Human Immunology 74 (2013) 1313–1320

Contents lists available at SciVerse ScienceDirect

journal homepage: www.elsevier.com/locate/humimm

# Six-locus high resolution HLA haplotype frequencies derived from mixed-resolution DNA typing for the entire US donor registry

Loren Gragert *, Abeer Madbouly, John Freeman, Martin Maiers

*Bioinformatics Research, National Marrow Donor Program, Minneapolis, MN, USA*

ARTICLE INFO

*Article history:*
Received 30 October 2012
Accepted 14 June 2013
Available online 24 June 2013

ABSTRACT

We have calculated six-locus high resolution HLA A∼C∼B∼DRB3/4/5∼DRB1∼DQB1 haplotype frequencies using all Be The Match® Registry volunteer donors typed by DNA methods at recruitment. Mixed resolution HLA typing data was inputted to a modified expectation–maximization (EM) algorithm in the form of genotype lists generated by interpretation of primary genomic typing data to the IMGT/HLA v3.4.0 allele list. The full cohort consists of 6.59 million subjects categorized at a broad race level. Overall 25.8% of the individuals were typed at the C locus, and 5.2% typed at the DQB1 locus, while all individuals were typed for A, B, DRB1. We also present a subset of 2.90 million subjects with detailed race/ethnic information mapped to 21 population subgroups, 64.1% of which have primary DNA typing data across at least A, B, and DRB1 loci. Sample sizes at the detailed race level range from 1,242,890 for European Caucasian to 1,376 Alaskan Native or Aleut. Genetic distance measurements show high levels of HLA genetic divergence among the 21 detailed race categories, especially among the eight Asian–American populations. These haplotype frequencies will be used to improve match predictions for donor selection algorithms for hematopoietic stem cell transplantation and improve the accuracy in modeling registry match rates.



## 1. Introduction

The National Marrow Donor Program (NMDP) manages a registry of volunteer donors, the Be The Match Registry®, to facilitate unrelated hematopoietic stem cell transplantation (HSCT), a curative therapy for blood malignancies and other disorders. Donors are selected based on matching alleles with the recipient for several human leukocyte antigen (HLA) genes. Transplants that are HLA-matched have the best outcome because they prevent immune rejection of foreign tissue and facilitate immune reconstitution [1]. HLA genes are highly polymorphic, located in the major histocompatibility complex (MHC) on chromosome 6, with frequency of alleles and linkage of alleles into haplotypes varying widely among human populations.

Identification of alleles in registry HLA typing produces a typing result which relies on DNA-based assays that are not always able to precisely identify the alleles present (i.e. allelic ambiguity). Typing methodology has evolved over time, with earlier low resolution methods such as serology resulting in thousands of potential genotypes, while newer technology has reduced this ambiguity significantly. These mixed resolution HLA assignments have been a challenge to efforts at characterizing HLA haplotypic diversity from registry data.

The expectation–maximization (EM) algorithm takes in HLA genotypes as input to estimate population haplotype frequencies. Early implementations of the EM algorithm resolved only phase ambiguity, or linkage of alleles along a chromosome [2,3], but could not handle HLA assignments with allelic ambiguity, where some alleles are not distinguished from one another. We have previously presented high resolution HLA A∼C∼B∼DRB1∼DQB1 haplotype frequency data from four US population categories (Caucasian, African American, Asian or Pacific Islander, and Hispanic) [4]. The study population was limited by including only individuals typed without allelic ambiguity, and used broad race/ethnic categorization that did not distinguish among genetically distinct subpopulations. While the minority populations were typed in randomized prospective studies, the Caucasian HLA typings were performed on behalf of patient searches, which may have biased results towards HLA alleles commonly found in searching patients.

Kollman et al. has since implemented an EM algorithm that simultaneously resolves both phase and allelic ambiguity seen in mixed-resolution assignments [5]. Kollman estimated A∼B∼DRB1 frequencies with high resolution typings calculated only for the DRB1 locus, because of computational limitations. High resolution

* Corresponding author. Address: National Marrow Donor Program, 3001 Broadway St NE, Suite 100, Minneapolis, MN 55413, USA.
*E-mail address:* lgragert@nmdp.org (L. Gragert).

0198-8859 
http://dx.doi.org/10.1016/j.humimm.2013.06.025

estimates incorporating more loci using HLA assignments with allelic ambiguity becomes exponentially more difficult [6].

Significant limitations of previous frequency studies remain in terms of sample size, the number of HLA loci, and coverage of diverse world populations. These limitations manifest themselves in the performance of NMDP's matching algorithm, HapLogic™, which uses haplotype frequencies to predict the likelihood of allele-level matches between patient and donor. Since the initial release of HapLogic™, matched donors are now more rapidly identified; however, further improvements could be realized with utilization of the full complement of registry HLA data.

Here we describe a haplotype frequency estimation method that can process millions of mixed resolution typed samples and addresses previous limitations. We calculated high resolution haplotype frequencies at six loci (HLA-A∼C∼B∼DRB3/4/5∼DRB1∼DQB1) in 21 populations, including all DNA typed donors in the registry. Because HLA frequencies can differ substantially between subpopulations, this expansion of population categories improved the accuracy of allele predictions in matching algorithms. Match likelihood estimates are also improved as frequency generating population sample size is increased, because the fraction of multilocus genotypes that cannot be explained by any pair of high resolution haplotypes is decreased.

## 2. Materials and methods

### 2.1. HLA typing methods and primary data interpretation

The NMDP collects donor HLA typing from laboratories including primary data which contain extensive details on the exact tests performed and the presence and absence of specific oligonucleotide sequences. Alleles are described with the first two fields of HLA allele nomenclature, representing protein level assignment. We combine alleles with amino acids identical in the antigen recognition site (ARS) since these sets of alleles (listed in Supplementary Table 1) are often not distinguished by current typing systems, and genomic regions are not defined outside the ARS for many of the alleles [7].

