Title
HLA alleles and haplotypes observed in 263 US families

Journal
Human Immunology, 80(9)

ISSN
0198-8859

Authors
Osoegawa, Kazutoyo
Mallempati, Kalyan C
Gangavarapu, Sridevi
et al.

Publication Date
2019-09-01

DOI
10.1016/j.humimm.2019.05.018

Peer reviewed

HHS Public Access

Author manuscript

Hum Immunol. Author manuscript; available in PMC 2020 September 01.

Published in final edited form as:

Hum Immunol. 2019 September ; 80(9): 644–660. doi:10.1016/j.humimm.2019.05.018.

HLA Alleles and Haplotypes Observed in 263 US Families

Kazutoyo Osoegawa[1], **Kalyan C. Mallempati**[1], **Sridevi Gangavarapu**[1], **Arisa Oki**[2], **Ketevan Gendzekhadze**[2], **Susana R. Marino**[3], **Nicholas K. Brown**[3], **Maria P. Bettinotti**[4], **Eric T. Weimer**[5], **Gonzalo Montero-Martín**[1,6], **Lisa E. Creary**[1,6], **Tamara A. Vayntrub**[1], **Chia-Jung Chang**[7], **Medhat Askar**[8], **Steven J. Mack**[9], **Marcelo A. Fernández-Viña**[1,6]

[1]Histocompatibility, Immunogenetics & Disease Profiling Laboratory, Stanford Blood Center, Palo Alto, CA, USA

[2]HLA Laboratory, City of Hope, Duarte, CA, USA

[3]Transplant Immunology Laboratory, The University of Chicago Medicine, Chicago, IL, USA

[4]Immunogenetics Laboratory, Johns Hopkins University, Baltimore, USA

[5]Department of Pathology & Laboratory Medicine, UNC Chapel Hill School of Medicine, Chapel Hill, NC, USA

[6]Department of Pathology, Stanford University School of Medicine, Palo Alto, CA, USA

[7]Stanford Genome Technology Center, Palo Alto, CA, USA

[8]Baylor University Medical center, Dallas, TX, USA

[9]Center for Genetics, Children's Hospital Oakland Research Institute, Oakland, CA, USA

Abstract

The 17th International HLA and Immunogenetics Workshop (IHIW) conducted a project entitled "The Study of Haplotypes in Families by NGS HLA". We investigated the HLA haplotypes of 1,017 subjects in 263 nuclear families sourced from five US clinical immunogenetics laboratories, primarily as part of the evaluation of related donor candidates for hematopoietic stem cell and solid organ transplantation. The parents in these families belonged to five broad groups – African (72 parents), Asian (115), European (210), Hispanic (118) and "Other" (11). High-resolution HLA genotypes were generated for each subject using next-generation sequencing (NGS) HLA typing systems. We identified the HLA haplotypes in each family using HaplObserve software that builds haplotypes in families by reviewing HLA allele segregation from parents to children. We calculated haplotype frequencies within each broad group, by treating the parents in each family as unrelated individuals. We also calculated standard measures of global linkage disequilibrium (LD)

Corresponding Author: Kazutoyo Osoegawa, kazutoyo@stanford.edu, Histocompatibility, Immunogenetics & Disease Profiling Laboratory, Stanford Blood Center, 3155 Porter Drive Palo Alto, CA 94304 USA.

Conflicts of Interests

The authors declared no conflicting interests.

Online Resources

and conditional asymmetric LD for each ethnic group, and used untruncated and two-field allele names to investigate LD patterns. Finally we demonstrated the utility of consensus DNA sequences in identifying novel variants, and confirming them using HLA allele segregation at the DNA sequence level.

Keywords

HLA haplotype; Linkage disequilibrium; Family

1. INTRODUCTION

The classical Human Leukocyte Antigen (HLA) genes are located in a 3.6 Mb region at chromosome 6p21.3, and show the highest density of single nucleotide polymorphisms (SNPs) in the human genome [1, 2]. The high-levels of allelic diversity within these genes have evolved through intra- and intergenic recombination and short-tract gene conversions [3-5]. It is unmanageably complex to characterize HLA gene diversity in terms of SNPs. To foster the scientific interpretation of the highly complex HLA gene in a clinically meaningful manner, the HLA nomenclature committee established a nomenclature system that assigns unique allele names based on the constellation of polymorphism within the genes [6]. As of March 2019, more than 21,000 alleles have been catalogued in the IPD-IMGT/HLA Database (release version 3.35.0) for 11 classical HLA genes (*HLA-A, HLA-C, HLA-B, HLA-DRB3, HLA-DRB4, HLA-DRB5, HLA-DRB1, HLA-DQA1, HLA-DQB1, HLA-DPA1*, and *HLA-DPB1*) [7]. HLA alleles of neighboring genes (e.g., *HLA-C* and *HLA-B*) often display strong linkage disequilibrium (LD) [8], forming haplotype blocks [9] (sets of alleles at two or more loci that share chromosomal phase). In allogeneic hematopoietic transplantation, high-resolution HLA genotype and haplotype matching between donors and patients correlates highly with improved clinical outcomes [10-12]. In addition to the role in transplantation, many studies have identified that certain HLA alleles and haplotypes associate with susceptibility or resistance to development of autoimmune diseases [13].

HLA haplotypes can be estimated from unrelated individuals in various ethnic groups using the expectation-maximization (EM) algorithm [14-18]. Accurate haplotype frequencies are vital for hematopoietic stem cell patient/donor match predictions and facilitate the identification of suitable HLA-matched donors for more patients [19, 20]. Clinical specialists frequently use published EM haplotypes as reference data when reviewing HLA typing data in attempts to verify and predict the expected haplotypes from HLA types. In general, high-frequency EM haplotypes have been found to be reliable [18, 19]. However, HLA haplotype frequencies typically follow a heavy-tailed distribution across all population/ethnic groups, which indicates the presence of rare haplotypes in all populations [21]. This skewed distribution results in difficulties to accurately predict low-frequency haplotypes via the EM approach [22]. In addition, rare haplotypes may not be present in published reference tables, and in some cases, published haplotype frequencies may be overestimated [23].

