Immunogenetics (2019) 71:589–604
https://doi.org/10.1007/s00251-019-01144-7

ORIGINAL ARTICLE

# Single haplotype admixture models using large scale HLA genotype frequencies to reproduce human admixture

Alexandra Litinsky Simanovsky[1] · Abeer Madbouly[2] · Michael Halagan[2] · Martin Maiers[2] · Yoram Louzoun[1]

Received: 17 April 2019 / Accepted: 19 October 2019 / Published online: 18 November 2019

## Abstract

The human leukocyte antigen (HLA) is the most polymorphic region in humans. Anthropologists use HLA to trace populations' migration and evolution. However, recent admixture between populations can mask the ancestral haplotype frequency distribution. We present a statistical method based on high-resolution HLA haplotype frequencies to resolve population admixture using a non-negative matrix factorization formalism and validated using haplotype frequencies from 56 world populations. The result is a minimal set of source components (SCs) decoding roughly 90% of the total variance in the studied admixtures. These SCs agree with the geographical distribution, phylogenies, and recent admixture events of the studied groups. With the growing population of multi-ethnic individuals, or individuals that do not report race/ethnic information, the HLA matching process for stem-cell and solid organ transplants is becoming more challenging. The presented algorithm provides a framework that facilitates the breakdown of highly admixed populations into SCs, which can be used to better match the rapidly growing population of multi-ethnic individuals worldwide.

**Keywords** HLA · Genetic admixture · Non-negative matrix factorization · Stem-cell donor registry · Unsupervised learning

Alexandra Litinsky Simanovsky and Abeer Madbouly contributed equally to this work.

**Author Summary** Anthropologists frequently use HLA to trace migration and evolution of different populations. This is due to the high linkage among HLA genes leading to the transmission of intact haplotypes from parents to offspring, hence preserving key population ancestral features. Such a linkage has been proposed to be the result of positive selection. However, over the last century, human kind is going through a rapid population mixture, producing admixtures of original populations, and algorithms are required to disentangle these populations for population genetics and for improving matching for stem-cell or solid organ transplants.

To address these challenges, we developed a new HLA-based method to identify population-level admixture using high-resolution HLA haplotype frequencies. This is in contrast with existing methods estimating individual-level admixture based on genomic distribution of SNPs.

We show that 90% of the genetic variance of the HLA haplotypes can be explained using a much-reduced set of 8 source components (SCs). The estimated SCs and admixture models agree with the geographical distribution, population phylogenies, and recent historic admixture events of the studied populations.

**Electronic supplementary material** The online version of this article (https://doi.org/10.1007/s00251-019-01144-7) contains supplementary material, which is available to authorized users.

✉ Yoram Louzoun
louzouy@math.biu.ac.il

Extended author information available on the last page of the article

## Introduction

Human Leukocyte Antigen (HLA) plays a crucial role in both the adaptive and innate immune system and is proven instrumental in multiple medical disciplines, including matching for solid organ and hematopoietic stem-cell transplantation (HSCT) (Chinen and Buckley 2010; Brand et al. 2013; Geneugelijk et al. 2017; Phillips and Callaghan 2017). As generations of humans migrated throughout the world, HLA has evolved with each population conserving key ancestral features (Sanchez-Mazas and Thorsby 2012), and with more than 15,000 defined alleles to date (Robinson et al. 2014) and over 1,000,000 haplotypes (Slater et al. 2015), it stands as the most polymorphic region in the human genome.

The study of human admixture and population stratification addresses a growing need in multiple disciplines, including large-scale disease and drug association studies (Gragert et al. 2014b; Ferrell and McLeod 2008), planning and modeling volunteer adult donor stem-cell registries (Gragert et al. 2014a), and personalized and consumer genetics (Chua and Kennedy 2012). Typical population genetics algorithms interrogate Single-Nucleotide Polymorphisms (SNPs) in large genomic or genome-wide regions to track concepts like population size, migration, selection, admixture, recombination, or genetic drift (Balding 2006; Voight et al. 2006; Sabeti et al. 2007; Williamson et al. 2007; Hindorff et al. 2009; Pickrell et al. 2009) or study the population genetics of specific ethnic groups (Kennedy et al. 2003; Fujimura and Rajagopalan 2010). The need for large genomic regions stems from the limited polymorphism in bi-allelic SNPs. While the HLA locus is indeed highly polymorphic, it is of limited size, allowing for different measures of admixture. With the evolution of HLA over millions of years, distinctive features of the MHC locus became evident, such as gene density, diversity, and low recombination rate (Traherne 2008; Shiina et al. 2009). Multiple models have been proposed to explain HLA polymorphism (Parham and Ohta 1996; Aguilar et al. 2004), the most prominent being pathogen-driven balancing selection (Apanius et al. 1997). This extensive polymorphism facilitates tracking human migration patterns, population admixture, allelic diversity, pathogen evolution, and selection through the study of HLA allele and haplotype frequencies (Tokunaga et al. 1996; Abi-Rached et al. 2011).

Importantly, because of the high-gene density, diversity, and low recombination rates, HLA alleles are inherited mostly as intact haplotypes, preserving a high degree of linkage disequilibrium (LD) (Lam et al. 2013). Recent evidence shows a clear purifying selection in HLA haplotypes (Alter et al. 2017), i.e., many combinations of HLA alleles are seldom separated over multiple generations, possibly as a result of positive selection (Alter et al. 2017; Lobkovsky et al. 2019). One interpretation is that certain allele groupings on a haplotype are tuned to work together. Hence, the absence of detectable recombination may reflect maintenance of haplotypes with favored immunological function through selection against unfavorable recombinant haplotypes (Traherne 2008), creating founder haplotypes which have not degraded over the development time scale of current self-defined human ethnicities. The high occurrence of a handful of distinct long-range HLA haplotypes has long been of interest because of their prevalence and strong associations with complex diseases (Simmonds and Gough 2007; Holoshitz 2013) (Sjakste et al. 2016).

Population genetic traits were studied through the estimation and analysis of HLA allele and haplotype frequencies. Klitz et al. (Klitz et al. 2010) studied detailed Expectation Maximization based haplotype frequencies across three HLA loci to estimate the phylogeny of Jewish populations using correlations between haplotype frequencies among populations. More detailed correlation based analysis of high resolution six-locus haplotype frequencies was performed by Gragert et al. (Gragert et al. 2013) where population groups and phylogeny was detected among the main US populations as defined in the Be The Match® US volunteer donor stem-cell registry. Recently, we studied the effect of selection on HLA haplotypes (Slater et al. 2015; Alter et al. 2017). However, to our knowledge there is no published work dividing HLA haplotypes into original components that break down the admixture in haplotype distributions.

Population substructure is often estimated using admixture models, where each population is broken down into a mixture of original populations (Pritchard et al. 2000; Alexander et al. 2009). This is formally a hidden state model, where the hidden states and the associated probability distribution are estimated in each population. To resolve the admixture into its primary components, most algorithms use a maximum likelihood approach (Alexander et al. 2009; Zhou et al. 2011). Multiple methods have been developed to detect population admixture (e.g., ADMIXTURE (Alexander et al. 2009) and STRUCTURE (Pritchard et al. 2000; Porras-Hurtado et al. 2013)). However, hidden state model methods can produce spurious results in the presence of background LD, which often happens in admixed populations due to multiple effects, such as genetic drifts and in some highly polymorphic markers, such as HLA and microsatellites (Kaeuffer et al. 2007). Additionally, these models are not adapted to a single highly polymorphic region in contrast to large regions of limited polymorphism. We here present a method to estimate population admixture, in the presence of background LD, based on Non-negative-Matrix Factorization (NMF). Using HLA haplotype frequencies from 56 world populations, we decompose the frequency distributions into the non-negative composition of a small number of Source Components (SCs) and test the precision of our results using a large dataset of high-resolution HLA haplotype frequencies, estimated from over 3.5 million individuals. Data were collected from volunteer unrelated adult stem-cell donor registries worldwide for haplotype frequency analysis and registry modeling studies. We then show that these SCs can be used to detect the admixture composition of new populations.

Simply stated, the proposed factorization presumes that each observed frequency is the weighted combinations of frequencies in a small number of SC. However, since the SCs are unknown, we solve an algebraic problem of finding the composition of a small number of components whose weights best explain the observed frequencies in all populations. This would translate to a decomposition problem. Since haplotype frequencies are always positive, and each current population cannot have a negative fraction of one of the seeding populations, all elements in the admixture must be either zero or positive. We thus use a variant of matrix decomposition denoted as non-negative matrix factorization.

Current HLA matching practices in HLA registries mainly rely on predefined ethnic groups driven by self-reported race and ethnicity (SIRE). The admixture method presented here can augment the current matching practices by contributing a genetically oriented approach, particularly in cases when SIRE is unavailable or with individuals of multi-ethnic origin. This would facilitate reducing the total number of ethnic groups being matched and permit the development of cross-registries haplotype frequency estimation. As the fraction of people with unknown SIRE or mixed SIRE keeps increasing, modeling approaches become increasingly important. The development of admixture methods in combination with a Bayesian classification scheme that would associate donors with different SCs, would allow both imputation and matching of donors in the absence of SIRE (except for a prior on their probability to be associated with each SC). Moreover, when the number of different populations is increased, using SCs would limit the number of combinations to be tested for donors with complex ancestry.

## Materials and methods

### Study populations and haplotype frequency estimation

HLA haplotype frequencies from 56 global populations were used in this study (Table 1). These populations were predominantly volunteer donors in stem-cell donor registries across the globe. We used haplotype frequencies from the Ezer-Mizion Registry in Israel (Manor et al. 2016) multiple registries in India (Maiers et al. 2014), the US Be The Match registry® (Gragert et al. 2013), a number of European registries that list their donors within the US registry (The Netherlands, Norway, Sweden, Wales) and the Australian registry, including two registries that list within it: New Zealand and Thailand. Haplotype frequencies from the Canadian registry were used for method validation (Table 2 (Gragert et al. 2013)). Five-locus HLA-A~C~B~DRB1~DQB1 allele-level haplotype frequencies were used for all populations except for the Canadian groups for which only four-locus data were available and HLA-A~C~B~DRB1 frequencies were estimated. Phased HLA haplotype frequencies were estimated using an implementation of the Expectation Maximization (EM) algorithm that resolves phase, allelic, and missing allele ambiguities (Kollman et al. 2007). The applied EM algorithm was shown to cope with varying levels of typing resolution or missing loci data, even if the level of typing resolution is not independent of the HLA type (for example higher rate of missing typing at particular loci like HLA-C and HLA-DQB1) .