For sequence specific oligonucleotide (SSO) and sequence specific primer (SSP) methods, primary data consists of a list of the probes used in the kit and their sequences, and the positive or negative result for each probe. For sequence based typing (SBT) methods, primary data includes the annealing location of the amplification primers, the diploid sequence read and any hemizygous reads resulting from group specific PCR or group specific sequencing primers (GSSP). Primary data are transmitted electronically to NMDP in an XML-based message format for interpretation [8].

Given the primary HLA typing data and the list of all described alleles for the IMGT/HLA database version 3.4.0, we calculated a genotype list for each locus of all possible allelic combinations that were consistent with the typing result for each subject for input into EM [9]. Primary typing data have been reported only by laboratories that have typed newly recruited donors since 1997. Of the remaining HLA typings, those that were reported as unambiguous alleles or in the NMDP allele code format were converted to genotype list format [9]. The typings reported in allele code format only include alleles that existed when the typing was performed, and do not consider newly described alleles that would have been consistent with the result obtained.

We were unable to test for Hardy–Weinberg equilibrium (HWE) on these samples because there are no methods to test for HWE using ambiguous HLA typing data. We also could not calculate HWE at the allele family level because many HLA typings contain possible genotypes that include more than two allele families.

### 2.2. Ambiguity reduction and haplotype frequency estimation

Haplotype frequencies were calculated from genotype list data using the expectation–maximization (EM) algorithm [2,3,5,10]. Some HLA typings have extremely high ambiguity, with as many as $10^{22}$ possible six-locus haplotype pairs in the genotype list. Because of the computational challenges inherent in calculating haplotype frequencies from these long genotype lists, we applied some methods to reduce ambiguity of the genotype list input to EM.

We first calculated a minimum set of alleles that explain all HLA typings in a population. Because many vanishingly rare alleles have been described in IMGT/HLA [11,12], and many ambiguous HLA typings contain possible alleles have never been reported unambiguously to NMDP for a given population, we removed these unlikely alleles from consideration. Our algorithm started with a set of common alleles (frequency> = 1/2000) for each broad race category [4] and attempted to interpret all primary data. During each iteration, we calculated which allele(s) allowed the most previously uninterpretable HLA typings to be assigned, then all genotype lists in the population sample were reinterpreted with the addition of the new allele(s). Only haplotype pairs containing alleles in the abridged allele list were included in the new genotype lists. Alleles removed from consideration are very unlikely to exist in the population because they are not required to interpret any of the HLA typings in the large population sample.

After reducing the allele list, the HLA typing data were still far too ambiguous to compute six-locus haplotype frequencies in a single EM run, so we broke up the calculations into tractable sub-problems. We ran EM on two-locus blocks beginning with C∼B and DRB3/4/5∼DRB1 loci where linkage disequilibrium was highest [13]. Using these haplotype frequencies, each subject in the sample is imputed to get a list of their possible haplotype pairs and probabilities up to a threshold of 99% cumulative probability. Next these reduced C∼B and DRB3/4/5∼DRB1 genotype lists were treated as a single locus, or block, for the next EM step where the A∼(C∼B) and (DRB3/4/5∼DRB1)∼DQB1 frequencies are calculated. The final two-locus EM step combined the class I and class II blocks to calculate six-locus haplotype frequencies (A∼(C∼B))∼((DRB3/4/5∼DRB1)∼DQB1). This blocks/imputation approach significantly reduced the number of possible haplotype pairs as haplotype blocks were extended, so only unlikely haplotype pairs were removed from consideration Using 16 3.0 GHz Intel Xeon CPUs, haplotype frequencies for the entire Be The Match registry, consisting of 6.59 million donors, was completed within one week.

Copy-number variation in HLA for DRB3/4/5 loci presents difficulties for HLA haplotype analysis. DRB1 can be found on the same chromosome as either DRB3, DRB4, DRB5, or none of the DRB3/4/5 genes [14]. Typings at these loci may have ambiguity for implicit possible heterozygosity, where it is unknown if a subject is homozygous for the DRB3/4/5 allele, or if they are heterozygous and lack a DRB3/4/5 gene. In the NMDP registry, we roughly estimate that 25% of donors are typed for the DRB3 and DRB5 loci, and 5% for the DRB4 locus; these estimates are inexact because reporting of typing for the DRB3/4/5 genes is not routinely carried out by HLA laboratories. DRB4 was inconsistently typed in our sample, with some labs typing only for DRB1, DRB3 and DRB5, resulting in complex heterogeneity in both typing methods and data. To generate haplotype frequencies, the expectation–maximization (EM) algorithm was modified to account for the structural and allelic ambiguity of DRB3/4/5 loci. In the EM algorithm we treated DRB3/4/5 as a single locus since a maximum of one of these genes occurs per chromosome.

Because no DRB3/4/5 typing intent was available, we developed a novel method for practical calculation of haplotype frequencies

that include DRB3/4/5. We restricted DRB3/4/5∼DRB1 genotype lists to only include DRB3/4/5 genes in associations that agreed with the common DRB3/4/5∼DRB1 linkages listed in Supplementary Table 2 [14]. The small fraction (<0.1%) of donors with HLA typings that could not possibly align with the common linkage rules were kept in with the fully enumerated genotype list, as exceptions to these linkage rules have been described, for example Tautz et al. [15].

### 2.3. Evaluation of genetic variation and population categories

We chose 21 population categories for analysis with the goal of reflecting the genetic diversity of the populations in the USA registry and to have adequate sample size to characterize the HLA genetics of each category. Over the history of NMDP donor recruitment, three different race/ethnicity questionnaires were used, with 49 race categories and three ethnic categories (Hispanic, Not Hispanic, and Ethnicity Not Asked) that totaled 147 race/ethnic combinations. To reduce this number of combinations to categories that were distinct and statistically relevant, several techniques based on Nei's genetic distance [16], sample size, and population definitions were applied. Subjects who self described in a way that was too general or not widely applicable were included in the broad categories (European American, African American, Asian or Pacific Islander, Hispanic, and Native American). Because donors from international donor centers are also listed in the US registry, international donors were included in the broad categories. Individuals who indicated multiple race/ethnic categories were put in a single Multiple Race category in the broad study, and were not assigned to any of the detailed race categories.