A family-based approach offers a different strategy for phasing HLA haplotypes based on HLA allele segregation [24]. This approach makes it feasible to accurately determine chromosomal phase in samples of small size, even a single family, although the resulting haplotype frequencies may not be accurate. We have high confidence in haplotypes inferred from family studies because the haplotypes are built based on observations of HLA allele segregation within a single family. Ha plotypes built from families are *observed* haplotypes, and thus all inferred haplotypes may exist in the human population. A family-based approach also makes it possible to identify newly generated haplotypes that result from crossover events.

Five US clinical HLA Immunogenetics laboratories collected anonymized DNA specimens originally drawn for evaluating candidates for hematopoietic stem cell and solid organ transplantation and their corresponding related donor. The DNA samples were sequenced and genotyped using NGS HLA typing systems in each laboratory. Three computational programs were developed for building HLA haplotypes from families, calculating haplotype frequencies, estimating standard measures of global (locus-level) LD, and comparing the family-based haplotypes with EM-based haplotypes as part of the 17th IHIW Informatics of Genomic Data component [25]. An HLA DNA sequence alignment tool (hlaPoly) was developed for reporting novel variants in the consensus DNA sequences relative to reported HLA alleles [26]. We applied these computational tools to analyze HLA genotypes and consensus DNA sequences generated for the Study of Haplotypes in Families by NGS HLA (Family haplotype) project in the 17th IHIW database [26]. Here, we report the inferred haplotypes derived from nuclear or single-parent families, and present a strategy for the identification and confirmation of novel variants using HLA allele sequence segregation. This work was performed as a project of the HLA of NGS component of the 17th IHIW, and preliminary results were presented at the 17th IHIW in September of 2017.

2. MATERIALS AND METHODS

2.1 Families and subjects

Five HLA Immunogenetics laboratories in the United States collected blood specimens from 1017 subjects in 263 families as part of related bone marrow and solid organ transplantation donor recruitments (Supplemental Table 1). De-identified subjects were registered in the 17th IHIW database by each participating laboratory, and double-blinded IHIW subject IDs were automatically generated by the database [26]. Each participating laboratory entered the self-identified ethnicity [African American (AFA), Asian American (ASI), European American (EUR), Hispan ic American (HIS) and Other (OTH)] and familial relationships for their subjects into the IHIW database (Table 1). The analyses of HLA genotype data with the double-blinded sample IDs were conducted at the Stanford Blood Center and Stanford University under the Stanford University Institutional Review Board (IRB) eProtocol titled, “17th International HLA and Immunogenetics Workshop” (# : 38899).

2.2 HLA genotype data

DNA was extracted from blood, and the *HLA-A, HLA-C, HLA-B, HLA-DRB3/4/5, HLA-DRB1, HLA-DQA1, HLA-DQB1, HLA-DPA1* and *HLA-DPB1* genes were amplified. DNA

sequencing libraries were prepared using commercially available NGS HLA typing reagents [MiaFora FLEX (Immucor Inc.), TruSight HLA (Illumina Inc.) or NGSgo (GenDx Inc.)], and sequencing was performed on MiSeq (Illumina Inc.) or NextSeq (Illumina Inc.) instruments at each laboratory. HLA alleles were assigned based on the reference sequences in IPD-IMGT/HLA Database version 3.25.0, using the corresponding commercially available HLA typing software [MiaFora v3.1 (Immucor Inc.), TruSight HLA (Illumina Inc.) or NGSengine (GenDx Inc.)] by each laboratory. Supplemental Table 1 includes the NGS HLA genotyping systems used by each laboratory. Currently, there is no universally accepted standard format for reporting novel alleles. The workshop organizers designed the 17th IHIW database to collect standardized reports of novel alleles [26], but no 17th IHIW participant reported novel alleles in a computer analyzable format. It was not realistic to perform genetic analysis with unstructured, human readable “comments” using software. The various NGS HLA software vendors have different ways of reporting novel alleles, and different laboratories have their own “internal” way of indicating new alleles. There may be multiple novel alleles for the same “closest” known allele. It was not possible to collect unofficial novel alleles from multiple laboratories, and analyze the data in a standard fashion. To make the HLA genotype data analyzable using software, we accepted only untruncated official HLA allele names in IPD-IMGT/HLA Database version 3.25.0 in the 17th IHIW database [26]. When perfectly matched HLA alleles were not found (novel alleles), the closest HLA alleles that had the least numbers of mismatches were selected. The novel alleles were identified using consensus DNA sequences as described in section 2.9. HLA genotypes and corresponding consensus sequences (when available) were imported into the IHIW database in Histoimmunogenetics Makeup Language (HML) [27] or eXtensible Markup Language (XML) formats [26].

2.3 Inferred parents and their imputed genotypes

We used HaplObserve, a Java application that infers parent to child allele segregation [25] [A], to build HLA haplotypes from families (see section 2.4). HaplObserve requires genotypes from two parents and at least one child per family [25]. We also treated parents as unrelated individuals to calculate haplotype frequencies (see section 2.6). Thirty-six of the 263 families included a single parent, and one family included no parents. To include every *observed* parental HLA allele in the study, we manually assigned the 38 missing parental genotypes using segregation analysis from the available genotype data (Table 2). This practice is routinely performed in clinical HLA laboratories when dealing with families that are missing one or two parents [28, 29]. We defined the missing parent as an “inferred parent” (Supplemental Table 1, and Table 2). There were several cases in which only one parental allele at a locus could be assigned, because the second allele for the inferred parent was not found in the offspring’s genotypes, leaving the second allele unknown (Table 2B). In these cases, “No Type (NT)” (with the corresponding locus prefix, e.g., HLA-DPB1*NT) was used to represent the unknown allele. The alleles represented with NT were used as placeholders to build haplotypes using HaplObserve. Tables 2A and 2B describe how HLA genotypes were assigned for an inferred parent.

