations: 1,213 for HLA-A, 1,193 for HLA-B and 1,363 for HLA-DRB1, Fig. 4c), corresponding to about 24,000-36,000 years depending on the generation time considered (20-30 years). The posterior distributions for the three estimated values are relatively large, but agree among the loci (95 % HPD Interval: 217-2,219 for HLA-A, 200-2,167 for HLA-B and 290-2,276 for HLA-DRB1, Fig. 4c). Nevertheless, even the lower limits of these ranges are greater than 4,000 years, reflecting that a short-term genetic exchange is very unlikely to have created the existing genetic structure. These results contradict the assumption of some scholars who explain the observed genetic patterns by recent gene flow from Northwest Asia through the Silk Road, about only 2,000 years ago [3, 5]. Actually, they suggest a longer and more extensive overlap between NEA and SEA populations than estimated so far with neutral data, which may either reflect a stronger influence of a Northern route that is not captured by these neutral data or extensive geographic differences in pathogen selection at HLA genes that is not captured by the selection models used in our study.

For reasons of simplification, we decided to simulate the human expansions by the two routes simultaneously,

but our results do not exclude that the Southern route was older, as suggested by some genetic [5, 64] and archaeological [65] data. However, the T values that we estimated imply that human migrations via the Northern route date back at least to the Neolithic, and more probably to the Upper Paleolithic. The genetic admixture between the NEAs and SEAs may have initiated just after, or even before the Last Glacial Maximum (23–21 kya) [66]. Genetic differences between the early populations originating from the two migration routes would be, to some extent, reflected by the recent analyses of ancient human DNA from Tianyuan Cave (~40 kya) in northern China [67], on one side, and from Mal'ta (~24 kya) and Afontova Gora-2 (~17 kya) in Southern Siberia [68], on the other side (Fig. 1 and Additional file 2: Table S2). Also, the high degree of morphological diversity observed between individuals found in the Upper cave (~18 kya) [69], several kilometers from Tianyuan cave, would be explained by early contact of different populations in NEA, as suggested by the Overlapping model. This scenario is further consistent, on a larger geographical scale, to a model of multiple dispersals Out-of-Africa recently tested and supported by genomic and cranial phenotype data [70].

Conclusions

In this study, we simulated different scenarios of human peopling history in East Asia, i.e. the Southern-origin model (one-route), the Pincer model (two-route) and the Overlapping model (also two-route). Our quantitative analyses of 100,000 simulations for each model show that the Southern-origin model is virtually unable to reproduce genetic data compatible with the observations. The two-route models taking into account human colonization into the region via both a Southern and a Northern route are significantly more likely. Moreover, the results are in favor of the Overlapping model compared to the Pincer model.

In addition, our study reveals that evolutionary mechanisms related to the demographic history of East Asian populations had major effects on the observed genetic patterns compared to natural selection acting on HLA, which is weak for the three HLA loci (and especially so for HLA-A). In particular, we found a significantly lower migration rate among populations in SEA than in NEA, suggesting that SEAs underwent severe bottlenecks or periods of substantial isolation, likely due to the mountainous reliefs of this region. Our results also indicate the emergence of a genetic boundary characterized by limited gene flow between NEAs and SEAs around the Qinling-Huaihe line or Yangtze river (29–33°N zone), representing a limit of an “Altaicization” of northern Chinese languages proposed by some linguists [71] and likely related to imperial age political subdivisions, while such a boundary would have existed further north (41–45°N zone) in a more remote period. Finally, the genetic exchanges

between NEAs and SEAs, which took place when the two migration routes overlapped in East Asia, probably lasted a very long time which may trace back at least to the Neolithic but more probably to the Upper Paleolithic. All these events eventually gave rise to the pronounced but continuous HLA genetic variation currently observed between Northern and Southern East Asian populations.

Availability of supporting data

The references of all genetic data used in this study are given as supporting Additional file 1: Table S1. The program SELECTOR is freely available at <http://ua.unige.ch/en/agp/tools/selector/>.