### Non-negative matrix factorization

To estimate SCs, we use NMF. Such a factorization is used when the observed values are all non-negative, and we make two assumptions:

A) Each observed haplotype frequency in a given population is a combination of their frequencies in SC, and the fraction of each SC in the current population is always non-negative.

B) The SCs themselves have only non-negative frequencies.

Specifically, a non-negative matrix **C** would be factorized into two non-negative matrices **A** and **B**, where **C** is a $\boldsymbol{nXm}$ matrix, representing the haplotype frequencies in $\boldsymbol{n}$ populations, each with $\boldsymbol{m}$ haplotypes. Note that many haplotypes are missing in each population, these values are represented as zeros in the appropriate positions. **A** is $\boldsymbol{nXr}$ non-negative mixing matrix and **B** is $\boldsymbol{rXm}$ non-negative "original" populations (OP) matrix. Each $\boldsymbol{n}$-dimensional vector represents one out of $\boldsymbol{m}$ populations. In the current application, $\boldsymbol{m}$ constitutes the 56 studied populations and $\boldsymbol{n}$ is of the order of a million different haplotypes. The variable $\boldsymbol{r}$ is chosen to be much smaller than $\boldsymbol{n}$ and $\boldsymbol{m}$. The range of **r** values, representing the count of possible SCs, was tested between 1 and 23. The results of the decomposition are a function of the chosen cost function (Lee and Seung 2001). We have tested the Frobenius, Kullback-Leibler (KL), Lee, Offset, and LS-NMF (least squares non-negative matrix factorization) cost function as implemented in the R package NMF (Gaujoux and Seoighe 2010).

The factorization is initialized with a seed (e.g., $A_0$ and/or $B_0$), which initiates the expectation maximization (EM) process. In each iteration, new non-negative A and B matrices are calculated. The process is repeated until it converges to a locally optimal matrix factorization. There are currently multiple standard algorithms for NMF (Lee and Seung 2001). These algorithms approximate a non-negative matrix (a matrix where all elements are non-negative) as a product of two low-rank non-negative matrices. The results of the approximation are affected by the loss function used and by the weighting of the input columns. In the current context, we represent the haplotype frequencies in 56 populations as the positive admixture of 8 components.

### Phylogenetic analysis

To test that the admixture reproduces the known genetic groups of populations, we compared the admixture-based grouping to a fixation index (Fst) based phylogenetic tree. We applied a Neighbor-Joining (NJ) algorithm to construct a phylogenetic tree of the studied populations (Saitou and Nei 1987; Costa et al. 2017). Following the tree construction, populations were grouped by the SCs obtained from the NMF analysis, grouping highly overlapping SCs (e.g., two SCs frequent in all European populations). To validate the NMF algorithm, we investigated

**Table 1** Study populations, acroynyms used, and the sample size used to generate haplotype frequencies for each population

| Population | Abbreviation | Counts | Population | Abbreviation | Counts |
|---|---|---|---|---|---|
| US African American | US_AAFA | 416581 | Australia Aboriginal | AU_ABORIGINAL | 990 |
| US African | US_AFB | 28557 | Australia China | AU_CHINA | 869 |
| US South Asian Indian | US_AINDI | 185391 | Australia India | AU_INDIA | 1795 |
| US American Indian South or Central Am. | US_AISC | 5926 | Australia Jewish | AU_JEWISH | 2597 |
| US Alaska native or Aleut | US_ALANAM | 1376 | Australia Middle Eastern | AU_MIDEAST | 1265 |
| US North American Indian | US_AMIND | 35791 | Australia Northern Europe | AU_NEUR | 42486 |
| US Caribbean black | US_CARB | 33328 | Australia Northwestern Europe | AU_NWEUR | 2573 |
| US Caribbean Hispanic | US_CARHIS | 115374 | Australia South Europe | AU_SEUR | 4153 |
| US Caribbean Indian | US_CARIBI | 14339 | Australia Sri Lanka | AU_SRILANKA | 1796 |
| US European Caucasian | US_EURCAU | 1242890 | Ezer Mizion Arab | IL_ARAB | 12,300 |
| US Filipino | US_FILI | 50614 | Ezer Mizion Argentina | IL_ARGENTINA | 4307 |
| US Hawaiian or other Pacific Islander | US_HAWI | 11499 | Ezer Mizion Ashkenazi | IL_ASHKENAZI | 4625 |
| US Japanese | US_JAPI | 24582 | Ezer Mizion Bukhara | IL_BUKHARA | 2317 |
| US Korean | US_KORI | 77584 | Ezer Mizion Druze | IL_DRUZE | 5914 |
| US Middle Eastern or N. Coast of Africa | US_MENAFC | 70890 | Ezer Mizion Ethiopia | IL_ETHIOPIA | 5928 |
| US Mexican or Chicano | US_MSWHIS | 261235 | Ezer Mizion Georgia | IL_GEORGIA | 4471 |
| US Chinese | US_NCHI | 99672 | Ezer Mizion Iran | IL_IRAN | 8153 |
| US Hispanic South or Central American | US_CAHIS | 146714 | Ezer Mizion Iraq | IL_IRAQ | 13,270 |
| US Black South or Central American | US_SCAMB | 4889 | Ezer Mizion Israel | IL_ISRAEL | 69,716 |
| US Southeast Asian | US_SCSEAI | 27978 | Ezer Mizion Kavkaz | IL_KAVKAZ | 2840 |
| US Vietnamese | US_VIET | 43540 | Ezer Mizion Lybia | IL_LYBIA | 3739 |
| Netherlands | NL | 30516 | Ezer Mizion Morocco | IL_MOROCCO | 36,718 |
| New Zealand | NZ | 4385 | Ezer Mizion Poland | IL_POLAND | 13,871 |
| Sweden | SE | 14645 | Ezer Mizion South East European | IL_SEEUR | 11,179 |
| Thailand | TH | 136498 | Ezer Mizion Tunisia | IL_TUNISIA | 9070 |
| Wales | GB_WLS | 47870 | Ezer Mizion USA | IL_USA | 6058 |
| Norway | NO | 14460 | Ezer Mizion USSR | IL_USSR | 45,681 |
| India | IN | 26240 | Ezer Mizion Yemen | IL_YEMEN | 15,542 |

the similarity between the tree structure and the phylogenetic grouping.

The NJ distance matrix was based on the genetic distance between the population haplotype frequencies. We used the pairwise Fst measure (Weir and Hill 2002; Excoffier et al. 2005) calculated for all combinations of population pairs. Typically, the Fst value ranges from 0.0 to 1.0, where 0.0 indicates identical population frequencies and high Fst values indicate significantly different populations. We use the Fst algorithm as implemented

**Table 2** Populations from the Canadian OneMatch registry used for validation

| Population | Abbreviation | Count |
|---|---|---|
| Aboriginal~East | CA_E_ABORIGINAL | 1030 |
| Aboriginal~West | CA_W_ABORIGINAL | 2099 |
| ArabWestAsian | CA_ARAB_W_ ASIAN | 1997 |
| Asian | CA_ASIANC | 8091 |
| Black | CA_BLACK | 2542 |
| Caucasian | CA_CAUCASIAN | 268,990 |
| Chinese | CA_CHINESE | 19,425 |
| Filipino | CA_FILIPINO | 1026 |
| Hispanic | CA_HISPANIC | 2384 |
| Japanese | CA_JAPANESE | 190 |
| Korean | CA_KOREAN | 489 |
| SouthAsian | CA_S_ASIAN | 13,076 |
| SoutheastAsian | CA_S_E_ASIAN | 2312 |

by Weir et al. (Weir and Cockerham 1984; Excoffier et al. 1992; Weir and Ott 1997; Bird et al. 2011).

## Principal coordinate analysis

Principal Coordinate Analysis (PCOA) (also denoted multi-dimensional scaling) was applied to translate the distance matrix into a 3-dimensional Euclidean space. The projection was performed using the wcmdscale function implemented in the R package Vegan.

## Comparison with hidden state model algorithm

We compared the performance of our NMF-based admixture algorithm to the software STRUCTURE (Pritchard et al. 2000) that estimates population substructure using a Markov Chain Monte Carlo (MCMC) Model. Given that STRUCTURE produces spurious results for markers with excessive background LD (such as HLA), we ran STRUCTURE on select genome-wide SNP genotypes from the reference 1000 Genomes phase I dataset (Consortium 2015) (Table 3) and compared the output admixtures to those estimated by HLA-based SCs in populations of similar ancestry (in our study) as the 1000 Genome populations. The 500 selected genome-wide SNP ancestry informative marker panel was previously published and validated for delineation of population substructure within European populations as well as globally across world continents (Paschou et al. 2008). STRUCTURE v2.3.4 was run using K = 8 clusters (like the count of SCs), with 7000 replicates and 50,000 burnin' cycles.

## Likelihood estimate

Given the overall observed haplotype frequencies, each population haplotype frequency likelihood was estimated assuming independence of sampling of haplotypes under a Poisson model (assuming the composition reproduces frequencies, and that the total population size is much larger than the registry size). The log likelihood of a reconstruction $\widetilde{H} = AP$ of $H$ was estimated as $\sum_{i,j} \log\left(p\left(h_{i,j}|\widetilde{h}_{i,j}, N_i\right)\right)$, where $N_i$ is the registry size of population $i$, $j$ is the index of a haplotype and $p\left(h_{i,j}|\widetilde{h}_{i,j}, N_i\right) = \frac{e^{-\widetilde{h}_{i,j}*N_i}}{h_{i,j}!}\left(\widetilde{h}_{i,j}*N_i\right)^{h_{i,j}}$ using a Poisson distribution assumption. The Bayesian Information Criteria (BIC) and the Akaike information criterion (AIC) were

**Table 3** Phase I 1000 Genome population acronym details

| Population code | Population description |
|---|---|
| CHB | Han Chinese in Beijing, China |
| JPT | Japanese in Tokyo, Japan |
| CHS | Southern Han Chinese |
| CEU | Utah Residents (CEPH) with Northern and Western European Ancestry |
| TSI | Tuscany in Italia |
| FIN | Finnish in Finland |
| GBR | British in England and Scotland |
| IBS | Iberian Population in Spain |
| YRI | Yoruba in Ibadan, Nigeria |
| LWK | Luhya in Webuye, Kenya |
| ASW | Americans of African Ancestry in SW USA |
| MXL | Mexican Ancestry from Los Angeles USA |
| PUR | Puerto Ricans from Puerto Rico |
| CLM | Colombians from Medellin, Colombia |

then computed using the number of free parameters in A and P.