[A] HaplObserve: https://github.com/ihiw/haplObserve

2.4 Building haplotypes

HLA haplotypes were built from families using HaplObserve. We observed many untruncated allele name discordances in the originally submitted data; most of these cases were for class II allele names, which were concordant at the two-field level. When an HLA allelic discordance was detected for the segregation of alleles within a family, we carefully reviewed all the discordant alleles in the HLA genotyping software. If the discordances were identified to be HLA genotyping errors, HLA allele calls were manually reassigned on the basis of the polymorphic positions by carefully reviewing DNA sequence alignments in the graphical user interface of the HLA genotyping software used by each laboratory. The other obvious allele inconsistencies arose from complete HLA allele call dropouts (when no HLA allele was reported at a locus). When a single allele dropout was suspected for a discordant allele due to allelic imbalance, we corrected HLA genotypes if possible by carefully reviewing DNA sequence alignments in the graphical user interface of the pertinent HLA genotyping software. When the allele inconsistency was not resolved due to suspected DNA contamination or other artifacts, HLA genotyping was repeated. If the allele inconsistency was not resolved after repeating NGS HLA genotyping, the family was excluded from analysis. HaplObserve reports “NoMatch” when HLA allele segregation is discordant [25]. When multiple discordances were observed for a given allele, and these discordances were unresolvable and true, the discordant subject was treated as an unrelated family member and removed from the analyses. These manual haplotype inspections and revisions were typically performed multiple times until the allelic inconsistencies were completely resolved for a given family. When it was not feasible to automatically assign the correct haplotypes using HaplObserve, due to uninformative HLA genotypes within a family (for example, identical genotypes at a locus for all the family members), we manually adjusted haplotypes as described previously [25]. These haplotypes were “locked” on March 4, 2019 for this publication. The HLA alleles were reported in telomeric to centromeric order in the resulting haplotypes. The haplotypes were reported in Genotype List (GL) String format, which consists of two tildes (~) delimited HLA haplotypes connected by a plus (+) sign (*HLA-A~HLA-C~HLA-B~HLA-DRB3/4/5~HLA-DRB1~HLA-DQA1~HLA-DQB1~HLA-DPA1~HLA-DPB1*) [30]. The Supplemental Materials (FAM_Haplotype_Summary_GL_String_2019-03-04.csv) contains haplotypes for each family member. The resulting haplotypes were converted into comma separated value format for easy viewing (Family_Haplotype_Summary_Table_2019-03-04.csv).

2.5 Merging potentially identical haplotypes

Different NGS HLA genotyping vendors used different PCR primer sets, different sequencing library preparation protocols and different HLA genotyping software algorithms. During haplotype review, we observed potentially identical haplotypes that were distinguished by specific alleles, which were most likely due to these differences (Supplemental Table 2). For example, two slightly different haplotypes were reported as follows: *HLA-DRB1*07:01:01:01/HLA-DRB1*07:01:01:02~HLA-DQB1*02:02:01:01* and *HLA-DRB1*07:01:01:01/HLA-DRB1*07:01:01:02/HLA-DRB1*07:01:01:03~HLA-DQB1*02:02:01:01.* These haplotypes differ by the presence of *HLA-DRB1*07:01:01:03* in the ambiguity string for the second haplotype. The haplotype frequency estimation is affected because these potentially identical haplotypes are treated as distinct haplotypes.

When we observed potentially identical haplotypes that were distinguished due to differences that originated from the different NGS HLA genotyping protocols, we carefully reviewed HLA genotypes, and merged haplotypes if appropriate. Supplemental Table 2 shows the haplotypes that were merged.

2.6 Estimating HLA allele and haplotype frequencies

We treated parents as unrelated individuals and used their haplotypes to calculate frequencies within each broad continental origin group. Haplotypes containing missing alleles ("NT") were excluded when haplotype counts and frequencies were calculated. The fully phased haplotypes were divided into individual loci (*HLA-A, HLA-C, HLA-B, HLA-DRB3/4/5, HLA-DRB1, HLA-DQA1, HLA-DQB1, HLA-DPA1* and *HLA-DPB1*) or smaller haplotype blocks for calculating frequencies (Supplemental Table 3). Haplotype frequency tables are found in Supplemental Materials. For example, the haplotype ID for *HLA-DRB3/4/5~HLA-DRB1~HLA-DQA1~HLA-DQB1* haplotype is "DRDQ" as shown in Supplemental Table 3. Therefore, the *HLA-DRB3/4/5~HLA-DRB1~HLA-DQA1~HLA-DQB1* haplotype frequency table is found in "Global_DRDQ_Haplotype_Summary_2019-03-04.csv". The haplotype frequencies are also available on the 17th IHIW web site [B].

2.7 Evaluation of linkage disequilibrium

We calculated standard measures of global (locus-level) linkage disequilibrium (LD) (*D'* [31], *Wn* [32], and the conditional asymmetric LD (cALD) measures [33, 34]) using the "Phased Or Unphased Linkage Disequilibrium: pould" R package [25] [C]. The current version of pould (0.8.1.9000) also generates heatmaps of LD values for these measures. Any allelic ambiguities were collapsed to the lowest-field unambiguous name for use with pould. LD measures for two-locus phased and EM (unphased) haplotypes were calculated with untruncated and two-field allele names for each population. The Supplemental Materials contain pould input and output files for each ethnic group. For example, the input file of unambiguous alleles for the AFA group is "UnambiguousAllele_Haplotype_AFA_2019-03-04.csv"; its corresponding output file for phased haplotypes is "UnambiguousAllele_Haplotype_AFA_2019-03-04_Phased_LD_results.csv". Detailed documentation of "pould" output files can be obtained in IHIW GitHub pould repository [C].

2.8 Deviations from expected Hardy-Weinberg Equilibrium (HWE)

Python for Population Genomics (PyPop) version 0.7.0 [35] was used to investigate Hardy-Weinberg Equilibrium (HWE) via the Guo and Thompson test [36], and homozygosity via Slatkin's implementation of the Ewens–Watterson test [37, 38], for each broad continental group. For each test, statistical significance was evaluated at the p-value < 0.05 level. PyPop output files for each broad continental group are found in Supplemental Materials, for

[B] HLA haplotype frequency table: http://17ihiw.org/17th-ihiw-ngs-hla-data/
[C] Pould: https://github.com/IHIW/pould

example, “AFA_pop_v7_out.txt”. Detailed documentation of PyPop output files can be obtained in PyPop web site [D].

2.9 Analyzing consensus sequences

We developed hlaPoly [E], a web-based software tool that identifies sequence variants in consensus sequences relative to (1) user-specified reference HLA alleles, a.k.a. “closest allele”, and (2) locus-specific full-length reference alleles defined by the 17th IHIW Informatics Component [26]. All consensus sequences collected for this project were evaluated using hlaPoly with IPD-IMGT/HLA Database release version 3.25.0.

