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

Running title: Single haplotype based admixture.

Alexandra Litinsky Simanovsky^{1†}, Abeer Madbouly^{2†}, Michael Halagan², Martin Maier², Yoram Louzoun^{1,*}

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

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

[†]These authors have contributed equally to this work

*Corresponding author: Yoram Louzoun, Department of Mathematics and Gonda brain research institute, Bar-Ilan University, Ramat-Gan 52900, Israel. Email – louzouy@math.biu.ac.il

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 masks the ancestral haplotype frequency distribution.

We present an HLA-based 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 populations. The result is a minimal set of original populations decoding roughly 90% of the total variance in the studied admixtures. These original populations agree with the geographical distribution, phylogenies and recent admixture events of the studied groups.

With the growing population of multi-ethnic individuals, the 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 original groups, 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

Author Summary

Human Leukocyte Antigen (HLA) is known to be the most polymorphic region in the human genome. 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.

We developed a new HLA-based method to identify admixture models in mixed populations using high-resolution HLA haplotype frequencies. Our results highlight that a single highly polymorphic locus can contain enough information to map clearly human admixture and the population genetics of the different human populations, and reproduces results based on SNP arrays.

The presented algorithm is validated using haplotype frequencies sampled from 56 worldwide populations. Under such factorization we demonstrate that 90% of the variance in these populations can be explained using a much-reduced set of 8 ethnic groups. We demonstrate that the estimated ethnic groups and admixture models agree with the geographical distribution, population phylogenies and recent historic admixture events of the studied populations.

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) [1-4]. As generations of humans migrated throughout the world, HLA has evolved with each population conserving key ancestral features [5], and with more than 15,000 defined alleles to date [6] and over 1,000,000 haplotypes [7], 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 [8][9], planning and modeling volunteer adult donor stem-cell registries [10] and personalized and consumer genetics [11]. 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 [12-18] or study the population genetics of specific ethnic groups [17, 19, 20]. The need for large genomic regions stems from the limited polymorphism in bi-allelic SNPs. This, however, is not the case for HLA. 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 [21, 22]. Multiple models have been proposed to explain HLA polymorphism [23, 24], the most prominent being pathogen-driven balancing selection [25]. 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 [26] [27].

Importantly, because of the high-gene density, diversity and low recombination rates, HLA alleles are inherited as haplotypes, preserving a high degree of linkage disequilibrium (LD) [28]. Recent evidence shows a clear purifying selection in HLA haplotypes [29], i.e. many combinations of HLA alleles

are seldom separated over multiple generations. 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[21], creating founder haplotypes with insufficient evolutionary time to degrade. 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 [30, 31] [32].

Multiple researchers have studied population genetic traits through the estimation and analysis of HLA allele and haplotype frequencies. Klitz et al [33] studied detailed Expectation Maximization based on three-locus low resolution haplotype frequencies to estimate the phylogeny of Jewish populations using shared haplotypes and correlations between haplotype frequencies among populations. More detailed correlation based analysis of high resolution six-locus haplotype frequencies was performed by Gragert et al [34] 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 selection affecting HLA haplotypes [29, 35]. However, to our knowledge there is no published work dichotomizing HLA haplotypes into original populations 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 [36] [37]. This is formally a hidden state model, where the hidden states and the probability distribution of each hidden state in each population are estimated. To resolve the admixture into its primary components, most algorithms use a maximum likelihood approach [37, 38]. Multiple methods have been developed to detect population admixture (e.g. ADMIXTURE [37] and STRUCTURE[36, 39]). 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 polymorphic markers, such as HLA and microsatellites [40]. 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 original populations (OP) 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 original populations can be used to detect the admixture composition of new populations.

Materials and Methods

Study Populations

HLA haplotype frequencies from 56 global populations were used in this study (table 1-a). 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 [41] multiple registries in India[42] , the US Be The Match registry® [34], 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 1-b[34]). 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 HLA-A~C~B~DRB1 frequencies were available.