### Inference of unlabeled populations

A possible application of the work presented here is to dissect an unknown admixed population into a set of SCs. Starting with a set of estimated SC haplotype frequencies, one can decompose a new population of unknown admixture using a regression of these frequencies over the used original frequencies.

Often the new population may have lower resolution (e.g., 4 locus information instead of 5). In such a case, the original frequencies are reduced to a similar resolution prior to regression. We also show that this reduction does not impact the accuracy of the results.

## Model

We present a model to infer the basic components of a population using haplotype frequency distributions from multiple populations. The 56 studied populations originate from multiple ethnic groups in four continents with ancestral roots encompassing all six world continents. Some populations were relatively homogeneous (e.g., Vietnamese) while others were formed of relatively recent admixtures (e.g., African-American and Latin American) (Paschou et al. 2008; Behar et al. 2010). To estimate the components of a population admixture, we assume the existence of K SCs, each represented by a haplotype frequency denoted as $P_j$, where $j$ranges from 1 to K SCs. Each observed population is assumed to be a positive normalized combination of such SCs:

$$H_i = \sum_{j=1}^{K} a_{ij} P_j + \varepsilon_i, 0 \le a_{ij} \le 1, \sum_{j=1}^{K} a_{ij} = 1, \qquad (1)$$

where $\varepsilon_i$is a noise vector that is affected by the sample size and may vary for each haplotype. To translate (1) into a matrix decomposition problem, we rewrite it as:

$$H \approx AP, h_{ij}, a_{ik} p_{kj} \ge 0, \sum_{k=1}^{K} a_{ik} = 1, \qquad (2)$$

where A is the matrix of all $a_{ij}$. To estimate the number of SCs best describing the studied populations, an appropriate cost function is required. Such a function can be based on the error in the reconstructed populations ($E = H - AP$) and is scaled by the sample size of each population. Such scaling may be crucial, since some populations are over-represented in the study cohort. Moreover, if the population is not scaled, splitting a population into two sub-populations would double its contribution to the error. Other cost functions were also tested (See Supp. Mat. Table S1 for a list of tested cost functions and weightings). The relative error of the decomposition was computed for each algorithm as a function of the number of used OPs:

$$Err = \frac{\sum_i N_i \| E_i \|}{\sum_i N_i \| H_i \|}, \qquad (3)$$

Where $E_i$ and $H_i$ are column vectors of E and H respectively representing a single population, $N_i$is the population size, and the norm used is Euclidean. In all tested cost functions, the errors plateaued at 7-10 SCs (Fig. 1A). Note that non-scaled solutions produce SCs strongly biased toward highly admixed groups (Supp. Mat. Fig. S1) and produce a higher error rate (Fig. 1A). To test that 7–10 SCs is actually the optimal solution, we computed the Bayesian Information Criteria (BIC) and the Akaike information criterion (AIC) for each solution (Inset in Fig. 1A). Indeed, the optimal AIC and BIC are between 6 and 10 SCs (See methods for likelihood function).

Each SC is assigned a name according to the principal observed populations containing it, with the following main groups: European, African, Israeli, Latin American, South Asian, and East Asian. When more SCs are used, the result is typically the splitting of one of these groups into sub-groups (Supp. Mat. Fig. S1). Given the results of the admixture, we assign each observed component one main group according to the main SC composing them. In the following results, the observed components are grouped according to the SC composing the largest fraction of the admixture vector $A_i$(A single column of A) (Fig. 1B).

$$PP(i) = \underset{k}{\operatorname{argmax}} (a_{ik}). \qquad (4)$$

In all plots, populations are colored according to the SC with the $PP(i)$ most contributing to this population.

The model and core are available at: https://github.com/louzounlab/Haplotype-Admixture

## Results

Our results show that seven to ten SCs are the optimal range for the admixture in the 56 studied populations (Fig. 1A). The lowest AIC is obtained for seven SCs and the lowest BIC is obtained for 10 (see methods for likelihood estimate and inset in Fig. 1A). We have used an intermediate value of Eight SCs which reduces the error to 10% of the original matrix variance (as defined by Eq. 3), i.e., a minimal set of SCs captures roughly 90% of the total variance in the studied admixtures. We observed that the diversity level varied widely among populations; in some groups the admixture was predominantly represented by one SC as in the case of the US Japanese and Filipino

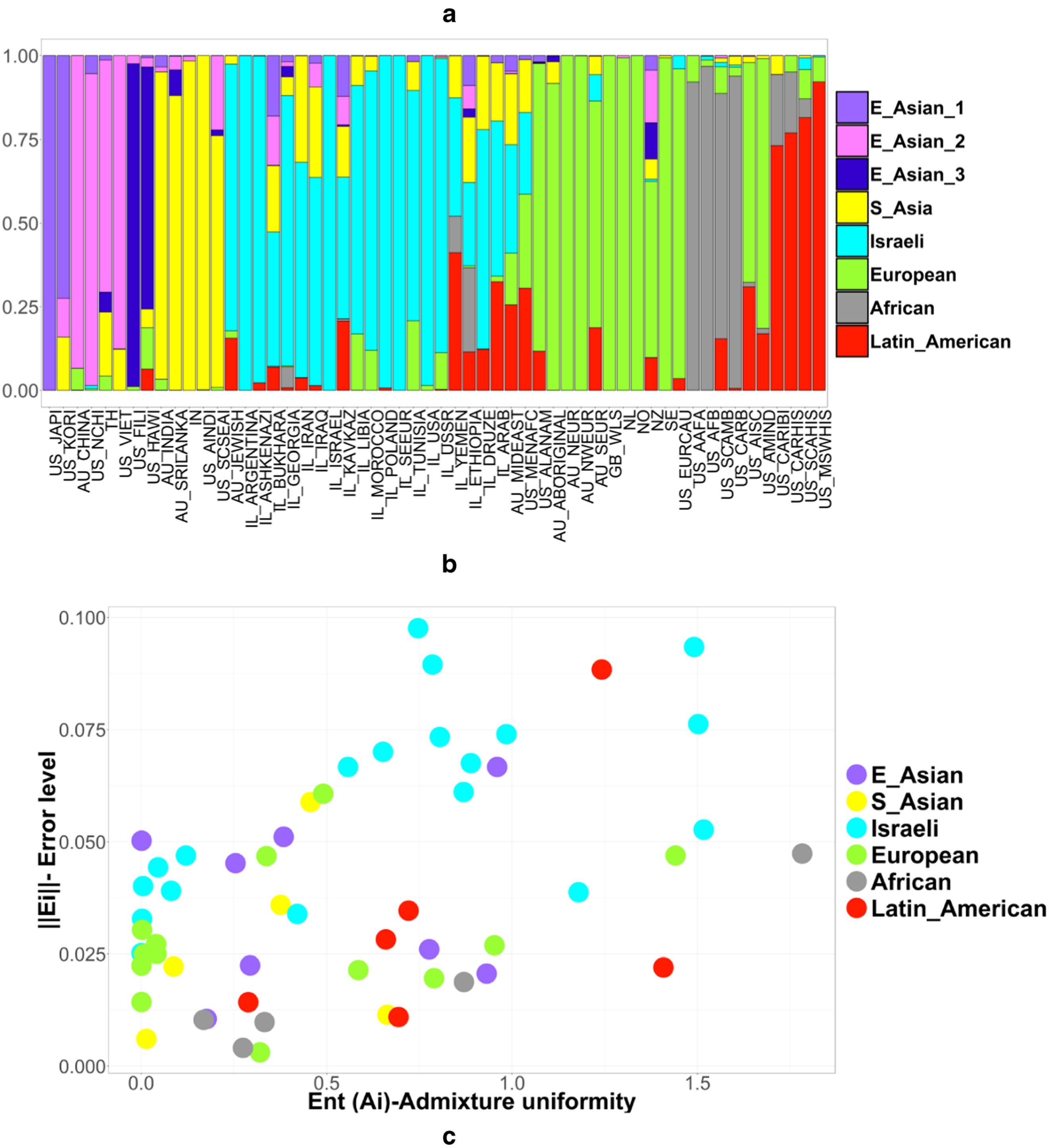
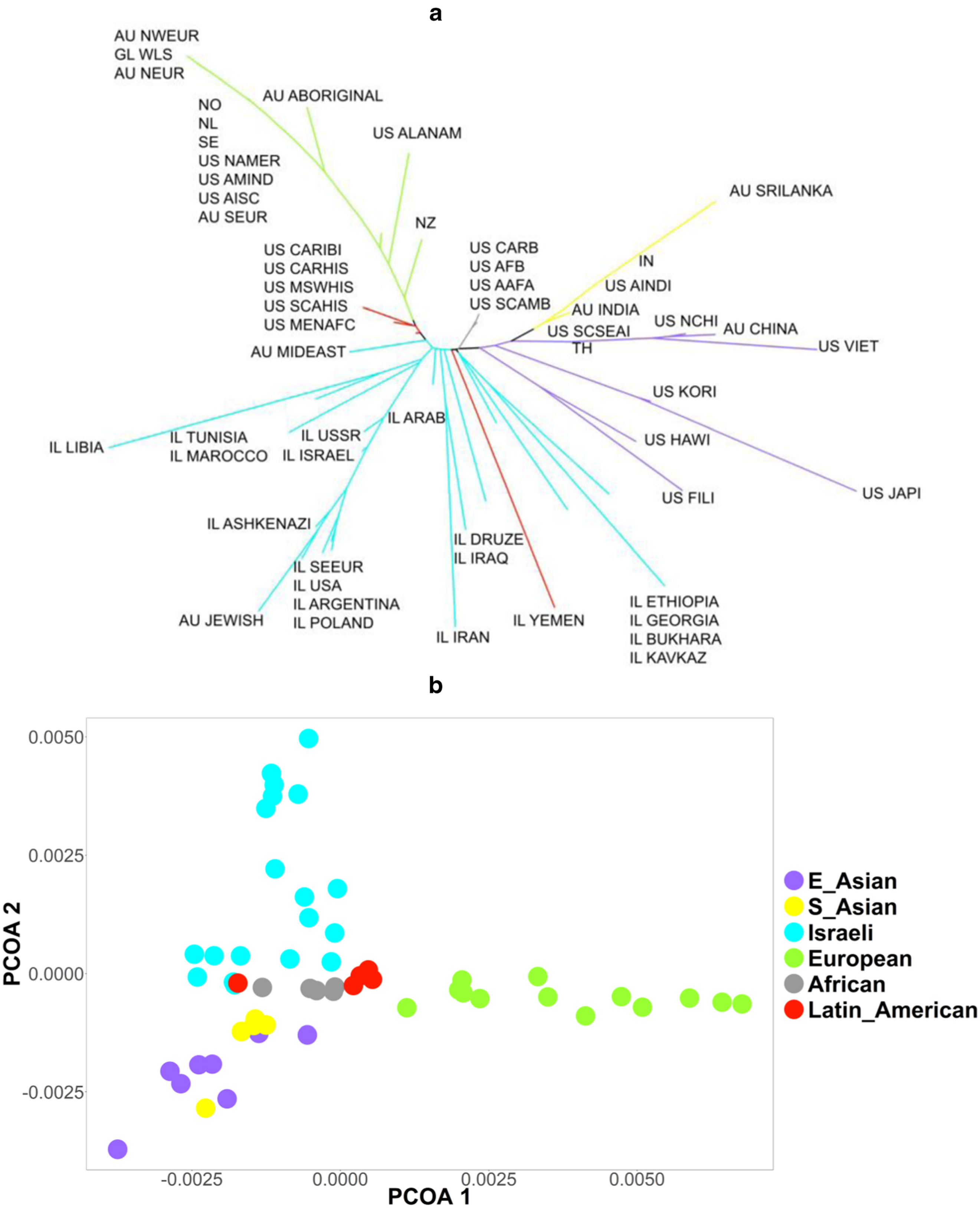