Depending on the typing method applied, a given consensus sequence may be longer or shorter than the available reference sequence. hlaPoly masks non-overlapping sequence regions using values of “N” for each position. We first systematically removed artificial variants containing “N” for the subsequent analyses.

Second, we analyzed variants that were only identified using the closest allele as a reference; in other words we eliminated variants that were identified using locus-specific full-length reference allele sequences, because these reference alleles were used when the closest alleles did not contain reference sequences for certain features. For example, the *HLA-DPA1*02:01:07* allele contained only exon 2 reference sequence. For this allele, the *HLA-DPA1*02:01:02* reference sequence was used for DNA sequence alignment for the missing exons and introns. When a novel variant is identified in the features other than exon 2 of *HLA-DPA1*02:01:07*, we excluded these variants.

In the nuclear family study, each child inherits a paternal and a maternal haplotype/allele set. To identify the likely true novel variants, we used a two-step filtering system in which the novel variants are first identified in the parental consensus sequences (first step), and then the same variants are identified in at least one of the children (second step). When the same novel variant is identified in a parent and a child who each have different second allele combinations, we consider the novel variant more likely to be true than when the novel variant is identified only once.

We reviewed the identified novel variants in the NGS HLA genotyping software and confirmed their validity. After we confirmed novel variants, the confirmed novel variants were used to locate individuals with same novel variants but who did not initially pass the two-step filtering. Finally, we compared the confirmed novel variants against IPD-IMGT/HLA Database release version 3.35.0 to determine whether allele names for these variants had been assigned or not.

[D] PyPop: http://pypop.org
[E] hlaPoly: http://hlapoly.immunogenomics.org

3. RESULTS

3.1 Hardy-Weinberg Equilibrium (HWE) Test

No statistically significant deviations (all *p*-values < 0.05) from expected HWE proportions were observed for the 11 loci in the AFA and EUR groups (AFA_pop_v7_out.txt and EUR_pop_v7_out.txt). For the HIS group (HIS_pop_v7_out.txt), the *HLA-DPB1* locus displayed a significant overall deviation from HWE expectations (*p*-value = 0.0278), but the other 10 loci did not. The two major contributing *HLA-DPB1* genotypes were *HLA-DPB1*04:02:01:02+HLA-DPB1*04:01:01:01* (10 observed, 19.08 expected, Chen's *p*-value = 0.0095) and *HLA-DPB1*04:02:01:02+HLA-DPB1*04:02:01:02* (21 observed, 13.22 expected, Chen's *p*-value = 0.0017). We also performed HWE analysis of the *HLA-DPA1~HLA-DPB1* haplotypes. The *HLA-DPA1~HLA-DPB1* haplotypes display a deviation from HWE in the direction of homozygote excess (27 observed, 16.17 expected, *p*-value = 0.0071). The three principal contributing *HLA-DPA1~HLA-DPB1* haplotypes are: *HLA-DPA1*01:03:01:02~HLA-DPB1*04:01:01:01+HLA-DPA1*01:03:01:02~HLA-DPB1*04:01:01:01* (7 observed, 3.40 expected, Chen's *p*-value = 0.0230), *HLA-DPA1*01:03:01:05~HLA-DPB1*04:02:01:02+HLA-DPA1*01:03:01:02~HLA-DPB1*04:01:01:01* (5 observed, 12.37 expected, Chen's *p*-value = 0.0136), and *HLA-DPA1*01:03:01:05~HLA-DPB1*04:02:01:02+HLA-DPA1*01:03:01:05~HLA-DPB1*04:02:01:02* (19 observed, 11.29 expected, Chen's *p*-value = 0.0012). For the ASI group (ASI_pop_v7_out.txt), only the *HLA-DQA1* and *HLA-DRB3/4/5* loci display no overall deviation from HWE, and all other 7 loci display significant deviations from HWE. Of the 115 *HLA-DQA1* genotypes, no genotype was observed more than 5 times, or expected more than 3.33 times, resulting in little deviation between observed and expected values. However, *HLA-DQA1* displays many genotypes that are only seen once or twice (70/79), and few common genotypes. This pattern at *HLA-DQA1* suggests admixture from multiple distinct populations.

This pattern for *HLA-DQA1* genotypes is even more pronounced for the 7 loci displaying HWE deviations, suggesting that the families that constitute the ASI group are derived from multiple distinct populations; the US Asian American population is known to have immigrated from many Asian countries and islands [39]. It was not possible to sub-divide the ASI group using information recorded by the contributing laboratories.

3.2 Haplotypes

The haplotype blocks for which frequencies were calculated are shown in Supplemental Table 3. Allele/haplotype counts, allele/haplotype frequencies, and allele/haplotype rankings are summarized in CSV files (Supplemental Materials: Global_"HaplotypeID"_Haplotype/Locus_Summary_2019-03-04.csv). Each file name contains the Haplotype ID shown in Supplemental Table 3. For example, *HLA-C~HLA-B* haplotype frequencies are fond in "Global_CB_Haplotype_Summary_2019-03-04.csv". We were able to build unambiguously phased haplotypes using the strategy described previously [25]. To explore *putative* founder haplotypes, we compared *HLA-A~HLA-C~HLA-B~HLA-DRB1~HLA-DQB1* haplotypes from this family study with those estimated from EM haplotypes in unrelated individuals (Creary et al., manuscript in preparation) [B] for each ethnic group. We selected 13

haplotypes that were identified in both studies (Table 3). The haplotype frequencies for these haplotypes from this study are found in Supplemental Materials: Global_ACBDRB1DQB1_Haplotype_Summary_2019-03-04.csv file.

Tables 4A-D show the top five observed haplotype counts and frequencies for *HLA-C~HLA-B, HLA-DRB3/4/5~HLA-DRB1~HLA-DQA1~HLA-DQB1* and *HLA-DPA1~HLA-DPB1* haplotypes for the AFA, ASI, EUR and HIS groups. The resulting untruncated allele name haplotypes were compared to NMDP g group haplotypes [18]. *HLA-DPA1~HLA-DPB1* haplotypes were not included in the NMDP dataset. The majority of the 17th IHIW haplotypes were identified as high-ranking in the NMDP dataset.