Additional files

Additional file 1: Table S1. Information of populations sampled for HLA-A, -B, -DRB1 loci and cited in this study. (PDF 81 kb)

Additional file 2: Table S2. Information of majorModel comparison using the Latitudinal Overdominant Selection model (LOS Upper Paleolithic archaeological sites with human remains in East Asia. (PDF 87 kb)

Additional file 3: Table S3. Distribution of simulated, retained and observed values of each pair of statistics for each HLA locus. (PDF 1316 kb)

Additional file 4: Table S4. Model comparison using the Latitudinal Overdominant Selection model (LOS). (PDF 59 kb)

Additional file 5: Table S5. Model comparison using the Negative Frequency-Dependent Selection model (NFDS). (PDF 75 kb)

Additional file 6: Table S6. RMSEP charts of the parameters under the Overlapping model. (PDF 158 kb)

Additional file 7: Table S7. Results of Kolmogorov-Smirnov test against a uniform distribution. (PDF 115 kb)

Abbreviations

ABC: approximate Bayesian computation; HLA: human leukocyte antigen; LOS: latitudinal overdominant selection; NEA: North of East Asia; NEAs: Northern East Asian populations; NFDS: negative frequency-dependent selection; PLS: partial least squares; RMSEP: root mean squared error prediction; SEA: South East Asia; SEAs: Southern East Asian populations; SNPs: single nucleotide polymorphisms; SOS: symmetric overdominant selection; STRs: short tandem repeats.

Competing interests

The authors declare that they have no competing interests.

Authors' contributions

DD, ASM and MC designed the study. DD collected the HLA data and carried out simulations and ABC analyses. MC designed and modified the SELECTOR program. DD, ASM and MC wrote and approved the final manuscript.

Acknowledgements

We are indebted to Nabil Abdennadher and Mohamed Ben Belgacem for their help during our use of computation platform EZ-Grid. We also thank José Manuel Nunes for his invaluable advices on the development of efficient algorithms in SELECTOR. We are grateful for Daniel Wegmann's helpful suggestions concerning ABC estimations and ABCtoolbox. We would also like to thank Laurent Sagart and Li Fa-jun for their fruitful discussions in linguistics and archaeology. This study was supported by the Swiss National Foundation, grants No. 31003A-112651, 127465 and 144180 to ASM and No. 31003A-156853 to MC.

Author details

¹Department of Genetics and Evolution - Anthropology Unit, Laboratory of Anthropology, Genetics and Peopling history (AGP lab), University of Geneva, 12 rue Gustave-Reveillod, Geneva CH-1211, Geneva 4, Switzerland. ²Institute of Genetics and Genomics in Geneva (IGE3), University of Geneva Medical Centre (CMU), 1 rue Michel-Servet, Geneva CH-1211, Geneva 4, Switzerland.