**Fig. 1** A) Relative decomposition error computed for each evaluated cost function as a function of the number of SCs (Source Components). In all tested cost functions, the errors plateaued at 7–10 SCs (Inset shows minimum Bayesian Information Criteria (BIC) and Akaike information criterion (AIC) reached for the same range of Ops indicating minimal information loss in this range). B) Population Admixture Analysis: individual admixture proportion of 56 population samples with 8 SCs. The color composition reflects the admixture (i.e., the fraction of a given population explained by the appropriate SC. Some of the population are composed of a single SC. All population acronyms are listed in Table 1. C) Scatter plot of admixture error $\|E_i\|$ vs. the admixture entropy$Ent(A_i)$. Each population has an assorted color, depending on the group it belongs to. The correlation between the entropy and admixture is unaffected by the population size

populations, while other populations, such as US Middle Eastern, New Zealand, and Australian Aboriginal were quite admixed (Fig. 1B). These results are consistent when the admixture method and number of SCs are varied



(Supp. Mat. Fig. S1). The difference in the composition can be observed in the entropy of vector $A_i$ vs.$Ent(A_i)$ (Ordinate axis in Fig. 1C).

The highly admixed populations with higher diversity in haplotype frequency may be challenging to properly classify, as can be seen from the correlation between the

◀ **Fig. 2** A) Unrooted neighbor-joining (NJ) phylogenetic trees based on the Fst matrix calculated from five-locus haplotype frequencies of the 56 studied populations. The tree is divided into branches delineating the main ethnic groups in the current analysis: Jewish, African, European, Hispanic, East and South Asians. Color coding shows concordance between the estimated admixture and tree branches. The only exceptions are highly admixed populations, such as Ethiopian Jews. B) PCOA (Principal Coordinate Analysis) analyses of the distances as defined by pairwise population Fst. The population has been grouped into broad regions: South Asian, East Asian, Pacific Islanders, Jews, European, African, and Hispanic. More plots are provided in supplementary materials

admixture error $\|E_i\|$and the admixture entropy$Ent(A_i)$ (Fig. 1C). The error results with average and std are given in Supp. Mat. Fig. S2.

Note that the observed correlation may be the result of a cofactor such as the registry size N. There is a negative correlation between N and Ent(A) since populations with high N values tend to have a single population explaining them to minimize the loss. For similar reasons there is a negative correlation between N and the error. To neutralize such an effect a partial correlation was performed (i.e., a correlation over the residual of the regression on the population size (Nunnally and Bernstein 1994), with a Spearman partial correlation of 0.4 ($p < 1.e{-}4$.). Interesting correlations also emerge between the entropy of the haplotype frequencies and the error, with a low error for high entropy (more uniform distribution of haplotypes). Moreover, as could be expected, size is correlated with a lower error and a lower entropy of A, representing the bias toward the more precise representation of larger populations (Supp Mat. Fig. S3). Note that we have used the sampled population as a proxy for the real population size. However, as discussed later, the results are quite robust to changes in the input, and thus are not expected to be affected by limited sampling bias.

The admixture results (Fig. 1B) are consistent with the phylogenetic analysis based on the *Fst* distance between the HLA haplotype frequencies of the observed populations (Fig. 2A), where leaves are colored per$PP(i)$. The tree splits into primary ethnic groups. Similar populations from different registries (e.g., populations of African ancestry from the Australian and US registries) are neighbors in the lineage tree and have similar $PP(i)$values, showing that technical differences between the methods used by different registries have a limited effect on the resulting admixture. The trees based on HLA frequencies effectively reproduce most details of current known population admixture, including the fine details of Asian and Jewish populations, as well as the admixture of Latin American populations (Gragert et al. 2013; Verdu et al. 2014) .

Some interesting trends in the tree, beyond reproducing the main human lineages, are the proximity between American Indians and European populations, which is previously reported (Wang et al. 2007; Bryc et al. 2015). This can be seen also by the significant European SC component of Native American populations (Fig. 1B). The same trend holds even more for Australian Aboriginal populations which agrees with the high admixture between these groups and settlers reported in the literature (McEvoy et al. 2010; Bryc et al. 2015; Malaspinas et al. 2016). Israeli Arabs and Druze are closer to their Jewish neighbors than to Middle Eastern or North African populations in the US or Australia. This is probably the result of the broad classification of these groups in the US and Australian registries (Gragert et al. 2013).

The classification of populations based on their leading SCs agrees with PCOA analyses of the pairwise population Fst distances based only on HLA. The PCOA is also in agreement with the geographical spread of the populations. The first PC (horizontal axis in Fig. 2A) represents an East-West division of populations, separating the European from all other populations. The second PC (vertical axis in Fig. 2A) separates Middle Eastern and Israeli groups from Asian populations. The African populations are in the center of the PCOA and the phylogenetic tree is as expected from the out of Africa evolutionary theories of human ancestry (Hammer et al. 1998; Moorjani et al. 2011). Note the agreement of $PP(i)$with the PCOA map, further showing that the admixture captures the main axes of variance in the human genetic development (Other dimensions of the PCOA are given in Supp. Mat. Fig. S4).

To test the robustness of our method, we performed multiple comparisons of the admixtures obtained when varying the populations or reducing the number of loci. Specifically, 4 different tests were performed (Fig. 3.). In all cases the resulting admixture was practically not affected. First, the admixture analysis was repeated after removing all HLA-C and HLA-DQB1 loci, i.e., on HLA-A~HLA-B~HLA~DRB1 haplotype (upper plot labeled “3 locus”). We then repeated the analysis with most Israeli populations removed leaving only 37 populations to analyze (second plot labeled “37 pop”). Finally, we replaced the Israeli populations with the Canadian populations. The Canadian populations only had four-locus haplotype frequencies. Thus, the haplotypes lack HLA-DQB1 so other populations frequencies were also reduced to a four-locus resolution accordingly. In all cases very similar admixtures were obtained (Fig. 3). Finally, the entire analysis was repeated 50 times, where a mixture of two random populations was added and the original two populations were removed. We then compared the admixture of the new population with the linear combination of the SCs. The Euclidean distances between the appropriate admixture vectors were much lower than the distances between pairs of random populations (Fig. 3 lowest panel.).