3.3 Measures of LD

LD measures for all possible two-locus phased and unphased haplotypes were calculated for both untruncated and two-field allele names for each broad continental group. The pould output files were included in the Supplemental Materials. For example, for AFA, four pould output files are provided: UnambiguousAllele_Haplotype_AFA_2019-03-04_Phased_LD_results.csv, UnambiguousAllele_Haplotype_AFA_2019-03-04_Unphased_LD_results.csv, TwoFieldAllele_Haplotype_AFA_2019-03-04_Phased_LD_results.csv and TwoFieldAllele_Haplotype_AFA_2019-03-04_Unphased_LD_results.csv, respectively. We plotted *D'* and *Wn* values for 10 two-locus untruncated allele name haplotypes (*HLA-A~HLA-C, HLA-C~HLA-B, HLA-B~HLA-DRB1, HLA-DRB1~HLA-DQA1, HLA-DRB1~HLA-DPA1, HLA-DRB1~HLA-DPB1, HLA-DQA1~HLA-DQB1, HLA-DQB1~HLA-DPA1, HLA-DQB1~HLA-DPB1* and *HLA-DPA1~HLA-DPB1*) for the AFA, ASI, EUR and HIS groups along the genomic distance between each pair of loci based on Human Genome Assembly (hg38) (Figure 1).

The *HLA-DRB1~HLA-DQA1* haplotype showed the highest *D'* and *Wn* values, followed by *HLA-DQA1~HLA-DQB1, HLA-C~HLA-B*, and *HLA-DPA1~HLA-DPB1*. The short genomic interval between the loci in part explains the high LD for these haplotype blocks (Figure 1). Despite a third of the genomic distance between *HLA-DQB1~HLA-DPA1* (414 kb) compared to *HLA-A~HLA-C* (1.33 Mb) and *HLA-B~HLA-DRB1* (1.24 Mb), Figure 1 shows sharp drops in LD for the *HLA-DQB1~HLA-DPA1* haplotype, consistent with previous observations that suggested that this was due to the presence of recombination hotspot in the interval between the DR~DQ loci and the DP loci [40]. It is interesting to note that LD values for *HLA-DRB1~HLA-DPB1* are higher than those of *HLA-DQB1~HLA-DPA1, HLA-DQB1~HLA-DPB1*, and *HLA-DRB1~HLA-DPA1* (Figure 1), although these three two-locus haplotypes are within the *HLA-DRB1~HLA-DPB1* region. Interestingly, the evaluation of the most common 11 loci haplotypes shows that many individual haplotypes show even higher *D'* values between *HLA-A~HLA-C~HLA-B~HLA-DRB3/4/5~HLA-DRB1~HLA-DQA1~HLA-DQB1* haplotype and *HLA-DPA1~HLA-DPB1* haplotype (Supplemental Table 4).

We also extracted 10 measures of LD of two-locus phased haplotypes for untruncated and two-field allele name haplotypes for each population (Supplemental Tables 5ABCD). We found that LD measures for untruncated allele name haplotypes are higher than those of

two-field allele name haplotypes in almost all cases. This indicates that HLA haplotypes have more specific correlations when untruncated allele names are used.

3.4 Crossover

Of the 263 families analyzed, 31 families were trios, 208 were quartets, and 24 included more than two children (Table 5A). It is not possible to detect chromosomal crossover events in trios. It is feasible to identify a crossover event in quartet families, but not possible to determine which of the two parental haplotypes is recombinant. For the families that had more than two children, it is usually possible to identify the parental haplotypes that participated in crossover events where HLA alleles are heterozygous. Of the 232 families that had two or more children, we observed a total of 16 identifiable crossover events in 15 families. Of the 16 crossovers, five were observed between *HLA-A* and *HLA-C*, one between *HLA-C* and *HLA-B*, five between *HLA-B* and *HLA-DRB1*, and five between *HLA-DQB1* and *HLA-DPA1* loci (Table 5A). Of the five *HLA-B~HLA-DRB1* crossovers, two occurred in a family that had four children. In this family, we observed two independent *HLA-B~HLA-DRB1* crossovers: 1) one paternal *HLA-B~HLA-DRB1* recombinant HLA region was transmitted to a child with a non-recombinant maternal HLA region; 2) one maternal *HLA-B~HLA-DRB1* recombinant HLA region was transmitted to a child with a non-recombinant paternal HLA region; 3) two children inherited non-recombinant HLA regions from both parents. The remaining 14 crossover events were identified in 14 unrelated families. Among these 14 crossovers, it was impossible to recognize the original parental haplotypes for three events (one *HLA-A~HLA-C* and two *HLA-DQB1~HLA-DPA1)*, as these families had only two children (Table 5A). We were able to recognize the original parental haplotypes for the remaining 11 events, as these families had more than two children. The *HLA-C~HLA-B* crossover was not identified in randomly selected families, but was found during related bone marrow transplantation donor recruitments. This family was included to investigate this rare crossover event in detail. Tables 5B and 5C shows *HLA-A~HLA-C* and *HLA-C~HLA-B* crossover events, respectively.

3.5 Novel variants

We obtained a total of 1,009 unique variants after removing variants with “N” values and applying our two-step filtering system (See section 2.9). Of these 1,009 unique variants, 437 were in *HLA-DRB* genes; 94 of these were in *HLA-DRB* exon sequences. Of these 94 unique variants for *HLA-DRB* genes, we first carefully reviewed sequence alignment data for 21 variants in *HLA-DRB* exon2 and 3 sequences in the NGS HLA genotyping software, and found that all of these were false. We found that some NGS genotyping software reported multiple consensus sequences that include both correct and incorrect consensus sequences. The software created these variants by combining *HLA-DRB* sequences that were not necessarily in phase. These false variants all originated from incorrectly assembled consensus sequences. This demonstrates the challenges of assembling accurate consensus sequence from mixtures of short sequence reads from the highly complex homologues of the *HLA-DRB* gene family, which includes pseudogenes; when imprecise consensus sequences are built, it is difficult to identify novel variants. It was premature to expect to efficiently identify true novel variants using consensus sequences for the highly complex *HLA-DRB* genes from an extensive list of variants in the 17th IHIW data set. Therefore, we focused on