Received: 12 March 2015 Accepted: 19 October 2015

Published online: 04 November 2015

References

- McDougall I, Brown FH, Fleagle JG. Stratigraphic placement and age of modern humans from Kibish. *Ethiopia Nature*. 2005;433(7027):733–6.
- Henn BM, Cavalli-Sforza LL, Feldman MW. The great human expansion. *Proc Natl Acad Sci U S A*. 2012;109(44):17758–64.
- Zhang F, Su B, Zhang YP, Jin L. Genetic studies of human diversity in East Asia. *Philos Trans R Soc Lond B Biol Sci*. 2007;362(1482):987–95.
- Lowery RK, Uribe G, Jimenez EB, Weiss MA, Herrera KJ, Regueiro M, et al. Neanderthal and Denisova genetic affinities with contemporary humans: Introgression versus common ancestral polymorphisms. *Gene*. 2013;530(1):83–94.
- Shi H, Dong YL, Wen B, Xiao CJ, Underhill PA, Shen PD, et al. Y-chromosome evidence of southern origin of the East Asian-specific haplogroup O3-M122. *Am J Hum Genet*. 2005;77(3):408–19.
- Shi H, Zhong H, Peng Y, Dong YL, Qi XB, Zhang F, et al. Y chromosome evidence of earliest modern human settlement in East Asia and multiple origins of Tibetan and Japanese populations. *BMC Biol*. 2008;6:45.
- Abdulla MA, Ahmed I, Assawamakin A, Bhak J, Brahmachari SK, Calacal GC, et al. Mapping human genetic diversity in Asia. *Science*. 2009;326(5959):1541–5.
- Xiao CJ, Du R, Cavalli-Sforza LL, Minch E. Principal component analysis of the gene frequencies of Chinese populations. *Science in China (Series C)*. 2000;30(4):434–41.
- Ding YC, Wooding S, Harpending HC, Chi HC, Li HP, Fu YX, et al. Population structure and history in East Asia. *Proc Natl Acad Sci U S A*. 2000;97(25):14003–6.
- Di D, Sanchez-Mazas A. Challenging views on the peopling history of East Asia: the story according to HLA markers. *Am J Phys Anthropol*. 2011;145(1):81–96.
- Di D, Sanchez-Mazas A. The peopling history of continental East Asia revealed by the HLA system. *Commun Contemp Anthropol*. 2011;5(1):273–80.
- Di D, Sanchez-Mazas A. HLA variation reveals genetic continuity rather than population group structure in East Asia. *Immunogenetics*. 2014;66(3):153–60.
- Spurgin LG, Richardson DS. How pathogens drive genetic diversity: MHC, mechanisms and misunderstandings. *P R Soc B*. 2010;277(1684):979–88.
- Prugnolle F, Manica A, Charpentier M, Guegan JF, Guernier V, Balloux F. Pathogen-driven selection and worldwide HLA class I diversity. *Curr Biol*. 2005;15(11):1022–7.
- Sanchez-Mazas A, Lemaître 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(1590):830–9.
- Currat M, Poloni ES, Sanchez-Mazas A. Human genetic differentiation across the Strait of Gibraltar. *BMC Evol Biol*. 2010;10:237.
- Hartl DL, Clark AG, editors. *Principles of population genetics*. 4th ed. Sunderland, Massachusetts: Sinauer Associates, Inc; 2007.
- Ammerman A, Cavalli-Sforza LL. *The neolithic transition and the genetics of populations in Europe*. Princeton: Princeton University Press; 1984.
- Novembre J, Stephens M. Interpreting principal component analyses of spatial population genetic variation. *Nat Genet*. 2008;40(5):646–9.
- Reich D, Price AL, Patterson N. Principal component analysis of genetic data. *Nat Genet*. 2008;40(5):491–2.
- Currat M, Excoffier L. The effect of the Neolithic expansion on European molecular diversity. *Proc Biol Sci*. 2005;272(1564):679–88.
- Sanchez-Mazas A, Di D, Riccio ME. A genetic focus on the peopling history of East Asia: critical views. *Rice*. 2011;4(3-4):159–69.
- Sanchez-Mazas A, Buhler S, Nunes JM. A new HLA map of Europe: regional genetic variation and its implication for peopling history, disease-association studies and tissue transplantation. *Hum Hered*. 2013;76:162–77.
- Solberg OD, Mack SJ, Lancaster AK, Single RM, Tsai Y, Sanchez-Mazas A, et al. Balancing selection and heterogeneity across the classical human leukocyte antigen loci: A meta-analytic review of 497 population studies. *Hum Immunol*. 2008;69(7):443–64.
- Beaumont MA, Zhang W, Balding DJ. Approximate Bayesian computation in population genetics. *Genetics*. 2002;162(4):2025–35.
- Marjoram P, Tavaré S. Modern computational approaches for analysing molecular genetic variation data. *Nat Rev Genet*. 2006;7(10):759–70.
- Currat M, Excoffier L. Strong reproductive isolation between humans and Neanderthals inferred from observed patterns of introgression. *Proc Natl Acad Sci U S A*. 2011;108(37):15129–34.