Nonnegative Matrix Factorization

To estimate population admixture proportions using Nonnegative Matrix Factorization (NMF), a non-negative matrix $\mathbf{C}$ would be factorized into two non-negative matrices $\mathbf{A}$ and $\mathbf{B}$, where $\mathbf{C}$ is a $n \times m$ matrix, representing the haplotype frequencies in n populations, each with m haplotypes. $\mathbf{A}$ is $n \times r$ non-negative mixing matrix and $\mathbf{B}$ is $r \times m$ non-negative “original” populations (OP) matrix. Each n -dimensional vector represents one out of m populations. In the current application, m constitutes the 56 studied populations and n is of the order of a million different haplotypes. The variable r is chosen to be much smaller than n and m . The range of r values, representing the count of possible OPs, was tested between 1 and 23. The results of the decomposition are a function of the chosen cost function [43]. 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 [44].

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, a new non-negative matrix A or B is calculated. Then, the new obtained matrix is used for calculating the complementary matrix. Iterations are repeated until converging to a locally optimal matrix factorization [43].

There are currently multiple standard algorithms for NMF [43]. 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 produce a haplotype frequency matrix of all HLA haplotype frequencies for over 50 populations and present it as the product of eight original population frequency matrix and mixing factors.

Phylogenetic analysis

We applied a Neighbor-Joining (NJ) algorithm to construct a phylogenetic tree of the studied populations [45, 46]. Following the tree construction, populations were grouped by the most dominant OP obtained from the NMF analysis. To validate the NMF algorithm, we investigated 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 F_{st} measure [47, 48] calculated for all combinations of population pairs. Typically, the F_{st} value ranges from 0.0 to 1.0, where 0.0 indicates identical population frequencies and 1.0 indicates significantly different populations. We use the F_{st} algorithm as implemented by Weir et al [49-52].

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.

Hidden state model algorithm

We compared the performance of our NMF-based admixture algorithm to the software STRUCTURE[36] 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 SNP genotypes from the reference 1000 Genomes phase I dataset [53] and compared the output admixtures to admixtures estimated by OPs in populations of similar ancestry as the 1000 Genome populations in our study. The 500 selected genome-wide SNP ancestry informative marker panel was previously published and validated for delineation of population substructure [54]. STRUCTURE v2.3.4 was run using K = 8 clusters (like the count of OPs), with 7,000 replicates and 50,000 burnin cycles.

Inference of unlabeled populations

A possible application of the work presented here is to dissect an unknown admixed population into a set of OPs. Starting with a set of estimated OP 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.

Model

We present a model to infer the admixture 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)[54, 55]. To estimate the components of a population admixture, we assume the existence of K OPs, each represented by a haplotype frequency distribution denoted as P_j , where j ranges from 1 to K OPs. Each observed population is assumed to be a positive normalized combination of such OPs:

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

where ε_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:

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

where A is the matrix of all a_{ij} . To estimate the number of OPs 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-population will 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:

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

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 OPs (Fig. 1A). Note that non-scaled solutions produce OPs strongly biased toward highly admixed groups (Supp. Mat. Fig. S1) and produce a higher error rates (Fig. 1A).

Each OP 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 OPs 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 population one main group according to the main OP composing them. In the following results, the observed populations are grouped according to the OP composing the largest fraction of the admixture vector A_i (A single column of A) (Fig. 1B).

$$(4) PP(i) = \arg \max_k (a_{ik}).$$

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

Results

Our results show that eight OPs are sufficient to represent the admixture in the 56 studied populations. We observed that the diversity level varied widely among populations; in some groups the admixture was predominantly represented by one OP as in the case of the US Japanese and Filipino 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 OPs are varied (Supp. Mat. Fig. S1). The difference in the composition can be observed in the entropy of vector $A_i - Ent(A_i)$ (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 admixture error $\|E_i\|$ and the admixture entropy $Ent(A_i)$ (Fig. 1C).

Note that the observed correlation may be the result of a cofactor such as the population size. To neutralize such an effect a partial correlation was performed (i.e. a correlation over the residual of the regression on the population size), 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. S2).

The admixture results (Fig. 1B) are consistent with the phylogenetic analysis based on the F_{st} 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 alone 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[56, 57].

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[58, 59]. This can be seen also by the significant European OP 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[59-61]. 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[56].

The classification of populations based on their leading OP agrees with PCOA analyses of the pairwise population F_{st} 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 as expected from out of Africa evolutionary theories of human ancestry[62, 63]. 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. S2).

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.

3A) 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. 3A, 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 the European admixture in registry Latin American populations is usually less than Iberian populations because of the Amerindian influence.

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

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

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 OPs 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 OPs agreed with self-reported race and ethnicities and with similar groups from other registries in the US, Australia and Europe (Figure 3B). 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 of population migration and historical events.