identifying novel variants for the other seven non-DRB genes. Of 572 non-DRB gene variants, 21 were within coding sequences (CDS). Of these, we confirmed 10 novel CDS variants in eight different HLA alleles; three variants were found in the same HLA allele. Official HLA allele names had already been assigned to six alleles using IPD-IMGT/HLA Database version 3.35.0 (Table 6A). The remaining two alleles have not been reported in IPD-IMGT/HLA Database version 3.35.0. We also confirmed 18 non-CDS variants in nine HLA alleles, and official HLA allele names had been assigned to six alleles using IPD-IMGT/HLA Database version 3.35.0 (Table 6B). After confirming novel variants using our two-step filtering strategy, we refiltered the original list using the "confirmed" novel variants, and rescued additional variants that were not originally captured. For example, we were able to identify the *HLA-DQB1*03:01:01:07* allele for mother and child using the 2-step filtering system shown in Figure 2. Consensus sequences for *HLA-DQB1*03:01:01:07* were correctly built for the mother and child. The consensus sequence for *HLA-DQB1*03:01:01:12* was correctly built for the father, but not for the child. *HLA-DQB1*03:01:01:12* differs from *HLA-DQB1*03:01:01:07* by only one nucleotide in intron 2 (Figure 2B). The HLA genotyping software built a single consensus sequence of *HLA-DQB1*03:01:01:07* for the child, but did not build the second consensus sequence for *HLA-DQB1*03:01:01:12*, because these two sequences were too similar. The initial 2-step filtering system excluded *HLA-DQB1*03:01:01:12* from the output, but we manually rescued this allele after reviewing the consensus sequences of the father, and raw sequence read alignments of a child in the HLA genotyping software. It remains possible that some of the novel alleles may not have been identified using the two-step filtering system if consensus sequences were not built correctly in either child or parents.

4. DISCUSSION

The 17th IHIW was conducted to investigate NGS methods for generating full-length HLA gene sequences that include exons, untranslated regions (UTRs) and introns, and HLA genotypes with minimal ambiguities. We sequenced 11 HLA loci using commercially available NGS HLA genotyping systems in 1,017 subjects from 263 nuclear families and generated untruncated allele name HLA genotypes. Using the NGS HLA genotyping systems, we captured full-length HLA gene sequences for all class I genes, although it was not possible to capture complete full-length HLA genes for many HLA class II genes. HLA haplotypes were inferred based on HLA allele segregation after NGS HLA genotyping.

We observed some significant deviations from HWE for some of the groups studied here. This may have resulted from heterogeneous composition of the continental origin groups. When HWE was evaluated for each broad origin group, the Hispanic American group displayed significant deviation from expected HWE for the *HLA-DPB1* locus. This is a heterogeneous group including individuals of Native American, African, and European ancestries; as *HLA-DPB1*04:02:01:02* is the most common *HLA-DPB1* allele in Hispanic and Native North American groups [41, 42], it may not be unexpected to observe significant excesses of *HLA-DPB1*04:02:01:02* homozygotes, which may have resulted from non-random mating and/or population stratification. For the other HLA loci, alleles are balanced resulting in lower frequencies than that of *HLA-DPB1*04:02:01:02;* this effect may not be seen because the sample size is small.

We observed statistically significant deviations from expected HWE for most of the loci for the Asian American group. In the US, the “Asian” designation can be applied to people who emigrated from Japan, Korea, China, Thailand, Vietnam, Myanmar, The Philippines, Malaysia, Pakistan, India, and other countries. The six largest Asian-American subgroups (Indian, Chinese, Filipino, Japanese, Korean, and Vietnamese) comprise approximately 97% of the Asian American population [39]. This heterogeneous grouping poses health care problems for Asian American groups [39]. As such, this Asian American group is likely highly structured, and HWE should not be expected. Unfortunately, ancestral subgroup information is not available for these subjects. Gragert et al. described dramatic HLA variation between Asian populations, while populations in other broad geographic groups were more similar to those in the same group [18]. Therefore, while the haplotype frequencies for the Asian group may not be reliable, the individual haplotypes presented in this study definitely exist in the general US Asian population. Greater insight into Asian haplotype diversity may be made possible by collecting samples from more specific Asian regions as part of the 18th IHIW.

Building HLA haplotypes from families is especially advantageous for identifying rare haplotypes. For example, we identified a rare haplotype in a Hispanic family – *HLA-DRB4*01:03:01:01/HLA-DRB4*01:03:01:03~HLA-DRB1*04:07:01~HLA-DQA1*04:01:01~HLA-DQB1*04:02:01.* The common haplotypes for *HLA-DRB1*04:07:01* and *HLA-DQB1*04:02:01* are *HLA-DRB4*01:03:01:01/HLA-DRB4*01:03:01:03~HLA-DRB1*04:07:01~HLA-DQA1*03:01:01~HLA-DQB1*03:02:01* and *HLA-DRB1*08:02:01~HLA-DQA1*04:01:01~HLA-DQB1*04:02:01*, respectively. We confirmed that this haplotype was not the result of a *HLA-DRB1~HLA-DQA1* crossover event by reviewing HLA allele segregation. Our family-based approach for inferring HLA haplotype phase allows confident identification of rare haplotypes that exists in the general population.

In addition, we noticed that *HLA-C*04:01:01:01~HLA-B*35:17:01* was the third highest-ranking HIS haplotype in this study, while *HLA-C*04:01g~HLA-B*35:17* ranked 37th in the NMDP dataset. *HLA-C*04:01g~HLA-B*35:01g* was the highest ranking HIS haplotype in the NMDP dataset. Three exon 3 nucleotide differences distinguish *HLA-B*35:01:01:01* from *HLA-B*35:17.* The NMDP Hispanic data is derived from various geographic areas in the US that may include subjects with African, European and Native American ancestry, while the Hispanic subjects in this study are mostly from California, and may have principally Mexican ancestry.