We calculated pairwise Nei's genetic distance on each race/ethnic combination using population haplotype frequencies. To visualize the genetic distance between all population samples, we created a population dendrogram using the nearest neighbor algorithm in PHYLIP [17]. Populations that were highly similar to each other genetically were grouped together, for example "North American White" and "Western European" were grouped into "European Caucasian".

Another method we applied to visualize population variation was principal component analysis (PCA), which summarizes frequencies differences among thousands individual haplotypes into a smaller number of dimensions. PCA was performed on the entire haplotype frequency distribution of each population using MATLAB R2011b [18]. 2-D plots were created using the principal component pairs to display major trends of haplotype variation among populations.

### 2.4. Sample populations

Table 1 lists the sample sizes (number of individuals) for the 21 populations used for the detailed race analysis, while Table 2 lists the sample sizes for the broad race analysis. Samples in the detailed population analysis are also included in the broad analysis, as indicated by the "Broad race code" column in Table 2.

## 3. Results

High resolution HLA A∼C∼B∼DRB3/4/5∼DRB1∼DQB1 haplotype frequencies were estimated for 21 detailed and five broad populations and are available online (*http://bioinformatics.nmdp.org/haplotype2011*). Allele frequencies for each of the six HLA loci are also provided. We have also developed an online tool, HaploStats, that can predict the haplotypes contributing to an unphased HLA genotype based on these haplotype frequencies (http://www.haplostats.org).

The most common haplotype observed within any population was A∗24:02 g∼C∗12:02∼B∗52:01 g∼DRB5∗01:02 g∼DRB1∗15:02 ∼DQB1∗06:01, found in Japanese (JAPI) at a frequency of 7.8%. While also seen in Koreans (KORI) at 1.9%, this haplotype was uncommon in other populations, even among other Asian/Pacific Islanders. The haplotype A∗01:01 g∼C∗07:01 g∼B∗08:01 g∼DRB3∗01:01∼DRB1∗03:01∼DQB1∗02:01 g was the next most common within population haplotype at 6.5% in European Caucasians (EURCAU), but was also common across other populations except Asians. Supplementary Table 3 summarizes frequencies for the 100 most common haplotypes across populations.

The HLA-B locus had the highest allelic diversity ranging from 89 alleles observed in Alaskan Natives or Aleuts (ALANAM) to 530 in European Caucasians (EURCAU), while the DRB3/4/5 superlocus had the least diversity (Table 3). A large fraction of the alleles described in the IMGT/HLA database version 3.4.0 at the protein level were never seen in any of the 21 populations (510 of 1253 alleles observed in IMGT/HLA for A, 461 of 833 for C, 646 of 1703 for B, 32 of 75 for DRB3/4/5, 228 of 714 for DRB1, and 35 of 107 for DQB1). A more detailed summary of the number of common (frequency of >1/2000), rare (<1/2000), and alleles not observed by population is found in Supplementary Table 4. We also compare allele frequencies across populations by locus in Supplementary Table 5.

The number of haplotypes with an estimated count of greater than one varied from 627 in Alaskan Natives or Aleuts (ALANAM) to 37,215 in European Caucasians (EURCAU) (Table 3). Populations with larger sample sizes tended to have more alleles and haplotypes observed. The number of haplotypes required to reach a cumulative 50% frequency varied from 88 in Vietnamese to 871 in African Americans (AAFA), keeping in mind this metric is also sample size dependent. More isolated and more narrowly defined populations had fewer common haplotypes (1/2000 in frequency or greater), while populations with high genetic diversity and high admixture [19,20] had haplotype frequency distributions with more rare HLA haplotypes. A spreadsheet containing the ten most common haplotypes from the perspective of each population is available in Supplementary Table 6.

To validate the performance of our haplotype frequency estimation method, we compared five-locus A∼C∼B∼DRB1∼DQB1 haplotype frequencies in European Caucasians from our full registry dataset with a previous study of 12,768 Caucasian registry volunteers [4]. We found a very similar frequency distribution, indicating that our method is able resolve both phase and allelic ambiguity well for common haplotypes (Fig. 1). However, frequency estimates derived from the entire registry for a given haplotype were always somewhat lower, primarily because of much larger sample sizes. As sample size increases, newly observed haplotypes decrease the frequency of common haplotypes. Another contributor to the differences is that allelic ambiguity in the full registry typings also could not be resolved for all rare haplotypes, just as some phase ambiguity cannot be resolved in high resolution typings.

HLA polymorphism was studied for all 21 populations within and among the five major broad population categories. The distribution of haplotype frequencies by race/ethnicity is shown in Supplementary Fig. 7A–E. The height of each curve denotes the percentage of HLA haplotypes in the population with frequency less than the value on the horizontal axis. The median haplotype frequency (the value at which there is a 50% chance that a randomly selected haplotype would have a greater frequency) varied from 1.7E-3 in Alaskan Natives or Aleuts to 1.9E-4 in African-Americans. On average the HLA polymorphism is highest among African–Americans and lowest among Native Americans with Asian, European Caucasian and Hispanic populations falling at intermediate frequencies. The differences in polymorphism

**Table 1**
Size of haplotype frequency-generating samples by detailed race category. The "Count" column indicates the number of A, B, DRB1-typed samples, and the "Typed C", "Typed DQB1"; "Typed DRB3/4/5" columns indicate the number of those individuals with typing at those loci.