Discussion

We present an algorithm that dissects population genetic admixture, based on HLA haplotype frequencies, into OP components in the presence of background LD and high polymorphism. While traditional admixture and lineage models typically rely on many biallelic 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, is more computationally efficient than traditional admixture models and allows for 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 volunteer stem-cell registries representing over 3.5 million 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 populations had a single predominant OP 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 (Figure 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 [64] 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 (Figure 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 OP, but a relative fraction of OP cannot be assigned to an individual. Thus, the presented method is complimentary to existing unphased admixture models.

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 endogamy within the same SIRE. The probability to find an optimally HLA matched stem-cell donor is higher within the same ethnic group of a searching patient [10]. However, both admixture and phylogenetic results presented here show that some SIRE groups are composed of multiple endogamous populations and some have similarities in haplotype frequency distribution. The presented algorithm can therefore help dissect these multiple SIRE groups into a multitude of sub-populations based on defined haplotype distributions, which would ultimately streamline the matching process for HSCT.

314

315 **Acknowledgement**

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

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

325 This work was supported by a grant from the Department of the Navy, Office of Naval Research
326 (N00014-16-1-2020).

327

328 **Bibliography**

- 329 1. Geneugelijk K, Wissing J, Koppelaar D, Niemann M, Spierings E. Computational approaches to
330 facilitate epitope-based HLA matching in solid organ transplantation. *Journal of immunology research*.
331 2017;2017.
- 332 2. Chinen J, Buckley RH. Transplantation immunology: solid organ and bone marrow. *Journal of*
333 *Allergy and Clinical Immunology*. 2010;125(2):S324-S35.
- 334 3. Phillips BL, Callaghan C. The immunology of organ transplantation. *Surgery (Oxford)*. 2017.
- 335 4. Brand A, Doxiadis I, Roelen D. On the role of HLA antibodies in hematopoietic stem cell
336 transplantation. *HLA*. 2013;81(1):1-11.
- 337 5. Sanchez-Mazas A, Thorsby E. HLA in anthropology: the enigma of Easter Island. *Clinical*
338 *transplants*. 2012;167-73.
- 339 6. Robinson J, Halliwell JA, Hayhurst JD, Flicek P, Parham P, Marsh SG. The IPD and IMGT/HLA
340 database: allele variant databases. *Nucleic Acids Research*. 2014;43(D1):D423-D31.
- 341 7. Slater N, Louzoun Y, Gragert L, Maier M, Chatterjee A, Albrecht M. Power Laws for Heavy-Tailed
342 Distributions: Modeling Allele and Haplotype Diversity for the National Marrow Donor Program. *PLOS*
343 *Comput Biol*. 2015;11(4):e1004204.
- 344 8. Gragert L, Fingersh S, Albrecht M, Maier M, Kalaycio M, Hill BT. Fine-mapping of HLA
345 associations with chronic lymphocytic leukemia in US populations. *Blood*. 2014;124(17):2657-65.
- 346 9. Ferrell PB, McLeod HL. Carbamazepine, HLA-B* 1502 and risk of Stevens–Johnson syndrome and
347 toxic epidermal necrolysis: US FDA recommendations. 2008.
- 348 10. Gragert L, Eapen M, Williams E, Freeman J, Spellman S, Baitty R, et al. HLA match likelihoods for
349 hematopoietic stem-cell grafts in the US registry. *New England Journal of Medicine*. 2014;371(4):339-48.
- 350 11. Chua EW, Kennedy MA. Current state and future prospects of direct-to-consumer
351 pharmacogenetics. *Frontiers in pharmacology*. 2012;3.
- 352 12. Williamson SH, Hubisz MJ, Clark AG, Payseur BA, Bustamante CD, Nielsen R. Localizing recent
353 adaptive evolution in the human genome. *PLoS genetics*. 2007;3(6):e90. doi:
354 10.1371/journal.pgen.0030090. PubMed PMID: 17542651; PubMed Central PMCID: PMC1885279.