Two other factors that can affect the results are the sampling level, and the sensitivity to the initial conditions

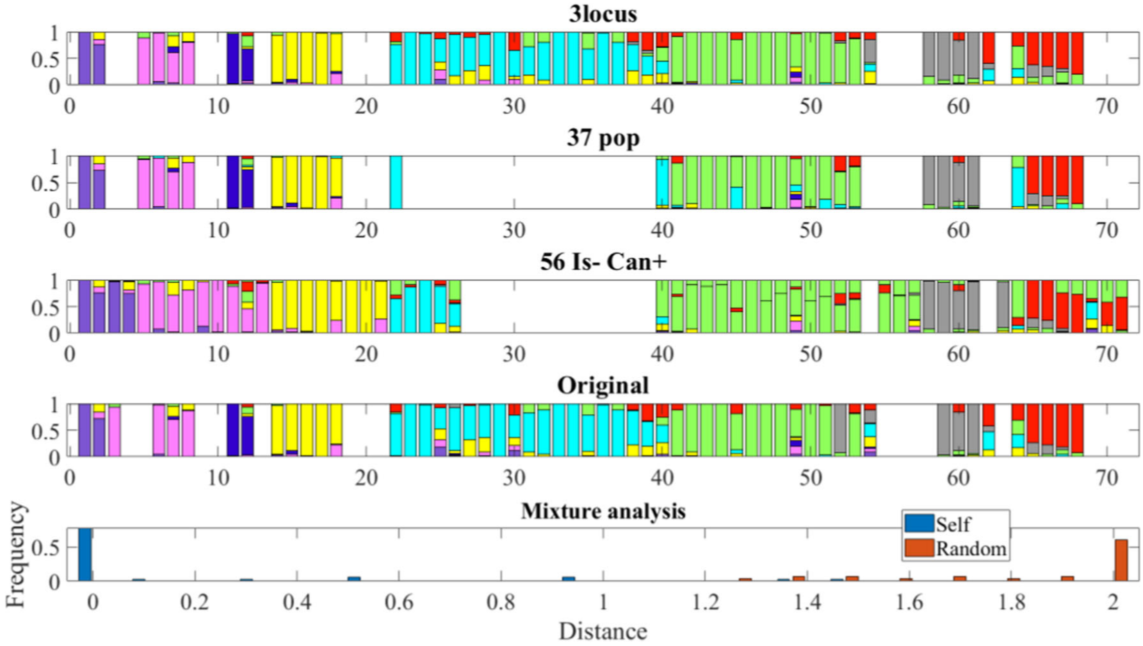


**Fig. 3** Validation of HLA based admixture. Four validations were performed. The upper 3 rows represent the results of the admixture analysis with alternative populations or alternative resolutions. The 4th row is the result of the original admixture analysis, and the last row is the distance distribution of simulated admixtures vs the admixture computed in the NMF. The first row is the NMF analysis using the same populations as the original analysis but with only 3 loci used for the analysis. The second row is the analysis with the original resolution, but without most Israeli populations. The third has most Israeli populations removed, and Canadian populations added. In addition, the resolution was lowered to 4 alleles. The fourth row is a repetition of Fig. 1. The last (fifth row) represent the result of simulations where a new population was created as a combination of two existing populations i and j. We then removed populations i and j, and computed the admixture with the mixed populations, we then computed the distance between the resulting admixture for the new population, and the combination of the original populations (blue bar for distribution), and the distance between the resulting population and another random population (red bars). One can clearly see that the blue distribution represents much shorter distances than the red one)

of the NMF. The NMF is a stochastic algorithm, and the two computed matrices are initiated with random positive values. The random seed may affect the results. To test that the results are not affected by the initial seed, we performed the optimization with different seeds and different numbers of SCs, with very similar results. (Fig. 4A). Similarly, while the population size is inherently incorporated in the optimization problem, the different sampling level may affect the results. To test for the sampling effect, we sampled each population to different levels (1000–25,000) with repetition (i.e., the same haplotype can be sampled multiple times). The results obtained from the sampled haplotype frequencies are very similar to the results from the full sample (Fig. 4B). The main difference observed between the different initial seeds, as well as the sampling is that in some solutions, the Filipino and Hawaian populations form a distinct SC, and in some solution it does not. The other populations are practically not affected.

To compare our results to existing methods, we performed an MCMC based admixture analysis on a subset of genome-wide ancestry informative marker SNPs in the 1000 genome (1KG) populations (Fig. 5A) and compared the result to a subset of geographically similar populations in our study where admixture was estimated using the NMF-based analysis and HLA haplotype frequencies. As seen in Fig. 5A, significant similarities can be seen between the NMF-based (top) and MCMC estimated admixtures (bottom), especially in Chinese, Japanese, African, and European populations. We observed more splits in the Latin-American 1KG populations (MXL, PUR, CLM) in the MCMC admixture. This could be the result of over representation of these samples in the 1KG project compared with the registry samples. Note that the registry data does not contain a specific Iberian Spanish population and therefore there is no dedicated haplotype frequency that particularly represents Iberian Spanish groups. Additionally, the European admixture in registry Latin American populations is usually less than in Iberian populations because it is partially replaced by Amerindian admixture from Native Americans. This resulted in Latin American SCs participating in the admixture of the Iberian Spanish 1000 Genomes population (IBS—Fig. 5A). This is by no means indicative of population migration trends (historically, migration happened from Spain to the Americas, but not necessarily in the other

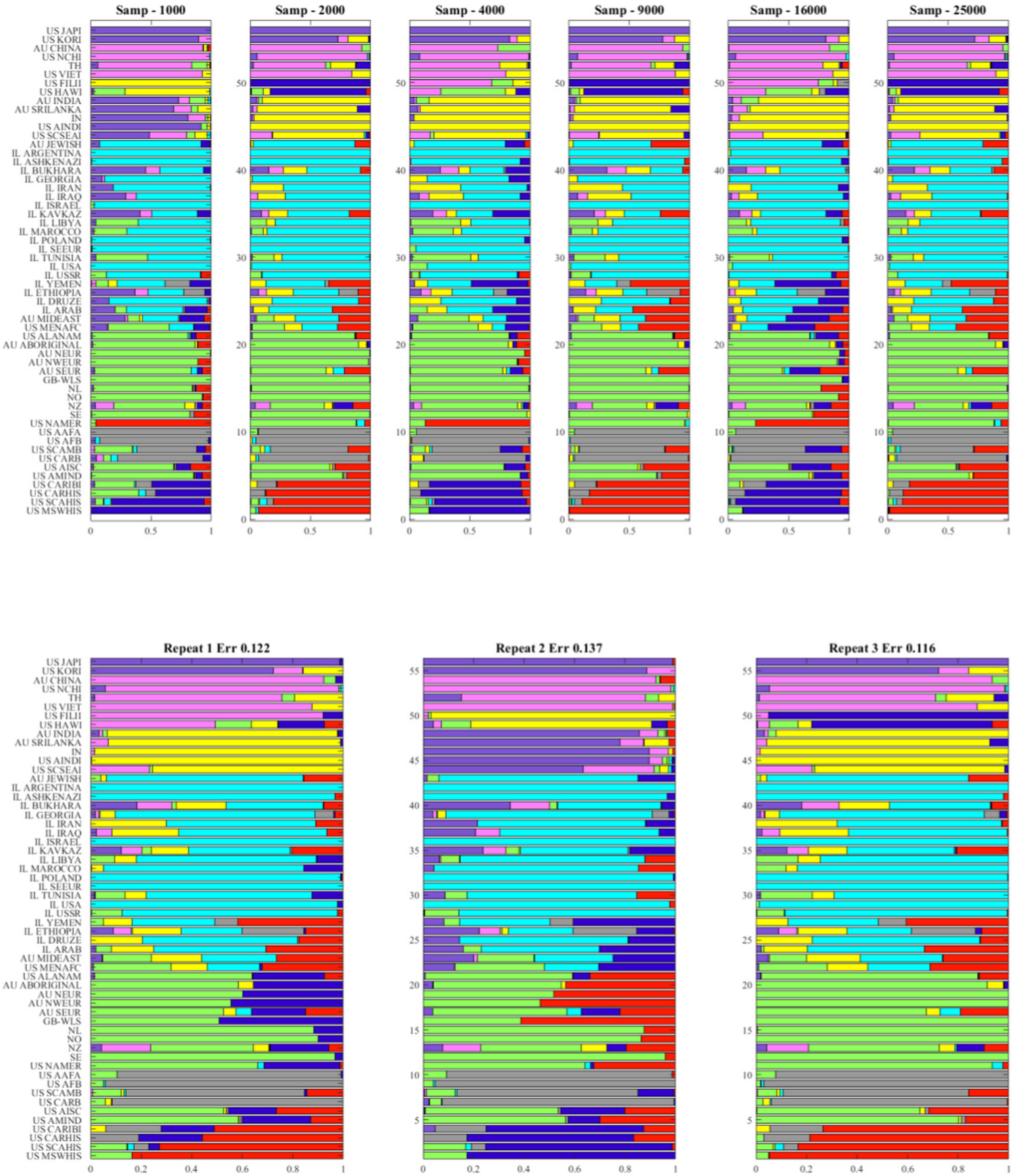
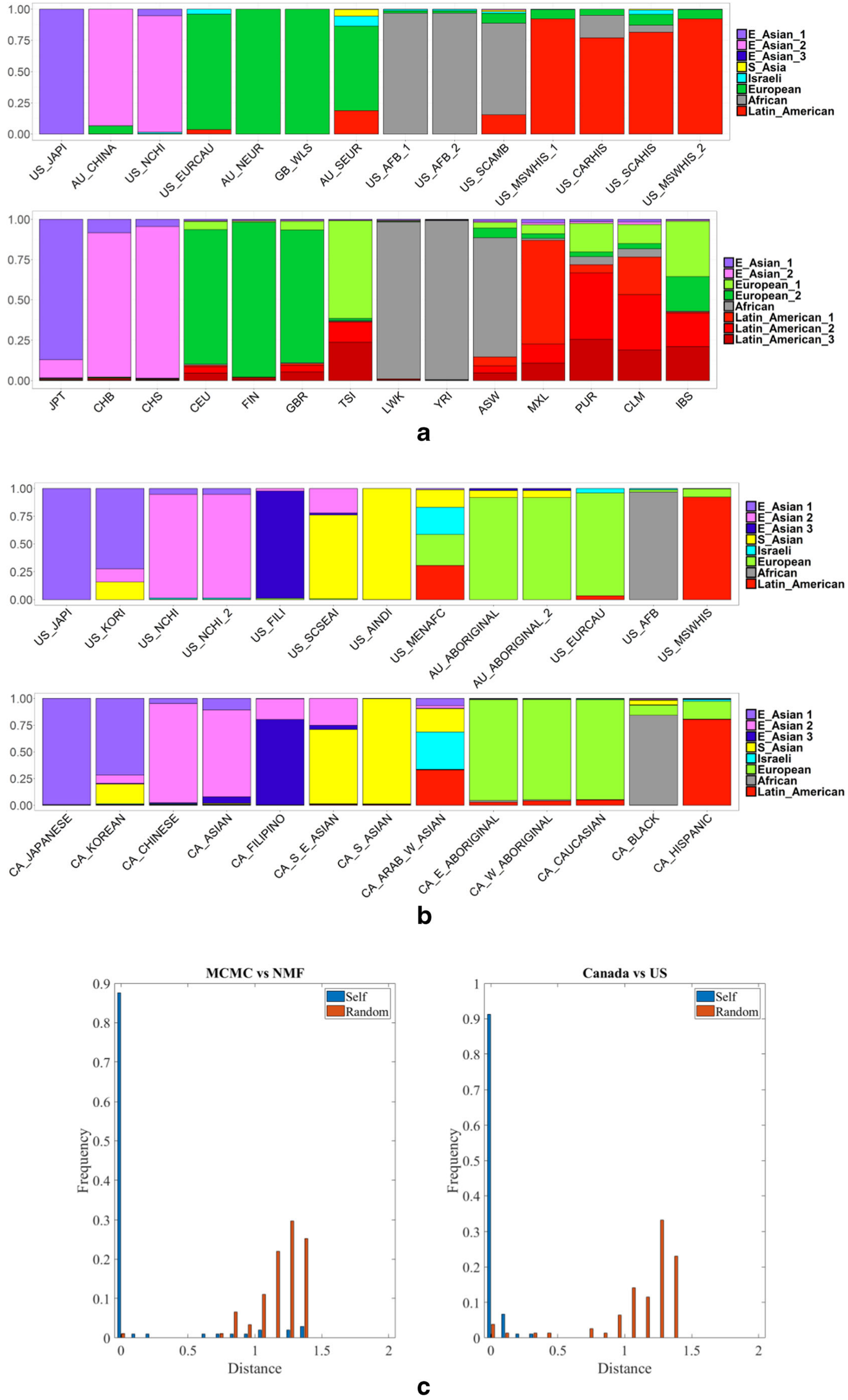