| Race code | Detailed race/ethnic description | Broad race group | Count | Typed C | Typed DQB1 | Typed DRB3/4/5 |
|---|---|---|---|---|---|---|
| AAFA | African American | AFA | 416581 | 99946 | 16178 | 134076 |
| AFB | African | AFA | 28557 | 6975 | 1488 | 8516 |
| AINDI | South Asian Indian | API | 185391 | 29635 | 8409 | 44484 |
| AISC | American Indian – South or Central Am. | NAM | 5926 | 1255 | 228 | 894 |
| ALANAM | Alaska native or Aleut | NAM | 1376 | 288 | 100 | 347 |
| AMIND | North American Indian | NAM | 35791 | 7006 | 2398 | 13821 |
| CARB | Caribbean black | AFA | 33328 | 10012 | 1856 | 9115 |
| CARHIS | Caribbean hispanic | HIS | 115374 | 21286 | 4420 | 31097 |
| CARIBI | Caribbean Indian | NAM | 14339 | 5631 | 937 | 1372 |
| EURCAU | European caucasian | CAU | 1242890 | 395676 | 81106 | 212472 |
| FILII | Filipino | API | 50614 | 15272 | 1919 | 14738 |
| HAWI | Hawaiian or other Pacific Islander | API | 11499 | 3110 | 505 | 3355 |
| JAPI | Japanese | API | 24582 | 3552 | 852 | 7886 |
| KORI | Korean | API | 77584 | 11656 | 2107 | 25082 |
| MENAFC | Middle Eastern or N. Coast of Africa | CAU | 70890 | 22337 | 4415 | 17609 |
| MSWHIS | Mexican or Chicano | HIS | 261235 | 50875 | 12721 | 85021 |
| NCHI | Chinese | API | 99672 | 16621 | 3753 | 23569 |
| SCAHIS | Hispanic – South or Central American | HIS | 146714 | 31446 | 5764 | 29331 |
| SCAMB | Black – South or Central American | AFA | 4889 | 927 | 203 | 1677 |
| SCSEAI | Southeast Asian | API | 27978 | 5579 | 1321 | 3946 |
| VIET | Vietnamese | API | 43540 | 10511 | 1032 | 2446 |

**Table 2**
Size of haplotype frequency-generating samples by broad race category. The "Count" column indicates the number of A, B, DRB1-typed samples, and the "Typed C", "Typed DQB1"; "Typed DRB3/4/5" columns indicate the number of those individuals with typing at those loci.

| Broad race code | Race/ethnic description | Count | Typed C | Typed DQB1 | Typed DRB3/4/5 |
|---|---|---|---|---|---|
| AFA | African American | 505 250 | 123 871 | 21 408 | 156 764 |
| API | Asian or Pacific Islander | 568 597 | 104 027 | 21 814 | 142 755 |
| CAU | Caucasian | 3 912 440 | 1 808 061 | 502 117 | 1 596 577 |
| HIS | Hispanic | 712 764 | 166 192 | 31 700 | 163 539 |
| NAM | Native American | 46 148 | 9 533 | 2 977 | 15 469 |

**Table 3**
Number of haplotypes and alleles observed in 21 detailed populations. "Haplos" – Number of estimated haplotypes, "Haplos > 1" – Number of haplotypes with an estimated count of greater than 1, "HaplosTo50" – Number of haplotypes required to reach 50% frequency.

| Race code | Haplos | Haplos > 1 | HaplosTo50 | Number of alleles observed | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | | | A | C | B | DRB3/4/5 | DRB1 | DQB1 |
| AAFA | 167349 | 32976 | 871 | 200 | 119 | 328 | 24 | 181 | 25 |
| AFB | 41256 | 8595 | 829 | 150 | 64 | 197 | 12 | 102 | 20 |
| AINDI | 64210 | 15594 | 318 | 167 | 70 | 242 | 15 | 160 | 23 |
| AISC | 15717 | 2415 | 279 | 77 | 39 | 194 | 10 | 77 | 16 |
| ALANAM | 6616 | 627 | 95 | 59 | 27 | 89 | 10 | 47 | 16 |
| AMIND | 27470 | 6193 | 238 | 123 | 60 | 211 | 15 | 113 | 20 |
| CARB | 45617 | 9004 | 721 | 133 | 59 | 191 | 14 | 110 | 19 |
| CARHIS | 54190 | 13975 | 295 | 156 | 55 | 243 | 16 | 145 | 19 |
| CARIBI | 18606 | 4253 | 265 | 91 | 46 | 170 | 10 | 88 | 18 |
| EURCAU | 132724 | 37215 | 203 | 317 | 184 | 530 | 32 | 278 | 41 |
| FILII | 26035 | 5722 | 98 | 101 | 63 | 173 | 13 | 106 | 18 |
| HAWI | 15903 | 2621 | 110 | 89 | 37 | 160 | 10 | 83 | 15 |
| JAPI | 18329 | 3301 | 104 | 92 | 37 | 147 | 12 | 73 | 15 |
| KORI | 30070 | 6713 | 141 | 127 | 47 | 174 | 15 | 93 | 18 |
| MENAFC | 53507 | 13176 | 573 | 166 | 93 | 255 | 18 | 157 | 17 |
| MSWHIS | 103378 | 21423 | 381 | 211 | 97 | 305 | 21 | 169 | 25 |
| NCHI | 48794 | 9557 | 169 | 131 | 57 | 228 | 14 | 135 | 17 |
| SCAHIS | 98162 | 20101 | 671 | 203 | 82 | 327 | 16 | 161 | 22 |
| SCAMB | 17566 | 2493 | 619 | 72 | 35 | 166 | 10 | 61 | 16 |
| SCSEAI | 34633 | 6716 | 414 | 92 | 54 | 175 | 14 | 113 | 18 |
| VIET | 28544 | 5186 | 88 | 87 | 47 | 150 | 11 | 76 | 17 |

can be attributed to a combination of sample size and the HLA diversity of the population being sampled. For smaller population samples such as Alaskan Natives or Aleuts, there were plateaus in frequency for haplotypes with a count of one or two, while for larger populations the frequency distribution is smoother at this scale. Because of the paucity of complete HLA typing, 6-locus estimates have lower accuracy when sample size is small, however the 3-locus A∼B∼DRB1 estimates may be more robust.