- Kimura M, Weiss GH. The Stepping Stone Model of Population Structure and the Decrease of Genetic Correlation with Distance. *Genetics*. 1964;49(4):561–76.
- Buhler S, Sanchez-Mazas A. HLA DNA sequence variation among human populations: molecular signatures of demographic and selective events. *PLoS One*. 2011;6(2), e14643.
- Klitz W, Hedrick P, Louis EJ. New reservoirs of HLA alleles: pools of rare variants enhance immune defense. *Trends Genet*. 2012;28(10):480–6.
- Bergstrom TF, Josefsson A, Erlich HA, Gyllensten U. Recent origin of HLA-DRB1 alleles and implications for human evolution. *Nat Genet*. 1998;18(3):237–42.
- von Salome J, Gyllensten U, Bergstrom TF. Full-length sequence analysis of the HLA-DRB1 locus suggests a recent origin of alleles. *Immunogenetics*. 2007;59(4):261–71.
- Wegmann D, Leuenberger C, Excoffier L. *Using ABCtoolbox (Manual)*. 2009.
- Chen S, Hu Q, Xie Y, Zhou L, Xiao C, Wu Y, et al. Origin of Tibeto-Burman speakers: evidence from HLA allele distribution in Lisu and Nu inhabiting Yunnan of China. *Hum Immunol*. 2007;68(6):550–9.
- Wang Z, editor. *Zhongguo Minzu Shi*. Beijing: Zhongguo Shehui Kexue; 1994.
- Nakaoka H, Mitsunaga S, Hosomichi K, Shyh-Yuh L, Sawamoto T, Fujiwara T, et al. Detection of ancestry informative HLA alleles confirms the admixed origins of Japanese population. *PLoS One*. 2013;8(4), e60793.
- Sagart L. Sino-Tibetan-Austronesian: an updated and improved argument. In: Sagart L, Blench R, Sanchez-Mazas A, editors. *The Peopling of East Asia: Putting Together Archaeology, Linguistics and Genetics*, vol. 1. London and New York: RoutledgeCurzon; 2005. p. 161–76.
- Sanchez-Mazas A, Poloni ES, Jacques G, Sagart L. HLA genetic diversity and linguistic variation in East Asia. In: Sagart L, Blench R, Sanchez-Mazas A, editors. *The Peopling of East Asia: Putting Together Archaeology, Linguistics and Genetics*, vol. 1. London and New York: RoutledgeCurzon; 2005. p. 273–96.
- Excoffier L, Lischer HE. Arlequin suite ver 3.5: a new series of programs to perform population genetics analyses under Linux and Windows. *Mol Ecol Resour*. 2010;10(3):564–7.
- Joyce P, Marjoram P. Approximately sufficient statistics and bayesian computation. *Stat Appl Genet Mol Biol*. 2008;7(1):Article26.
- Wegmann D, Leuenberger C, Excoffier L. Efficient approximate Bayesian computation coupled with Markov chain Monte Carlo without likelihood. *Genetics*. 2009;182(4):1207–18.
- Pritchard JK, Seielstad MT, Perez-Lezaun A, Feldman MW. Population growth of human Y chromosomes: a study of Y chromosome microsatellites. *Mol Biol Evol*. 1999;16(12):1791–8.
- Boulesteix AL, Strimmer K. Partial least squares: a versatile tool for the analysis of high-dimensional genomic data. *Brief Bioinform*. 2007;8(1):32–44.
- Leuenberger C, Wegmann D. Bayesian computation and model selection without likelihoods. *Genetics*. 2010;184(1):243–52.
- Cook SR, Gelman A, Rubin DB. Validation of software for Bayesian models using posterior quantiles. *J Comput Graph Stat*. 2006;15(3):675–92.
- Belgacem MB, Abdennadher N, Niinimäki M. Virtual EZ Grid: A Volunteer Computing Infrastructure for Scientific Medical Applications. In: *GPC 2010*. Taiwan; 2010: 385–394.
- Jeffreys H. *Theory of Probability*. 3rd ed. Oxford: Oxford University Press; 1998.
- Zhong H, Shi H, Qi XB, Duan ZY, Tan PP, Jin L, et al. Extended Y Chromosome Investigation Suggests Postglacial Migrations of Modern Humans into East Asia via the Northern Route. *Mol Biol Evol*. 2011;28(1):17–27.
- Zhao YB, Zhang Y, Zhang QC, Li HJ, Cui YQ, Xu Z, et al. Ancient DNA reveals that the genetic structure of the northern Han Chinese was shaped prior to 3,000 years ago. *PLoS One*. 2015;10(5), e0125676.
- Abdulla MA, Ahmed I, Assawamakin A, Bhak J, Brahmachari SK, Calacal GC, et al. Mapping Human Genetic Diversity in Asia. *Science*. 2009;326(5959):1541–5.
- Chen DZ, Zhang JZ. The physical characteristics of the early neolithic human in Jiahu site and the comparison with other neolithic humans and modern man. *Acta Anthropologica Sinica*. 1998;17(3):191–211.
- Wu XJ, Liu W, Bae CJ. Craniofacial variation between southern and northern Neolithic and modern Chinese. *Int J Osteoarchaeol*. 2012;22(2012):98–109.