**Fig. 4** Effect of sampling and random seeds on the decomposition. We tested the effect of sampling with repetition (upper plots) and of using the full sample, but with different seeds (lower plots) on the results. The effect of sampling is mainly the disappearance in some samples of the Filipino and Hawaiian SC (yellow bays in the rows 7–8. And the parallel split in the Hispanic population. The same happens when suboptimal solutions are used for the full populations (lower plots—two left columns)

direction), but rather of the use of available HLA haplotype frequencies to break down the admixture of certain populations for clinical matching purposes. To explain further, a methodological distinction between classical


1.00
0.75
0.50
0.25
0.00
US_JAPI
AU_CHINA
US_NCHI
US_EURCAU
AU_NEUR
GB_WLS
AU_SEUR
US_AFB_1
US_AFB_2
US_SCAMB
US_MSWHIS_1
US_CARHIS
US_SCAHIS
US_MSWHIS_2
E_Asian_1
E_Asian_2
E_Asian_3
S_Asia
Israeli
European
African
Latin_American
JPT
CHB
CHS
CEU
FIN
GBR
TSI
LWK
YRI
ASW
MXL
PUR
CLM
IBS
E_Asian_1
E_Asian_2
European_1
European_2
African
Latin_American_1
Latin_American_2
Latin_American_3
a
US_JAPI
US_KORI
US_NCHI
US_NCHI_2
US_FILI
US_SCSEAI
US_AINDI
US_MENAFC
AU_ABORIGINAL
AU_ABORIGINAL_2
US_EURCAU
US_AFB
US_MSWHIS
E_Asian 1
E_Asian 2
E_Asian 3
S_Asian
Israeli
European
African
Latin_American
CA_JAPANESE
CA_KOREAN
CA_CHINESE
CA_ASIAN
CA_FILIPINO
CA_S_E_ASIAN
CA_S_ASIAN
CA_ARAB_W_ASIAN
CA_E_ABORIGINAL
CA_W_ABORIGINAL
CA_CAUCASIAN
CA_BLACK
CA_HISPANIC
b
MCMC vs NMF
Canada vs US
Self
Random
Frequency
Distance
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1
0.5
1
1.5
2
c

◀ **Fig. 5** A) Comparison between admixture on 1000 genome data from STRUCTURE analysis (lower plot) and from the current NMF (Non-negative Matrix Factorization) based analysis (upper plot). The colors represent the SCs as produced in STRUCTURE and the NMF. For each population in the 1 K Genome data, the most similar population in the registry was used. B) Validation bar plot of Canadian population admixture based on the SCs producing the admixture in Fig. 1B. All haplotype frequencies were reduced to four-locus resolution like genotypes provided from the Canadian OneMatch registry. Lowest panels represent the distance distribution between matching (blue) and non-matching (orange) pairs. One can clearly see that matching pairs are much closer than non-matching pairs for the two comparisons above

admixture methods and the current decomposition is the absence of predefined "original" admixture components in the current analysis. The components are computed from the observed distributions in all populations. Thus, a small population flowing into a large population would be described as composed from the large population and not vice-versa (e.g., the Iberian population affected by American Hispanic population and not the opposite). While this is an impediment for the analysis of migration, this is an advantage when the goal is to find the most probable population to find a match.

To test similarities between the genomic MCMC and HLA based NMF method, we computed the distance between the admixture of matching populations and non-matching populations. The difference can be clearly observed (Fig. 5 lowest panels ($T$ test $p < 1.\text{e}{-}5$)). Note that the difference between the methods is in both the data used (genomic vs HLA) and method (MCMC vs. NMF). Still the results are very similar.

One application of the presented methods is the detection of the composition of groups with unknown SIRE to understand their sub-structure. To estimate the composition of unknown populations, we performed an admixture analysis using the estimated SCs to compute admixture in groups from the Canadian OneMatch registry while blinding SIRE, without the limit that the entire composition should be explained by a combination of the SCs:

$$H_i = \sum_{j=1}^{K} a_{ij} P_j + \varepsilon_i, 0 \le a_{ij}. \tag{5}$$

HLA haplotype frequencies from the Canadian registry were estimated at a four-locus resolution. To estimate the admixture of the Canadian populations, we dropped the resolution of the SCs haplotype frequency distribution to four-locus by summing over all five-locus haplotypes that differ only in HLA-DQB1. We then computed the admixture of the new observed populations using Eq. (5).

For most Canadian populations, the estimated admixture via SCs agreed with self-reported race and ethnicities and with similar groups from other registries in the US, Australia, and Europe (Fig. 5B). The similarity between the US and Canadian admixtures is clearly demonstrated in the admixture estimation of the Canadian Asian, African, European, Latin American, Korean, and Middle Eastern populations. However, some populations, such as Aboriginal and Filipino Canadians show a slightly different composition. For example, the Canadian Aboriginal population is composed not only of the Australian Aboriginal population components, but also has fragments from African and Hispanic populations. This could be attributed to different patterns and times of population migration and historical events. A test similar to the MCMC comparison above (comparison between self and non-self Euclidean distance of the admixtures) was performed with similar results (Fig. 5 lowest panels ($T$ test $p < 1.\text{e}{-}5$)).

## Discussion

We present an algorithm that dissects population genetic admixture, based on HLA haplotype frequencies, into SCs in the presence of background LD and high polymorphism. While traditional admixture and lineage models typically rely on many bi-allelic genome-wide loci, we demonstrate that HLA is polymorphic enough to allow for a clear delineation of population composition using a single genomic region. The admixture problem, shown equivalent to a Non-negative Matrix Factorization analysis, accommodates the admixture of a large number of populations with an extensive number of haplotypes (in the order of a million haplotypes). To our knowledge, this is the first algorithm to dissect population admixture with specific focus on HLA and a validation framework using a dataset with the presented magnitude.

We developed and applied our method to the haplotype frequencies of 56 populations from different adult stem-cell donor registries representing over 3.5 million volunteer donors. We showed that the resulting admixture is consistent with the known ethnic composition, recent history, and SNP based admixture.

The results of the admixture and phylogenetic analyses show a clear distinction between Asian, African, European, and Israeli populations. Expectedly, the Latin-American and Middle Eastern Arab populations were more admixed than other populations and contained a notable European component. Our method was also able to distinguish East Asian from south Asian and Pacific Islander populations such as New Zealand. Some of the analyzed populations were more admixed than others; for example, most East Asian groups had a single predominant SC while Caribbean and Middle Eastern groups were more admixed. A distinct difference emerged between Israeli and non-Israeli populations. Within the Jewish Israeli populations, phylogenetic analysis separated several distinct groups: Ashkenazi, North African, Central Asian, and Yemenite and Ethiopian Jews. Interestingly, the Jewish populations were mainly admixed with Caucasian and south Asian populations, probably representing historic migration and modern admixture events. Our results showed over 50% European admixture in Native American and Australian aboriginal populations, suggesting large recent admixtures (Fig. 1B). Additionally, we have reported in a previous study a degree of over reporting of self-identified Native-American race among Be The Match donors (Hollenbach et al. 2015) that did not completely coincide with the genetic composition of reporting individuals, some of whom were found to have substantial European admixtures. Our phylogenetic analysis showed comparable results (Fig. 2A). The general division of branches agreed with previous phylogenies of general human populations

A limitation of the current methods is its restriction to phased HLA haplotype frequencies. The current implementation cannot be applied to SNP or other unphased marker data. Additionally, the presented admixture assignments are estimated at the population level, i.e., each haplotype is assigned a most likely SC, but a relative fraction of SC cannot be assigned to an individual. Thus, the presented method is not necessarily a replacement but rather complimentary to existing unphased admixture models. Additionally, a relatively isolated population with mostly private haplotypes (haplotypes found in only one population) and less representation in the frequencies might not be accurately represented by SCs. We suspect this might be partly the reason an Australian Aborigine component was not explicitly picked up, while the New Zealand component was. Australian Aboriginals might be relatively more isolated than the New Zealand population that could share some roots with Hawaiians and Southeast Asians and therefore is better represented in the frequencies.

Haplotype frequencies were estimated using an implementation of the Expectation Maximization (EM) algorithm that resolves phase, allelic, and missing allele ambiguities (Kollman et al. 2007) and is robust against various levels of HLA typing ambiguities, including missing loci (Kollman et al. 2007). This allows the use of intermediate and high-resolution data from patient-directed registry typing (i.e., from subjects who were selected for testing on behalf of a specific patient) to extend haplotype frequency estimates to the allele level for loci with higher rates of missing data like HLA-C and HLA-DQB1. Additionally, other researchers have shown the robustness of the EM algorithm results against deviations from Hardy-Weinberg equilibrium proportions in the presence of larger sample sizes (Single et al. 2002). These findings were reached using a study cohort of about 250 samples for frequency estimation, therefore we anticipate robustness of the estimated frequencies used in our study given that our population sample sizes ranged from 900 to over 1 million samples per population.