Fig. 2 shows the Nei's genetic distance using haplotype frequencies obtained by the nearest neighbor approach. Major continental groups tended to appear in separate branches of the tree, while admixed Hispanics are in intermediate levels. The longer length of the branches of the tree between the Asian populations shows

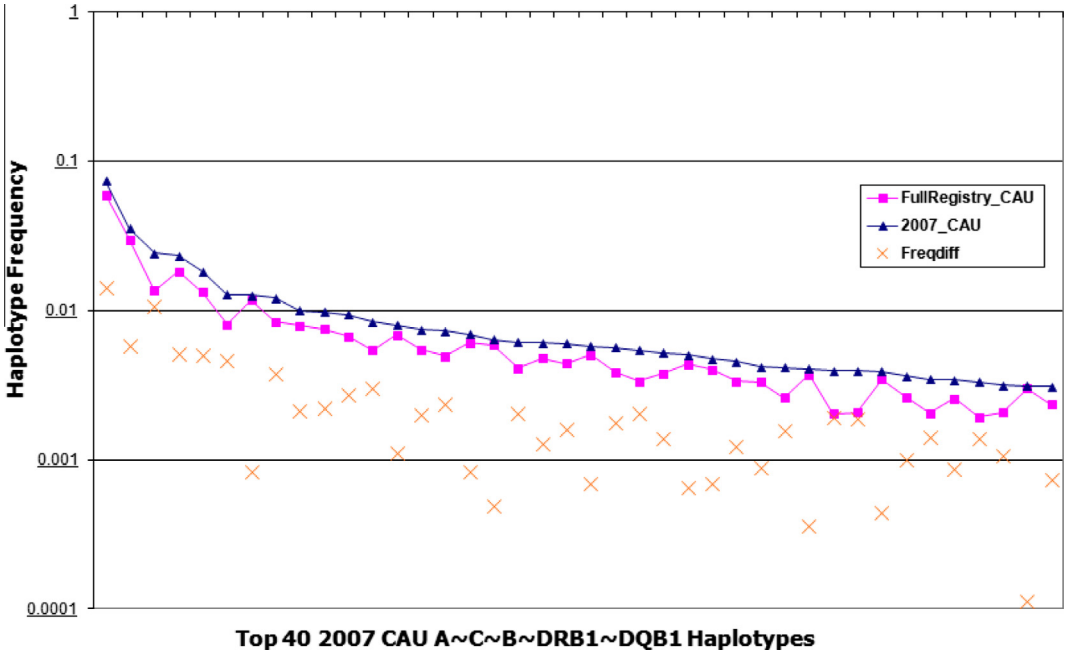


**Fig. 1.** Comparison of top 40 5-locus A~C~B~DRB1~DQB1 haplotype frequencies between Caucasians from the full registry dataset and the 2007 high resolution dataset [1]. "Freqdiff" is the difference in frequency between the 2 datasets.

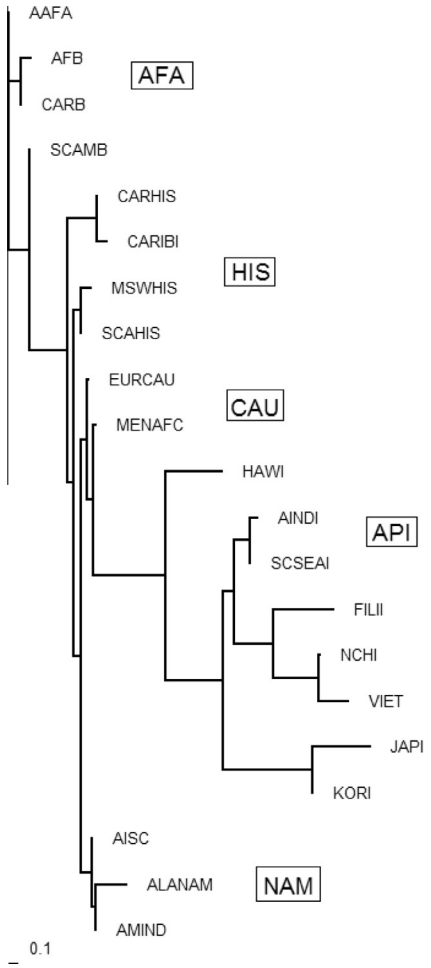


**Fig. 2.** Nei's Genetic Distance calculated using haplotype frequencies for 21 populations. Corresponding broad race categories for subtrees are shown. Populations are described in Table 1.

more dramatic HLA variation within Asia, while other populations are more similar to one another within broad race categories.

We used principal components to visualize haplotype frequency variation among the 21 populations in Fig. 3A and B. The first three principal components represent 29%, 19%, and 12% of the total variation in frequency. The first principal component does well at distinguishing between continental groups. The second principal component only separates Asian populations from Japanese on one end to Vietnamese on the other, again illustrating the relatively high level of population differentiation among the Asian–American groups.

The summed pairwise Kendall rank correlations of haplotype frequencies measures the similarity of each population against all others. This is illustrated by a depiction of pairwise correlations among haplotype frequencies (Fig. 4). Populations with high levels of admixture or European populations who are the ancestral sources

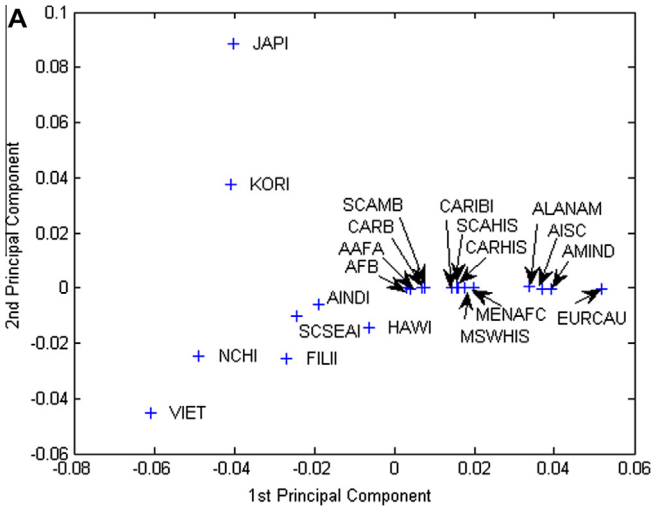


**Fig. 3.** (A) First two principal components of haplotype frequencies for 21 populations. (B) First and third principal components of haplotype frequencies for 21 populations.