13. Voight BF, Kudaravalli S, Wen X, Pritchard JK. A map of recent positive selection in the human genome. *PLoS biology*. 2006;4(3):e72. doi: 10.1371/journal.pbio.0040072. PubMed PMID: 16494531; PubMed Central PMCID: PMC1382018.
14. Sabeti PC, Varilly P, Fry B, Lohmueller J, Hostetter E, Cotsapas C, et al. Genome-wide detection and characterization of positive selection in human populations. *Nature*. 2007;449(7164):913-8. doi: 10.1038/nature06250. PubMed PMID: 17943131; PubMed Central PMCID: PMC2687721.
15. Pickrell JK, Coop G, Novembre J, Kudaravalli S, Li JZ, Absher D, et al. Signals of recent positive selection in a worldwide sample of human populations. *Genome research*. 2009;19(5):826-37. doi: 10.1101/gr.087577.108. PubMed PMID: 19307593; PubMed Central PMCID: PMC2675971.
16. Hindorff LA, Sethupathy P, Junkins HA, Ramos EM, Mehta JP, Collins FS, et al. Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. *Proceedings of the National Academy of Sciences of the United States of America*. 2009;106(23):9362-7. doi: 10.1073/pnas.0903103106. PubMed PMID: 19474294; PubMed Central PMCID: PMC2687147.
17. Ginodi I, Vider-Shalit T, Tsaban L, Louzoun Y. Precise score for the prediction of peptides cleaved by the proteasome. *Bioinformatics*. 2008;24(4):477-83.
18. Balding DJ. A tutorial on statistical methods for population association studies. *Nature reviews Genetics*. 2006;7(10):781-91. doi: 10.1038/nrg1916. PubMed PMID: 16983374.
19. Kennedy GC, Matsuzaki H, Dong S, Liu W-m, Huang J, Liu G, et al. Large-scale genotyping of complex DNA. *Nature biotechnology*. 2003;21(10):1233-7.
20. Fujimura JH, Rajagopalan R. Different differences: The use of 'genetic ancestry' versus race in biomedical human genetic research. *Social Studies of Science*. 2010;0306312710379170.
21. Traherne J. Human MHC architecture and evolution: implications for disease association studies. *International journal of immunogenetics*. 2008;35(3):179-92.
22. Shiina T, Hosomichi K, Inoko H, Kulski JK. The HLA genomic loci map: expression, interaction, diversity and disease. *Journal of human genetics*. 2009;54(1):15-39.
23. Parham P, Ohta T. Population biology of antigen presentation by MHC class I molecules. *Science*. 1996;272(5258):67-74. PubMed PMID: 8600539.
24. Aguilar A, Roemer G, Debenham S, Binns M, Garcelon D, Wayne RK. High MHC diversity maintained by balancing selection in an otherwise genetically monomorphic mammal. *Proceedings of the National Academy of Sciences of the United States of America*. 2004;101(10):3490-4. doi: 10.1073/pnas.0306582101. PubMed PMID: 14990802; PubMed Central PMCID: PMC373489.
25. Apanius V, Penn D, Slev PR, Ruff LR, Potts WK. The nature of selection on the major histocompatibility complex. *Critical reviews in immunology*. 1997;17(2):179-224. PubMed PMID: 9094452.
26. Tokunaga K, Imanishi T, Takahashi K, Juji T. On the origin and dispersal of East Asian populations as viewed from HLA haplotypes. *Prehistoric mongoloid dispersals*. 1996:187-97.
27. Abi-Rached L, Jobin MJ, Kulkarni S, McWhinnie A, Dalva K, Gragert L, et al. The shaping of modern human immune systems by multiregional admixture with archaic humans. *Science*. 2011;334(6052):89-94.
28. Lam T, Shen M, Chia J, Chan S, Ren E. Population-specific recombination sites within the human MHC region. *Heredity*. 2013;111(2):131.
29. Alter I, Gragert L, Fingerson S, Maiers M, Louzoun Y. HLA class I haplotype diversity is consistent with selection for frequent existing haplotypes. *PLOS Computational Biology*. 2017;13(8):e1005693.
30. Simmonds M, Gough S. The HLA region and autoimmune disease: associations and mechanisms of action. *Current genomics*. 2007;8(7):453-65.
31. Holoshitz J. The quest for better understanding of HLA-disease association: scenes from a road less travelled by. *Discovery medicine*. 2013;16(87):93.