Beyond the theoretical importance of NMF based admixture methods, the presented results are useful in the context of hematopoietic stem cell transplantation (HSCT). Current methods to estimate haplotype frequencies and detect optimally matched HLA donors are based on Self-Identified Race and Ethnicity (SIRE), with an assumption of within-group mixing and limited admixture. This definition is quite arbitrary and has no clear genetic rationale. We believe SIRE can be a proxy for ancestry but generally captures incomplete population information.

The concept of SIRE is crucial for defining population haplotype frequencies that are core to the imputation process (Madbouly et al. 2014), used to resolve HLA typing ambiguities. Volunteer donor registries have accumulated donors over decades where HLA typing has significantly evolved. However, the donor search process has to accommodate the legacy ambiguous HLA types as well as the new less ambiguous ones. Additionally, registries are now becoming more international with the search extending to include worldwide donors. Here the concept of SIRE gets more complicated and is frequently inaccurate. The modelling presented in this work becomes essential as the fraction of people with unknown SIRE or mixed SIRE increases. The development of admixture methods in combination with a Bayesian classification scheme that would associate donors with different SCs would allow both imputation and matching of donors in the absence of SIRE. Moreover, when the number of different populations is increased, using SCs would limit the number of combinations to be tested for donors with complex ancestry. Importantly, our continuous validation of the matching process (outside the scope of this work) has shown that SIRE, when available, will always provide more accurate imputation and matching (data not shonw). Therefore, our modeling is intended to compensate for absence of SIRE but not necessarily to replace it all together.

**Acknowledgment** We would like to thank the following registries for allowing their data to be used for this study: National Marrow Donor Program/Be The Match, USA; Ezer Mizion Bone Marrow Donor Registry, Israel; OneMatch Stem Cell and Marrow Network, Canada; Australian Bone Marrow Donor Registry, Australia; Matchis: the Dutch Centre for Stem Cell Donors, The Netherlands; Norwegian Bone Marrow Donor Registry, Norway; New Zealand Bone Marrow Donor Registry, New Zealand; Tobias Registry of Swedish Bone Marrow Donors, Sweden; Thai National Stem Cell Donor Registry, Thailand; Welsh Bone Marrow Donor Registry, Wales, UK.

**Data availability** The authors affirm that all data necessary for confirming the conclusions of the article are present within the article, figures, and tables or in published datasets.

## References

Abi-Rached L, Jobin MJ, Kulkarni S, McWhinnie A, Dalva K, Gragert L, Babrzadeh F, Gharizadeh B, Luo M, Plummer FA (2011) The shaping of modern human immune systems by multiregional admixture with archaic humans. Science 334(6052):89–94

Aguilar A, Roemer G, Debenham S, Binns M, Garcelon D, Wayne RK (2004) High MHC diversity maintained by balancing selection in an otherwise genetically monomorphic mammal. Proc Natl Acad Sci U S A 101(10):3490–3494

Alexander DH, Novembre J, Lange K (2009) Fast model-based estimation of ancestry in unrelated individuals. Genome Res 19(9):1655–1664

Alter I, Gragert L, Fingerson S, Maiers M, Louzoun Y (2017) HLA class I haplotype diversity is consistent with selection for frequent existing haplotypes. PLoS Comput Biol 13(8):e1005693

Apanius V, Penn D, Slev PR, Ruff LR, Potts WK (1997) The nature of selection on the major histocompatibility complex. Crit Rev Immunol 17(2):179–224

Balding DJ (2006) A tutorial on statistical methods for population association studies. Nat Rev Genet 7(10):781–791

Behar DM, Yunusbayev B, Metspalu M, Metspalu E, Rosset S, Parik J, Rootsi S, Chaubey G, Kutuev I, Yudkovsky G (2010) The genome-wide structure of the Jewish people. Nature 466(7303):238–242

Bird CE, Karl SA, Smouse PE, Toonen RJ (2011) Detecting and measuring genetic differentiation. Phylogeography and population genetics in Crustacea 19:31–55

Brand A, Doxiadis I, Roelen D (2013) On the role of HLA antibodies in hematopoietic stem cell transplantation. HLA 81(1):1–11

Bryc K, Durand EY, Macpherson JM, Reich D, Mountain JL (2015) The genetic ancestry of African Americans, Latinos, and European Americans across the United States. Am J Hum Genet 96(1):37–53

Chinen J, Buckley RH (2010) Transplantation immunology: solid organ and bone marrow. J Allergy Clin Immunol 125(2):S324–S335

Chua EW, Kennedy MA (2012) Current state and future prospects of direct-to-consumer pharmacogenetics. Front Pharmacol 3:152

Consortium GP (2015) A global reference for human genetic variation. Nature 526(7571):68

Costa CL,Schneider DM, Ramos MF, de Aguiar MA (2017) “Constructing phylogenetic trees in individual based models.” arXiv preprint arXiv:1709.04416

Excoffier L, Laval G, Schneider S (2005) Arlequin (version 3.0): an integrated software package for population genetics data analysis. Evol Bioinformatics Online 1:47

Excoffier L, Smouse PE, Quattro JM (1992) Analysis of molecular variance inferred from metric distances among DNA haplotypes: application to human mitochondrial DNA restriction data. Genetics 131(2):479–491

Ferrell PB, McLeod HL (2008). “Carbamazepine, HLA-B* 1502 and risk of Stevens–Johnson syndrome and toxic epidermal necrolysis: US FDA recommendations.”

Fujimura JH, Rajagopalan R (2010) Different differences: the use of ‘genetic ancestry’ versus race in biomedical human genetic research. Soc Stud Sci. https://doi.org/10.1177/0306312710379170

Gaujoux R, Seoighe C (2010) A flexible R package for nonnegative matrix factorization. BMC bioinformatics 11(1):367

Geneugelijk K, Wissing J, Koppenaal D, Niemann M, Spierings E (2017) Computational approaches to facilitate epitope-based HLA matching in solid organ transplantation. J Immunol Res

Gragert L, Eapen M, Williams E, Freeman J, Spellman S, Baitty R, Hartzman R, Rizzo JD, Horowitz M, Confer D (2014a) HLA match likelihoods for hematopoietic stem-cell grafts in the US registry. N Engl J Med 371(4):339–348

Gragert L, Fingerson S, Albrecht M, Maiers M, Kalaycio M, Hill BT (2014b) Fine-mapping of HLA ASSOCIATIONS with chronic lymphocytic leukemia in US populations. Blood 124(17):2657–2665

Gragert L, Madbouly A, Freeman J, Maiers M (2013) Six-locus high resolution HLA haplotype frequencies derived from mixed-resolution DNA typing for the entire US donor registry. Hum Immunol 74(10):1313–1320

Hammer MF, Karafet T, Rasanayagam A, Wood ET, Altheide TK, Jenkins T, Griffiths RC, Templeton AR, Zegura SL (1998) Out of Africa and back again: nested cladistic analysis of human Y chromosome variation. Mol Biol Evol 15(4):427–441

Hindorff LA, Sethupathy P, Junkins HA, Ramos EM, Mehta JP, Collins FS, Manolio TA (2009) Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. Proc Natl Acad Sci U S A 106(23):9362–9367

Hollenbach JA, Saperstein A, Albrecht M, Vierra-Green C, Parham P, Norman PJ, Maiers M (2015) Race, ethnicity and ancestry in unrelated transplant matching for the National Marrow Donor Program: a comparison of multiple forms of self-identification with genetics. PloS one 10(8):e0135960

Holoshitz J (2013) The quest for better understanding of HLA-disease association: scenes from a road less travelled by. Discov Med 16(87):93–101

Kaeuffer R, Réale D, Coltman D, Pontier D (2007) Detecting population structure using STRUCTURE software: effect of background linkage disequilibrium. Heredity 99(4):374

Kennedy GC, Matsuzaki H, Dong S, Liu W-m, Huang J, Liu G, Su X, Cao M, Chen W, Zhang J (2003) Large-scale genotyping of complex DNA. Nat Biotechnol 21(10):1233–1237

Klitz W, Gragert L, Maiers M, Fernandez-Viña M, Ben-Naeh Y, Benedek G, Brautbar C, Israel S (2010) Genetic differentiation of Jewish populations. Tissue Antigens 76(6):442–458

Kollman C, Maiers M, Gragert L, Müller C, Setterholm M, Oudshoorn M, Hurley CK (2007) Estimation of HLA-A,-B,-DRB1 haplotype frequencies using mixed resolution data from a national registry with selective retyping of volunteers. Hum Immunol 68(12):950–958

Lam T, Shen M, Chia J, Chan S, Ren E (2013) Population-specific recombination sites within the human MHC region. Heredity 111(2):131

Lee DD, Seung HS (2001) Algorithms for non-negative matrix factorization. Adv Neural Inf Proces Syst

Lobkovsky AE,Levi L,Wolf YI, Maiers M, Gragert L, Alter I, Louzoun Y, Koonin EV (2019) “Multiplicative fitness, rapid haplotype discovery, and fitness decay explain evolution of human MHC”. Proceedings of the National Academy of Sciences: 201714436

Madbouly A, Gragert L, Freeman J, Leahy N, Gourraud PA, Hollenbach JA, Kamoun M, Fernandez-Vina M, Maiers M (2014) Validation of statistical imputation of allele-level multilocus phased genotypes from ambiguous HLA assignments. Tissue antigens 84(3):285–92

Maiers M, Halagan M, Joshi S, Ballal HS, Jagannatthan L, Damodar S, Srinivasan P, Narayan S, Khattry N, Malhotra P (2014) HLA match likelihoods for Indian patients seeking unrelated donor transplantation grafts: a population-based study. The Lancet Haematol 1(2): e57–e63

Malaspinas A-S, Westaway MC, Muller C, Sousa VC, Lao O, Alves I, Bergström A, Athanasiadis G, Cheng JY, Crawford JE (2016) A genomic history of Aboriginal Australia. Nature 538(7624):207–214