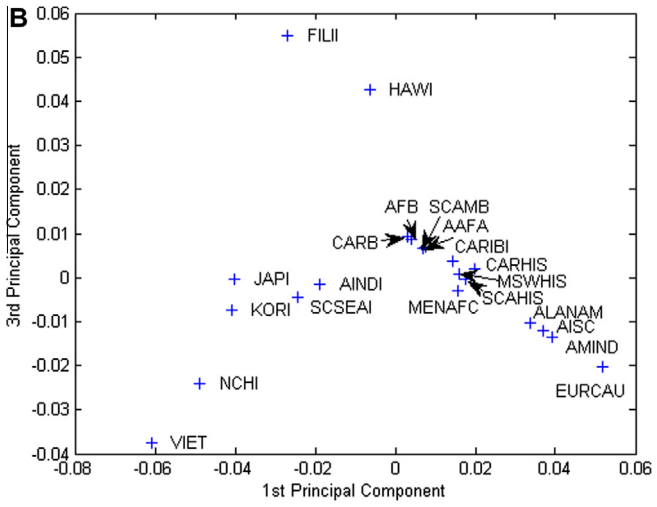


**Fig. 3** (*continued*)

of admixture, on the right and bottom of the figure, have relatively high frequency correlation with other populations. The Middle Eastern/North Coast of Africa population was the most similar to the others in terms of haplotype frequencies. Meanwhile, more isolated Asian populations and Alaskan Natives or Aleuts on the top and left of the figure were the most dissimilar from one another.

## 4. Discussion

The six-locus HLA A∼C∼B∼DRB1∼DRB3/4/5 haplotype frequencies of populations described here improve upon previous HLA haplotype studies in the dimensions of sample size, number of loci, population specificity, and in the comprehensive use of available typing data. Because of the large sample size, it is possible to evaluate the relative frequencies of many rare alleles and haplotypes for the first time. Interestingly, hundreds of alleles described in the IMGT/HLA database were not seen in our population study consisting of 6.59 million individuals, suggesting that they are unlikely to be seen again.

The accuracy of HLA haplotype frequency estimates depends on the typing method employed. Early HLA typing based on serology had high levels of mistyping and therefore was not deemed adequate for this study, which utilized only donors typed by DNA based methods. A large fraction of recruitment typing used oligonucleotide probe hybridization, which provides limited coverage of polymorphic sites in the HLA genes. As a result, we observed that some alleles were never distinguished from one another in any individual within a given population, for example DRB1∗03:18 and DRB1∗03:28 in European Caucasians. DNA sequencing based methods have lower ambiguity, although some genotypic ambiguity remains in diploid sequences. Higher resolution typing identifies alleles to the specific protein level, but still does not give haplotype phase information. As ambiguity decreases, ascertainment of haplotype frequencies improves.

One barrier to calculating haplotype frequencies has been variation in the form of HLA typing data due to changes in typing methods and reagents. HLA data representation also impacts the amount of information derived from an HLA assignment. Unfortunately, the NMDP allele letter code system [9], commonly used to compress lists of alternative genotypes, can result in inclusion of genotypes that were excluded by the typing method. To counter this we used HLA typings in genotype list format computed from re-interpreted primary data when available. Furthermore, without

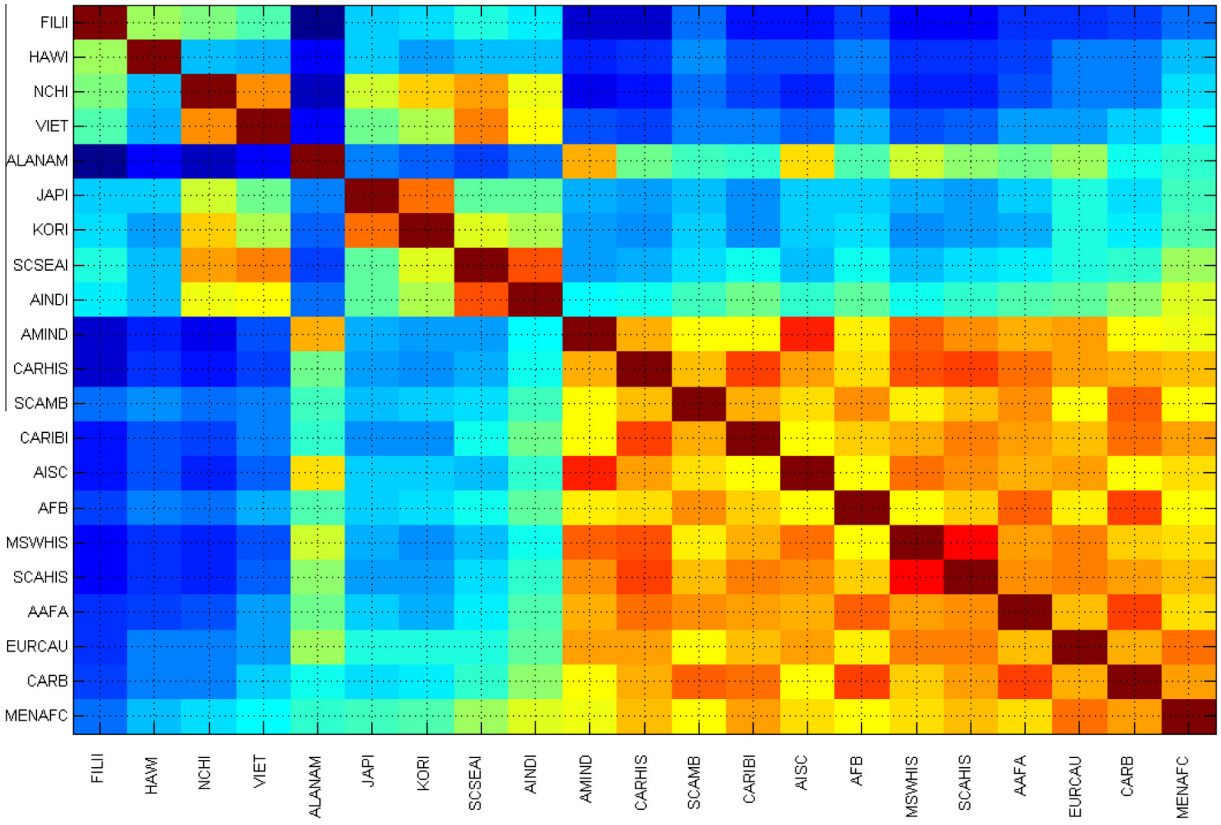


**Fig. 4.** Heat map of Kendall rank correlation of population haplotype frequencies ordered by the sum correlation ranging from high correlation (red) to low correlation (blue).

primary typing data, the only allelic possibilities considered in the laboratory assignment are those alleles described at the time the typing was performed, even though new alleles are constantly discovered, so should be included as possibilities. We encourage the use of HLA data standards, such as the genotype list format, that transmit the maximum amount of information about the tests performed [21].