32. Sjakste T, Kalnina J, Paramonova N, Nikitina-Zake L, Sjakste N. *Journal of Molecular and Genetic Medicine*. 2016.
33. Klitz W, Gragert L, Maiers M, Fernandez-Viña M, Ben-Naeh Y, Benedek G, et al. Genetic differentiation of Jewish populations. *Tissue antigens*. 2010;76(6):442-58.
34. Gragert L, Madbouly A, Freeman J, Maiers M. Six-locus high resolution HLA haplotype frequencies derived from mixed-resolution DNA typing for the entire US donor registry. *Human immunology*. 2013;74(10):1313-20. doi: 10.1016/j.humimm.2013.06.025. PubMed PMID: 23806270.
35. Slater N, Louzoun Y, Gragert L, Maiers M, Chatterjee A, Albrecht M. Power laws for heavy-tailed distributions: modeling allele and haplotype diversity for the national marrow donor program. *PLoS computational biology*. 2015;11(4):e1004204.
36. Pritchard JK, Stephens M, Donnelly P. Inference of population structure using multilocus genotype data. *Genetics*. 2000;155(2):945-59.
37. Alexander DH, Novembre J, Lange K. Fast model-based estimation of ancestry in unrelated individuals. *Genome research*. 2009;19(9):1655-64.
38. Zhou H, Alexander D, Lange K. A quasi-Newton acceleration for high-dimensional optimization algorithms. *Statistics and computing*. 2011;21(2):261-73.
39. Porras-Hurtado L, Ruiz Y, Santos C, Phillips C, Carracedo Á, Lareu MV. An overview of STRUCTURE: applications, parameter settings, and supporting software. *Frontiers in genetics*. 2013;4.
40. Kaeuffer R, Réale D, Coltman D, Pontier D. Detecting population structure using STRUCTURE software: effect of background linkage disequilibrium. *Heredity*. 2007;99(4):374.
41. Manor S, Halagan M, Shriki N, Yaniv I, Zisser B, Maiers M, et al. High-resolution HLA A~ B~ DRB1 haplotype frequencies from the Ezer Mizion Bone Marrow Donor Registry in Israel. *Human immunology*. 2016;77(12):1114-9.
42. Maiers M, Halagan M, Joshi S, Ballal HS, Jagannathan L, Damodar S, et al. HLA match likelihoods for Indian patients seeking unrelated donor transplantation grafts: a population-based study. *The Lancet Haematology*. 2014;1(2):e57-e63.
43. Lee DD, Seung HS, editors. Algorithms for non-negative matrix factorization. *Advances in neural information processing systems*; 2001.
44. Gaujoux R, Seoighe C. A flexible R package for nonnegative matrix factorization. *BMC bioinformatics*. 2010;11(1):367.
45. Costa CL, Schneider DM, Ramos MF, de Aguiar MA. Constructing phylogenetic trees in individual based models. *arXiv preprint arXiv:170904416*. 2017.
46. Saitou N, Nei M. The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Molecular biology and evolution*. 1987;4(4):406-25.
47. Excoffier L, Laval G, Schneider S. Arlequin (version 3.0): an integrated software package for population genetics data analysis. *Evolutionary bioinformatics online*. 2005;1:47.
48. Weir BS, Hill WG. Estimating F-statistics. *Annual review of genetics*. 2002;36(1):721-50.
49. Weir BS, Cockerham CC. Estimating F-statistics for the analysis of population structure. *evolution*. 1984;1358-70.
50. Weir BS, Ott J. Genetic data analysis II. *Trends in Genetics*. 1997;13(9):379.
51. Excoffier L, Smouse PE, Quattro JM. Analysis of molecular variance inferred from metric distances among DNA haplotypes: application to human mitochondrial DNA restriction data. *Genetics*. 1992;131(2):479-91.
52. Bird CE, Karl SA, Smouse PE, Toonen RJ. Detecting and measuring genetic differentiation. *Phylogeography and population genetics in Crustacea*. 2011;19:31-55.
53. Consortium GP. A global reference for human genetic variation. *Nature*. 2015;526(7571):68.