We calculated LD measures for each broad ethnic group (Supplemental Tables 5ABCD). It is fascinating to note that phased untruncated allele name haplotypes show higher LD than phased two-field allele name haplotypes in almost all cases. This demonstrates the presence of more specific correlations when untruncated allele names are used. In other words, non-coding variation may be able to correlate with specific HLA haplotype blocks. For example, we found 33 individuals carrying the *HLA-C*05:01:01:02~HLA-B*44:02:01:01* haplotype, and 12 European and Hispanic individuals carrying the *HLA-C*05:01:01:01~HLA-B*18:01:01:01* haplotype (Table 7). These very different *HLA-B* alleles are in LD with specific non-coding polymorphisms in otherwise identical *HLA-C* alleles. When these

haplotypes are reduced to two-field allele name haplotypes, *HLA-C*05:01* is phased with either *HLA-B*44:02* or *HLA-B*18:01*; the resulting in loss of specificity lowers the apparent LD between these loci. The 4-Mb human MHC region was sequenced from 8 cell lines [43], and more recently 95 human MHC genome sequences have become available [44]. The genome sequences supported the accuracy of some untruncated allele name haplotypes. Cell line "QBL" sequence (GenBank: GL000255.2) contains the complete sequence of the *HLA-C*05:01:01:01~HLA-B*18:01:01:01* haplotype [43]. Recently this specific haplotype was confirmed from the genome sequences of additional 4 cell lines: "DUCAF", "EJ32B", "JVM" and "L081785" [44]. The "SSTO" cell line's genomic sequence (GenBank: GL000256.2) contains complete sequence of *HLA-C*05:01:01:02~HLA-B*44:02:01:01* haplotype. This haplotype was also confirmed from the genome sequences of 4 cell lines: "AWELLS", "EK", "SP0010"and "WT47D" [44]. When we aligned genomic sequence containing *HLA-C*05:01:01:01~HLA-B*18:01:01:01* and *HLA-C*05:01:01:02~HLA-B*44:02:01:01* haplotypes using BLAT software in UCSC Genome Browser [45], there are only 7 SNP mismatches between these haplotypes when the HLA-B gene is excluded. However, there are over 100 mismatches found between the *HLA-B*18:01:01:01* and *HLA-B*44:02:01:01* alleles. It is also interesting to note that *HLA-B*18:01:01:02* forms haplotypes with *HLA-C*07:01:01:01* or *HLA-C*12:03:01:01* (Table 7). The *HLA-C*07:01:01:01~HLA-B*18:01:01:02* haplotype was confirmed from genome sequences of 2 cell lines: "31227ABO" and "BM16" [44]. These examples demonstrate the importance of assigning HLA alleles from full-length HLA gene sequences including UTRs, allowing assignment of specific haplotypes. This strong LD may also be useful for identifying HLA genotyping error, and for identifying matched donors for allogeneic hematopoietic transplantation [46].

Figure 1 indicates pronounced decreases in the values of LD for *HLA-DQB1~HLA-DPA1* haplotypes, consistent with a previously described recombination hotspot in this region [40, 47]. Conservation of specific haplotypes may have resulted from complementary functional selection among *HLA-A, HLA-C, HLA-B* and *HLA-DRB, HLA-DQA1, HLA-DQB1* alleles, while this complementarity may have not extended to *HLA-DP* associations with other HLA loci.

Traditionally, LD measures have been calculated using two-field allele name haplotypes, because it was not possible to determine untruncated allele names using Sanger sequencing or Sequence-Specific Oligonucleotide Probe (SSOP) methods when only limited exons were analyzed [48]. In addition, haplotypes have been estimated using the EM approach. We compared LD measures (*D'* and *Wn*) from phased untruncated allele name haplotypes with those from EM-estimated two-field allele name haplotypes (Figures 1C-F). We recently reported that the EM algorithm underestimates the frequency of rare (n<4) haplotypes, thereby overestimating the LD of 2-locus haplotypes [25]. Although we did not observe significant differences, phased untruncated allele name haplotypes still show higher LD measures than EM-estimated two-field allele name haplotypes in almost all cases, despite the expected overestimation of LD measures of EM-estimated haplotypes.

Conditional asymmetric LD (cALD) measures are valuable for interpreting which of two loci display greater variability in haplotypes. For example, we observed eight different *HLA-*

B alleles (*HLA-B*08:01:01:01, HLA-B*49:01:01, HLA-B*18:01:01:02, HLA-B*41:01:01, HLA-B*08:01:20, HLA-B*18:03, HLA-B*44:02:01:01* and *HLA-B*57:01:01*) in phase with *HLA-C*07:01:01:01* in European Americans (Global_CB_Haplotype_Summary_2018-11-02.csv). The haplotype counts for these 8 haplotypes are 36, 9, 9, 1, 1, 1, 1 and 1, respectively. While *HLA-C*07:01:01:01~HLA-B*08:01:01:01* is the most common *HLA-C*07:01:01:01~HLA-B* haplotype, *HLA-B*08:01:01:01* is only observed in this haplotype, and is not observed in phase with any other *HLA-C* allele. This example explains why the cALD value for *HLA-C* alleles (*Loc1*) on any of the haplotypes conditioned on the *HLA-B* alleles (*Loc2*) is higher ($W_{Loc2/Loc1}$ = 0.843) than that of *HLA-B* alleles on any of the haplotypes conditioned on the *HLA-C* alleles ($W_{Loc1/Loc2}$ = 0.701) in European Americans (Supplemental Table 5C).

We identified 16 detectable crossover events, including one intentionally selected *HLA-C~HLA-B* crossover event (Table 5A). This represents 15 out of 497 (416 + 81) children in 232 families (with at least two children: Table 5A) who had received at least one recombinant HLA haplo type from a parent; approximately 3% (15/497) of children received recombinant HLA haplotypes. This translates into a minimum of 15 crossover events in 994 (832 + 162) meioses (1.5%: Table 5A). Despite the lower LD values of *HLA-DQB1~HLA-DPA1* haplotypes than those of *HLA-A~HLA-C* and *HLA-B~HLA-DRB1* haplotypes, we identified equal numbers of *HLA-A~HLA-C, HLA-B~HLA-DRB1* and *HLA-DQB1~HLA-DPA1* crossover events (5 each). This may be an artifact of the small number of families analyzed; we would need to study larger numbers of families to determine a more precise recombination fraction for HLA haplotypes. It is important to note that this 1.5% rate represents a lower bound on the true recombination rate. Even in families with two or more children, there may be undetectable crossover events within homozygous HLA loci. This limits our ability to recognize crossover events purely on the basis of HLA allele names. Supplemental Table 6 shows parental allele counts for each locus in 232 families that had two or more children, and Supplemental Table 7 provides example instances in which we could have missed crossover events due to homozygosity in 95 of these families. For example, there were 4 parents who were homozygous for *HLA-A~HLA-C~HLA-B* haplotypes and heterozygous for *HLA-DRB1~HLA-DQA1~HLA-DQB1* haplotypes, and 4 parents who were heterozygous for *HLA-A~HLA-C~HLA-B* haplotypes and homozygous for *HLA-DRB1~HLA-DQA1~HLA-DQB1~HLA-DPA1~HLA-DPB1* haplotypes. It is not possible to detect *HLA-B~HLA-DRB1* crossover events in chromosomes inherited from these 8 parents (Supplemental Table 7). There were four subjects (H0000973, H000113D, H02761B9 and H0001150) who were completely homozygous for all loci (Supplemental Table 7). It is not possible to detect any HLA recombination event in chromosomes inherited from these subjects.