The runtime of frequency estimation algorithms is sensitive to population sample size, genetic diversity, and HLA typing ambiguity. Because of extensive computational challenges in estimating haplotype frequencies from highly ambiguous typing data, heuristics were required to reduce the amount of ambiguity. Abridging the allele list to only the alleles required to describe all subjects may result in a slight bias towards common alleles. To achieve computational tractability, low probability haplotype pairs were explicitly removed from consideration by EM as haplotype blocks were extended, which also biased the frequencies slightly towards common haplotypes.

Improved haplotype frequencies aid in donor selection processes, shortening the time-to-transplant from preliminary search, and reducing the number of samples fruitlessly tested for a match with extended typing. We observed such improvements after the HapLogic™ III matching algorithm, released in 2011, began utilizing the frequencies described here. This matching algorithm orders the list of potentially matched donors in a search report for a given patient by the donor's likelihood of being allele matched. HapLogic's™ match probability calculations begin with imputation, which produces a list of a subject's possible most likely phased multilocus genotypes and their corresponding genotype probabilities, given the HLA typing and population haplotype frequencies. Using more population categories gives more accurate predictions because HLA frequencies can vary dramatically within the broader race/ethnic categories [22]. Using haplotype frequency distributions calculated from more subjects decreases sampling error and results in fewer cases where no possible haplotype pairs can explain a subject's HLA typing. Typing information at the DRB3/4/5 locus, while not often considered in clinical matching between donor and recipient, can help infer alleles present at DRB1 and DQB1 loci. The practical operational use of these frequencies is being realized today.

Modeling of HLA match rates can be performed using HLA haplotype frequencies [23]. A population genetic model incorporating registry size and assuming random assortment of haplotypes in individuals is an effective predictor of match rates for a given patient population, assuming haplotype frequencies adequately reflect the overall population. Large population samples in the range of tens of thousands per population are required for modeling match rates because the shape of the haplotype frequency distribution in a large part determines the match rate, and poor sampling can truncate this distribution relative to the true population.

Many disease association and other research studies of large populations are limited by the high cost of HLA typing, the complexity of the data, and thus may only analyze the data at low resolution. Applying HLA imputation using haplotype frequencies can inexpensively reduce HLA typing ambiguities post hoc [24], providing high resolution associations. In addition, algorithms that have been developed to infer HLA by linkage with other SNPs in genome wide studies where HLA is not specifically typed can benefit from improved haplotype frequency reference data [25].

We plan on making regular public updates to these frequency estimates over time. While HLA frequencies would ideally change very little over time, continued accumulation of volunteer donors and higher resolution HLA typing methods will yield continually improving haplotype estimates, especially for rare types in minority populations. These frequencies complement a prior analysis in which each HLA allele was categorized as common or rare based on observations in laboratories around the world [11]. A larger proportion of high resolution or sequence based typing would give more accurate allelic determination within allele families. Because of the high level of allelic ambiguity in the input data, a large number of haplotypes had estimated fractional counts compared to what would be seen for high resolution typed datasets where only phase is estimated. Most donors lacked typing at HLA-C, HLA-DQB1, and HLA-DRB3/4/5, so more comprehensive typing could improve allelic determination especially at these loci. The DQ molecule also consists of the polymorphic DQA1 gene product, which is not considered here. Similarly inclusion of the rarely typed DPA1 and DPB1 genes would give complete extended haplotypes for these polymorphic HLA loci [26]. Public accessibility to the best available HLA data has significant benefits to the immunogenetics community.

The number of specific population categories may change over time as sample size increases and volunteer donor race/ethnicity information is improved. For example, the current race/ethnic categories in the recruitment questionnaire may not adequately capture the individual's ancestry. For this study we also did not address the issue of individuals listing more than one race, which are becoming increasingly common in the US [27]. Multi-race individuals are complex to analyze because the number of different population combinations reduces sample size and high proportions of first generation admixture leads to divergence from Hardy–Weinberg equilibrium assumed by the EM algorithm. Cryptic population substructure, such as the Ashkenazi Jewish subpopulation within Europeans, also confounds this analysis [28]. As improved methods for self reporting population identity and measuring genetic ancestry are applied to donor registries, more distinct subpopulations can be captured. Analyzing world populations more comprehensively through application of our described methods to the registries of Bone Marrow Donors Worldwide (BMDW) could substantially improve the global donor search process [29].

We conclude that these reference haplotype frequencies are of significant practical use in the hematopoietic stem cell registry, clinical transplant, and other immunologic research settings.

## Acknowledgments

Bioinformatics methods development has been funded by Office of Naval Research Grant N00014-11-1-0339. William Klitz from University of California-Berkeley for his work in helping to identify operationally useful race/ethnic categories. The International Histocompatibility and Immunogenetics Workshop (IHIW) Registry Diversity working group for helpful collaboration on estimating haplotype frequencies from registry data. Europdonor Foundation, Hadassah Medical Organization, DKMS, Knochenmarkspenderzentrale Dusseldorf, Tobias Registry of Swedish Bone Marrow Donors, Norwegian Bone Marrow Donor Registry, The Caitlin Raymond International Registry, and Gift of Life Bone Marrow Foundation for allowing us to integrate their public registry typing data in our analysis. Gift of Life is a US based affiliate registry of the NMDP with significant representation of Jewish donor populations, founded by transplant recipient Jay Feinberg. We also thank Carolyn Hurley from Georgetown for her review of the manuscript.

## Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at http://dx.doi.org/10.1016/j.humimm.2013.06.025.

## References


[1] Lee SJ, Klein J, Haagenson M, Baxter-Lowe LA, Confer DL, Eapen M, et al. High-resolution donor-recipient HLA matching contributes to the success of unrelated donor marrow transplantation. Blood 2007;110(13):4576–83.

[2] Excoffier L, Slatkin M. Maximum-likelihood estimation of molecular haplotype frequencies in a diploid population. Mol Biol Evol 1995;12(5):921–7.

[3] Long JC, Williams RC, Urbanek M. An E-M algorithm and testing strategy for multiple-locus haplotypes. Am J Hum Genet Am Soc Hum Genet 1995;56(3):799–810.

[4] Maiers M, Gragert L, Klitz W. High-resolution HLA alleles and haplotypes in the United States population. Hum Immunol 2007;68(9):779–88.

[5] Kollman C, Maiers M, Gragert L, Muller C, Setterholm M, Oudshoorn M, et al. Estimation of HLA-A, -B, -DRB1 haplotype frequencies using mixed resolution data from a national registry with selective retyping of volunteers. Hum Immunol 2007;68(12):950–8.

[6] Eberhard H-P, Feldmann U, Bochtler W, Baier D, Rutt C, Schmidt AH, et al. Estimating unbiased haplotype frequencies from stem cell donor samples typed at heterogeneous resolutions: a practical study based on over 1 million German donors. Tissue Antigens 2010;76(5):352–61.

[7] Marsh SGE, Albert ED, Bodmer WF, Bontrop RE, Dupont B, Erlich HA, et al. Nomenclature for factors of the HLA system, 2010. Tissue Antigens 2010;75(4):291–455.

[8] Maiers M. A community standard XML message format for sequencing-based typing data. Tissue Antigens 2007;69(s1):69–71.

[9] Maiers M, Hurley CK, Perlee L, Fernandez-Vina M, Baisch J, Cook D, et al. Maintaining updated DNA-based HLA assignments in the national marrow donor program bone marrow registry. Rev Immunogenet 2000;2(4):449–60.

[10] Dempster AP, Laird NM, Rubin DB. Maximum likelihood from incomplete data via the EM algorithm. J R Stat Soc Ser B (Methodological) R Stat Soc 1977;39(1):1–38.

[11] Cano P, Klitz W, Mack SJ, Maiers M, Marsh SGE, Noreen H, et al. Common and well-documented HLA Alleles Report of the Ad-Hoc committee of the American Society for histocompatiblity and immunogenetics. Hum Immunol 2007;68(5):392–417.

[12] Mack SJ, Cano P, Hollenbach JA, He J, Hurley CK, Middleton D, et al. Common and well-documented HLA alleles: 2012 update to the CWD catalogue. Tissue Antigens 2013;81(4):194–203.

[13] De Bakker PIW, McVean G, Sabeti PC, Miretti MM, Green T, Marchini J, et al. A high-resolution HLA and SNP haplotype map for disease association studies in the extended human MHC. Nat Genet 2006;38(10):1166–72.

[14] Fernandez-Viña MA, Gao XJ, Moraes ME, Moraes JR, Salatiel I, Miller S, et al. Alleles at four HLA class II loci determined by oligonucleotide hybridization and their associations in five ethnic groups. Immunogenetics 1991;34(5):299–312.

[15] Tautz C, Marsh DG, Baur X. A novel HLA-haplotype containing a DRB5 gene not associated with DRB1∗15 or DRB1∗16 alleles. Tissue Antigens 1992;39(2):91–4.

[16] Nei M. Genetic distance between populations. Am Nat 1972;106(949):282–92.

[17] Retief JD. Phylogenetic analysis using PHYLIP. Methods Mol Biol 2000;132:243–58.

[18] MATLAB version 7.13.0.564 Natick, Massachusetts: The Mathworks, Inc. Natick, Massachusetts: The Mathworks Inc; 2011.

[19] Klitz W, Gragert L, Maiers M, Tu B, Lazaro A, Yang R, et al. Four-locus high-resolution HLA typing in a sample of Mexican Americans. Tissue Antigens 2009;74(6):508–13.

[20] Hollenbach JA, Fernandez-Vina M, Thomson G, Cao K, Erlich HA, Bugawan TL, et al. HLA diversity, differentiation, and haplotype evolution in Mesoamerican Natives. Hum Immunol 2001;62(4):378–90.

[21] Hollenbach JA, Mack SJ, Gourraud P-A, Single RM, Maiers M, Middleton D, et al. A community standard for immunogenomic data reporting and analysis: proposal for a STrengthening the REporting of Immunogenomic Studies statement. Tissue Antigens 2011;78(5):333–44.

[22] 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.

[23] Kollman C, Abella E, Baitty RL, Beatty PG, Chakraborty R, Christiansen CL, et al. Assessment of optimal size and composition of the US National Registry of hematopoietic stem cell donors. Transplantation 2004;78(1):89–95.

[24] Gourraud P-A, Lamiraux P, El-Kadhi N, Raffoux C, Cambon-Thomsen A. Inferred HLA haplotype information for donors from hematopoietic stem cells donor registries. Hum Immunol 2005;66(5):563–70.

[25] Leslie S, Donnelly P, McVean G. A statistical method for predicting classical HLA alleles from SNP data. Am J Hum Genet 2008;82(1):48–56.

[26] Hollenbach JA, Madbouly A, Gragert L, Vierra-Green C, Flesch S, Spellman S, et al. A combined DPA1∼DPB1 amino acid epitope is the primary unit of selection on the HLA-DP heterodimer. Immunogenetics 2012;64(8):559–69 [13].

[27] US Census Bureau. Statistical Abstract of the United States: 2012 (131st edition). DC: Washington; 2011.

[28] 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 Genet 2008;4(7):e1000114.

[29] Van Rood JJ, Oudshoorn M. Eleven million donors in bone marrow donors worldwide! Time for reassessment? Bone Marrow Transplant 2007;41(1):1–9.