Manor S, Halagan M, Shriki N, Yaniv I, Zisser B, Maiers M, Madbouly A, Stein J (2016) High-resolution HLA A ~ B ~ DRB1 haplotype frequencies from the Ezer Mizion Bone Marrow Donor Registry in Israel. Hum Immunol 77(12):1114–1119

McEvoy BP, Lind JM, Wang ET, Moyzis RK, Visscher PM, van Holst Pellekaan SM, Wilton AN (2010) Whole-genome genetic diversity in a sample of Australians with deep Aboriginal ancestry. Am J Hum Genet 87(2):297–305

Moorjani P, Patterson N, Hirschhorn JN, Keinan A, Hao L, Atzmon G, Burns E, Ostrer H, Price AL, Reich D (2011) The history of African gene flow into Southern Europeans, Levantines, and Jews. PLoS Genet 7(4):e1001373

Nunnally JC, Bernstein I (1994) Psychometric Theory (McGraw-Hill Series in Psychology). McGraw-Hill, New York

Parham P, Ohta T (1996) Population biology of antigen presentation by MHC class I molecules. Science 272(5258):67–74

Paschou P, Drineas P, Lewis J, Nievergelt CM, Nickerson DA, Smith JD, Ridker PM, Chasman DI, Krauss RM, Ziv E (2008) Tracing substructure in the European American population with PCA-informative markers. PLoS Genet 4(7):e1000114

Phillips BL, Callaghan C (2017) The immunology of organ transplantation. Surgery (Oxford)

Pickrell JK, Coop G, Novembre J, Kudaravalli S, Li JZ, Absher D, Srinivasan BS, Barsh GS, Myers RM, Feldman MW, Pritchard JK (2009) Signals of recent positive selection in a worldwide sample of human populations. Genome Res 19(5):826–837

Porras-Hurtado L, Ruiz Y, Santos C, Phillips C, Carracedo Á, Lareu MV (2013) An overview of STRUCTURE: applications, parameter settings, and supporting software. Front Genet 4

Pritchard JK, Stephens M, Donnelly P (2000) Inference of population structure using multilocus genotype data. Genetics 155(2):945–959

Robinson J, Halliwell JA, Hayhurst JD, Flicek P, Parham P, Marsh SG (2014) The IPD and IMGT/HLA database: allele variant databases. Nucleic Acids Res 43(D1):D423–D431

Sabeti PC, Varilly P, Fry B, Lohmueller J, Hostetter E, Cotsapas C, Xie X, Byrne EH, McCarroll SA, Gaudet R, Schaffner SF, Lander ES, C. International HapMap, Frazer KA, Ballinger DG, Cox DR, Hinds DA, Stuve LL, Gibbs RA, Belmont JW, Boudreau A, Hardenbol P, Leal SM, Pasternak S, Wheeler DA, Willis TD, Yu F, Yang H, Zeng C, Gao Y, Hu H, Hu W, Li C, Lin W, Liu S, Pan H, Tang X, Wang J, Wang W, Yu J, Zhang B, Zhang Q, Zhao H, Zhao H, Zhou J, Gabriel SB, Barry R, Blumenstiel B, Camargo A, Defelice M, Faggart M, Goyette M, Gupta S, Moore J, Nguyen H, Onofrio RC, Parkin M, Roy J, Stahl E, Winchester E, Ziaugra L, Altshuler D, Shen Y, Yao Z, Huang W, Chu X, He Y, Jin L, Liu Y, Shen Y, Sun W, Wang H, Wang Y, Wang Y, Xiong X, Xu L, Waye MM, Tsui SK, Xue H, Wong JT, Galver LM, Fan JB, Gunderson K, Murray SS, Oliphant AR, Chee MS, Montpetit A, Chagnon F, Ferretti V, Leboeuf M, Olivier JF, Phillips MS, Roumy S, Sallee C, Verner A, Hudson TJ, Kwok PY, Cai D, Koboldt DC, Miller RD, Pawlikowska L, Taillon-Miller P, Xiao M, Tsui LC, Mak W, Song YQ, Tam PK, Nakamura Y, Kawaguchi T, Kitamoto T, Morizono T, Nagashima A, Ohnishi Y, Sekine A, Tanaka T, Tsunoda T, Deloukas P, Bird CP, Delgado M, Dermitzakis ET, Gwilliam R, Hunt S, Morrison J, Powell D, Stranger BE, Whittaker P, Bentley DR, Daly MJ, de Bakker PI, Barrett J, Chretien YR, Maller J, McCarroll S, Patterson N, Pe'er I, Price A, Purcell S, Richter DJ, Sabeti P, Saxena R, Schaffner SF, Sham PC, Varilly P, Altshuler D, Stein LD, Krishnan L, Smith AV, Tello-Ruiz MK, Thorisson GA, Chakravarti A, Chen PE, Cutler DJ, Kashuk CS, Lin S, Abecasis GR, Guan W, Li Y, Munro HM, Qin ZS, Thomas DJ, McVean G, Auton A, Bottolo L, Cardin N, Eyheramendy S, Freeman C, Marchini J, Myers S, Spencer C, Stephens M, Donnelly P, Cardon LR, Clarke G, Evans DM, Morris AP, Weir BS, Tsunoda T, Johnson TA, Mullikin JC, Sherry ST, Feolo M, Skol A, Zhang H, Zeng C, Zhao H, Matsuda I, Fukushima Y, Macer DR, Suda E, Rotimi CN, Adebamowo CA, Ajayi I, Aniagwu T, Marshall PA, Nkwodimmah C, Royal CD, Leppert MF, Dixon M, Peiffer A, Qiu R, Kent A, Kato K, Niikawa N, Adewole IF, Knoppers BM, Foster MW, Clayton EW, Watkin J, Gibbs RA, Belmont JW, Muzny D, Nazareth L, Sodergren E, Weinstock GM, Wheeler DA, Yakub I, Gabriel SB, Onofrio RC, Richter DJ, Ziaugra L, Birren BW, Daly MJ, Altshuler D, Wilson RK, Fulton LL, Rogers J, Burton J, Carter NP, Clee CM, Griffiths M, Jones MC, McLay K, Plumb RW, Ross MT, Sims SK, Willey DL, Chen Z, Han H, Kang L, Godbout M, Wallenburg JC, L'Archeveque P, Bellemare G, Saeki K, Wang H, An D, Fu H, Li Q, Wang Z, Wang R, Holden AL, Brooks LD, McEwen JE, Guyer MS, Wang VO, Peterson JL, Shi M, Spiegel J, Sung LM, Zacharia LF, Collins FS, Kennedy K, Jamieson R, Stewart J (2007) Genome-wide detection and characterization of positive selection in human populations. Nature 449(7164):913–918

Saitou N, Nei M (1987) The neighbor-joining method: a new method for reconstructing phylogenetic trees. Mol Biol Evol 4(4):406–425

Sanchez-Mazas A, Thorsby E (2012) HLA in anthropology: the enigma of Easter Island. Clin Transpl:167–173

Shiina T, Hosomichi K, Inoko H, Kulski JK (2009) The HLA genomic loci map: expression, interaction, diversity and disease. J Hum Genet 54(1):15–39

Simmonds M, Gough S (2007) The HLA region and autoimmune disease: associations and mechanisms of action. Curr Genomics 8(7):453–465

Single RM, Meyer D, Hollenbach JA, Nelson MP, Noble JA, Erlich HA, Thomson G (2002) Haplotype frequency estimation in patient populations: the effect of departures from Hardy-Weinberg proportions and collapsing over a locus in the HLA region. Genetic Epidemiology: The Official Publication of the International Genetic Epidemiology Society 22(2):186–195

Sjakste T, Kalnina J, Paramonova N, Nikitina-Zake L, Sjakste N (2016) "Journal of Molecular and Genetic Medicine."

Slater N, Louzoun Y, Gragert L, Maiers M, Chatterjee A, Albrecht M (2015) Power laws for heavy-tailed distributions: modeling allele and haplotype diversity for the national marrow donor program. PLoS Comput Biol 11(4):e1004204

Tokunaga K,Imanishi T, Takahashi K, Juji T (1996) "On the origin and dispersal of East Asian populations as viewed from HLA haplotypes". Prehistoric mongoloid dispersals: 187-197.

Traherne J (2008) Human MHC architecture and evolution: implications for disease association studies. Int J Immunogenet 35(3):179–192

Verdu P, Pemberton TJ, Laurent R, Kemp BM, Gonzalez-Oliver A, Gorodezky C, Hughes CE, Shattuck MR, Petzelt B, Mitchell J (2014) Patterns of admixture and population structure in native populations of Northwest North America. PLoS Genet 10(8):e1004530

Voight BF, Kudaravalli S, Wen X, Pritchard JK (2006) A map of recent positive selection in the human genome. PLoS Biol 4(3):e72

Wang S, Lewis CM Jr, Jakobsson M, Ramachandran S, Ray N, Bedoya G, Rojas W, Parra MV, Molina JA, Gallo C (2007) Genetic variation and population structure in Native Americans. PLoS Genet 3(11):e185

Weir BS, Cockerham CC (1984) Estimating F-statistics for the analysis of population structure. evolution:1358–1370

Weir BS, Hill WG (2002) Estimating F-statistics. Annu Rev Genet 36(1): 721–750

Weir BS, Ott J (1997) Genetic data analysis II. Trends Genet 13(9):379

Williamson SH, Hubisz MJ, Clark AG, Payseur BA, Bustamante CD, Nielsen R (2007) Localizing recent adaptive evolution in the human genome. PLoS Genet 3(6):e90

Zhou H, Alexander D, Lange K (2011) A quasi-Newton acceleration for high-dimensional optimization algorithms. Stat Comput 21(2):261–273

**Publisher's note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

## Affiliations

**Alexandra Litinsky Simanovsky**[1] · **Abeer Madbouly**[2] · **Michael Halagan**[2] · **Martin Maiers**[2] · **Yoram Louzoun**[1]

[1] Department of Mathematics and Gonda brain research institute, Bar-Ilan University, 52900 Ramat-Gan, Israel

[2] Bioinformatics Research, Center for International Blood and Marrow Transplant Research, Minneapolis, MN, USA