53. Bernatchez L, Landry C. MHC studies in nonmodel vertebrates: what have we learned about natural selection in 15 years? *J Evol Biol.* 2003;16(3):363–77.
54. Satta Y, O'HUigin C, Takahata N, Klein J. Intensity of natural selection at the major histocompatibility complex loci. *Proc Natl Acad Sci U S A.* 1994;91(15):7184–8.
55. Yasukochi Y, Satta Y. Current perspectives on the intensity of natural selection of MHC loci. *Immunogenetics.* 2013;65(6):479–83.
56. Hughes AL, Yeager M. Natural selection at major histocompatibility complex loci of vertebrates. *Annu Rev Genet.* 1998;32:415–35.
57. Saunders MA, Slatkin M, Garner C, Hammer MF, Nachman MW. The extent of linkage disequilibrium caused by selection on G6PD in humans. *Genetics.* 2005;171(3):1219–29.
58. Wood ET, Stover DA, Slatkin M, Nachman MW, Hammer MF. The beta -globin recombinational hotspot reduces the effects of strong selection around HbC, a recently arisen mutation providing resistance to malaria. *Am J Hum Genet.* 2005;77(4):637–42.
59. Hodgson JA, Pickrell JK, Pearson LN, Quillen EE, Prista A, Rocha J, et al. Natural selection for the Duffy-null allele in the recently admixed people of Madagascar. *Proc Biol Sci.* 2014;281(1789):20140930.
60. Gerbault P, Liebert A, Itan Y, Powell A, Currat M, Burger J, et al. Evolution of lactase persistence: an example of human niche construction. *Philos Trans R Soc Lond B Biol Sci.* 2011;366(1566):863–77.
61. Xue F, Wang J, Hu P, Ma D, Liu J, Li G, et al. Identification of spatial genetic boundaries using a multifractal model in human population genetics. *Hum Biol.* 2005;77(5):577–617.
62. Xue F, Wang Y, Xu S, Zhang F, Wen B, Wu X, et al. A spatial analysis of genetic structure of human populations in China reveals distinct difference between maternal and paternal lineages. *Eur J Hum Genet.* 2008;16(6):705–17.
63. Balaresque PL, Ballereau SJ, Jobling MA. Challenges in human genetic diversity: demographic history and adaptation. *Hum Mol Genet.* 2007;16 Spec No. 2:R134–139.
64. Macaulay V, Hill C, Achilli A, Rengo C, Clarke D, Meehan W, et al. Single, rapid coastal settlement of Asia revealed by analysis of complete mitochondrial genomes. *Science.* 2005;308(5724):1034–6.
65. Shen G, Wang W, Wang Q, Zhao J, Collerson K, Zhou C, et al. U-Series dating of Liujiang hominid site in Guangxi, Southern China. *J Hum Evol.* 2002;43(6):817–29.
66. Fagan B, editor. *The complete Ice Age: how climate change shaped the world.* London: Thames & Hudson; 2009.
67. Fu Q, Meyer M, Gao X, Stenzel U, Burbano HA, Kelso J, et al. DNA analysis of an early modern human from Tianyuan Cave, China. *Proc Natl Acad Sci U S A.* 2013;110(6):2223–7.
68. Raghavan M, Skoglund P, Graf KE, Metspalu M, Albrechtsen A, Moltke I, et al. Upper Palaeolithic Siberian genome reveals dual ancestry of Native Americans. *Nature.* 2014;505(7481):87–91.
69. Harvati K. Into Eurasia: a geometric morphometric re-assessment of the Upper Cave (Zhoukoudian) specimens. *J Hum Evol.* 2009;57(6):751–62.
70. Reyes-Centeno H, Ghirotto S, Detroit F, Grimaud-Herve D, Barbujani G, Harvati K. Genomic and cranial phenotype data support multiple modern human dispersals from Africa and a southern route into Asia. *Proc Natl Acad Sci U S A.* 2014;111(20):7248–53.
71. Sagart L. The formation of East Asian Language families: a partial scenario. In: *Languages and genes: recent work and emerging results:* 22–25 September Aussois. 2005.

Submit your next manuscript to BioMed Central and take full advantage of:

- Convenient online submission
- Thorough peer review
- No space constraints or color figure charges
- Immediate publication on acceptance
- Inclusion in PubMed, CAS, Scopus and Google Scholar
- Research which is freely available for redistribution

Submit your manuscript at
www.biomedcentral.com/submit