54. Paschou P, Drineas P, Lewis J, Nievergelt CM, Nickerson DA, Smith JD, et al. Tracing sub-structure in the European American population with PCA-informative markers. *PLoS genetics*. 2008;4(7):e1000114.
55. Behar DM, Yunusbayev B, Metspalu M, Metspalu E, Rosset S, Parik J, et al. The genome-wide structure of the Jewish people. *Nature*. 2010;466(7303):238-42.
56. Gragert L, Madbouly A, Freeman J, Maiers M. Six-locus high resolution HLA haplotype frequencies derived from mixed-resolution DNA typing for the entire US donor registry. *Human immunology*. 2013;74(10):1313-20.
57. Verdu P, Pemberton TJ, Laurent R, Kemp BM, Gonzalez-Oliver A, Gorodezky C, et al. Patterns of admixture and population structure in native populations of Northwest North America. *PLoS genetics*. 2014;10(8):e1004530.
58. Wang S, Lewis Jr CM, Jakobsson M, Ramachandran S, Ray N, Bedoya G, et al. Genetic variation and population structure in Native Americans. *PLoS genetics*. 2007;3(11):e185.
59. Bryc K, Durand EY, Macpherson JM, Reich D, Mountain JL. The genetic ancestry of African Americans, Latinos, and European Americans across the United States. *The American Journal of Human Genetics*. 2015;96(1):37-53.
60. McEvoy BP, Lind JM, Wang ET, Moyzis RK, Visscher PM, van Holst Pellekaan SM, et al. Whole-genome genetic diversity in a sample of Australians with deep Aboriginal ancestry. *The American Journal of Human Genetics*. 2010;87(2):297-305.
61. Malaspina A-S, Westaway MC, Muller C, Sousa VC, Lao O, Alves I, et al. A genomic history of Aboriginal Australia. *Nature*. 2016;538(7624):207-14.
62. Hammer MF, Karafet T, Rasanayagam A, Wood ET, Altheide TK, Jenkins T, et al. Out of Africa and back again: nested cladistic analysis of human Y chromosome variation. *Molecular biology and evolution*. 1998;15(4):427-41.
63. Moorjani P, Patterson N, Hirschhorn JN, Keinan A, Hao L, Atzmon G, et al. The history of African gene flow into Southern Europeans, Levantines, and Jews. *PLoS genetics*. 2011;7(4):e1001373.
64. Hollenbach JA, Saperstein A, Albrecht M, Vierra-Green C, Parham P, Norman PJ, et al. 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*. 2015;10(8):e0135960.

479 Table 1. A) Study populations, acronyms used and the sample size used to generate haplotype
480 frequencies for each population. B) Populations from the Canadian OneMatch registry used for
481 validation. C) Phase I 1000 Genome population acronym details

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_MENAF	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

482 (A)

Population	Abbreviation	Count
Aboriginal~East	CA_E_ABORIGINAL	1,030
Aboriginal~West	CA_W_ABORIGINAL	2,099
ArabWestAsian	CA_ARAB_W_ASIAN	1,997
Asian	CA_ASIANC	8,091
Black	CA_BLACK	2,542
Caucasian	CA_CAUCASIAN	268,990
Chinese	CA_CHINESE	19,425
Filipino	CA_FILIPINO	1,026
Hispanic	CA_HISPANIC	2,384
Japanese	CA_JAPANESE	190
Korean	CA_KOREAN	489
SouthAsian	CA_S_ASIAN	13,076
SoutheastAsian	CA_S_E_ASIAN	2,312

483 (B)

484

485

486

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

487
488
489
490
491
492
493
494
495
496

(C)

Figure 1: A) Relative decomposition error computed for each evaluated cost function as a function of the number of OPs (Original Population). In all tested cost functions, the errors plateaued at 7-10 OPs. B) Population Admixture Analysis: individual ancestry proportion of 56 population samples with 8 OPs. The color composition reflects the admixture (i.e. the fraction of a given population explained by the appropriate OP. Some of the population are composed of a single OP. 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.

(A)

(C)

Figure 2: A) Unrooted neighbor-joining (NJ) phylogenetic trees based on the F_{st} 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 F_{st} . The population have been grouped into broad regions: South Asian, East Asian, Pacific Islanders, Jews, European, African and Hispanic. More plots are provided in supplementary materials.

(A)

(B)

Figure 3. A) Comparison between admixture on 1,000 genome data from STRCUTURE analysis (lower plot) and from the current NMF (Non-negative Matrix Factorization) based analysis (upper plot). The colors represent the OPs as produced in STRCUTURE and the NMF. For each population in the 1K Genome data, the most similar population in the registry was used. B) Validation bar plot of Canadian population admixture based on the OPs (Original Populations) producing the admixture in Figure 1B. All haplotype frequencies were reduced to four-locus resolution like genotypes provided from the Canadian OneMatch registry.