We identified 17 novel HLA alleles using hlaPoly to analyze consensus DNA sequence in the context of IPD-IMGT/HLA database version 3.25.0 reference sequence. As of March 2019, 12 of the 17 novel alleles had already been assigned HLA allele name in IPD-IMGT/HLA Database version 3.35.0, but 5 alleles have not yet been named in the database (Table 6). Of 17 novel alleles identified, three include non-synonymous variants. Two alleles differ in exons 3 and 4, respectively, from the closest *HLA-DPA1* alleles that were assigned using IPD-IMGT/HLA Database version 3.25.0 (Table 6A). Those alleles that contain non-

synonymous variants may have clinical implications for transplantation. The *HLA-A*32:106* allele includes a single T->A substitution in exon 1 of *HLA-A*32:01:01* -- Tryptophan (W) to Arginine (R) at residue −2 in the leader peptide of HLA-A [49]. A partial leader peptide (residues −22 to −14 relative to the mature protein) of HLA-A, HLA-B and HLA-C molecules binds to the antigen binding groove of the HLA-E molecule, which serves as a ligand for the inhibitory CD94/NKG2A receptor molecule expressed on natural killer (NK) cells and T cells [50, 51]. Further investigations are required to fully comprehend the biological implications of W to R changes at residue −2 in the HLA-A*32:01:01 leader peptide. For the 5 alleles that include synonymous variants, more research is required to determine any clinical implications. A synonymous nucleotide substitution (GCG -> GCA) in exon 4 of *HLA-A*01:01:01:01* was reported to be responsible for lowering the expression of the resulting *HLA-A*01:01:38L* allele through aberrant splicing [52]. Similarly, the *HLA-A*24:01:03Q* allele includes the same synonymous nucleotide substitution (GCG -> GCA) in exon 4 of *HLA-A*24:01:01:01* as *HLA-A*01:01:38L* [53]. The Questionable allele name suffix was assigned because the protein expression level for *HLA-A*24:01:03Q* allele has not been confirmed. In addition, some intron variants cause splicing variants that result in non-expressed alleles [54, 55], or are associated with transcription [56]. These examples present important questions for the Histocompatibility and Immunogenetics community, and challenge us to systematically investigate the biological functions and/or clinical relevance of each variant that resides within the most polymorphic region in the human genome.

We found that haplotypes can be further refined when novel variants are assigned to specific HLA allele names. For example, the following two haplotypes include *HLA-DQB1*03:01:01:01 – HLA-DRB5*02:02~HLA-DRB1*16:02:01:02~HLA-DQB1*03:01:01:01* and *HLA-DRB3*02:02:01:01~HLA-DRB1*12:01:01:03/HLA-DRB1*12:10~HLA-DQB1*03:01:01:01.* From our novel variant analysis, we found that there are many non-coding variants of *HLA-DQB1*03:01:01:01* that were not recognized in IPD-IMGT/HLA Database version 3.25.0 (Table 6B). When we use the HLA allele names assigned under IPD-IMGT/HLA Database version 3.35.0, these haplotypes can be recognized as two distinct haplotypes that include either *HLA-DQB1*03:01:01:05* or *HLA-DQB1*03:01:01:06 – HLA-DRB5*02:02~HLA-DRB1*16:02:01:02~HLA-DQB1*03:01:01:06* and *HLA-DRB3*02:02:01:01~HLA-DRB1*12:01:01:03/HLA-DRB1*12:10~HLA-DQB1*03:01:01:05.* When untruncated allele names resulting from full-gene sequencing are used, we observe more distinct haplotypes.

Generating accurate and consistent HLA genotypes requires considerable effort in reviewing HLA genotypes, making it a labor-intensive process to recreate a haplotype table on a regular basis. However, it would be beneficial to reanalyze raw sequence data using newer versions of the IPD-IMGT/HLA Database, rebuild haplotypes, and update haplotype frequency tables on a regular basis, for example, as part of every IHIW. We captured HLA genotypes and consensus sequences for this 17th IHIW project, but we could not collect the raw sequence files (e.g., fastq files). However, even if fastq files were available, it would not be possible to generate HLA genotypes using different vendors' HLA genotyping software. We found that HLA genotyping software generates fairly accurate HLA genotype calls, but we detected many DNA sequence assembly errors when we attempted to identify novel

polymorphisms in consensus sequences. Given these shortcomings, it would also be valuable to re-analyze the raw sequence data (fastq files) using a given vendor's improved genotyping software versions. For this reason, it is important to maintain raw sequence data, for these typing reevaluations and haplotype revisions. As we have shown in Figure 2, HLA genotyping software did not generate accurate consensus sequences for individuals who had *HLA-DQB1*03:01:01:07* and *HLA-DQB1*03:01:01:12.* As shown in Supplemental Materials "Global_HLA-DPA1_Locus_Summary_2019-03-04.csv", *HLA-DPA1*01:03:01:01, HLA-DPA1*01:03:01:02, HLA-DPA1*01:03:01:03, HLA-DPA1*01:03:01:04* and *HLA-DPA1*01:03:01:05* are very common *HLA-DPA1* alleles across any ethnic group. *HLA-DPA1* alleles were often reported as homozygous in the original HLA genotyping report, though an individual may possess two distinct alleles. As stated section 2.4, many *HLA-DPA1* allele homozygous calls were determined to be heterozygous during the haplotype building and reviewing process. In this case, the HLA typing software generates only one consensus sequence. It would be very useful to analyze better assembled consensus sequences in the 18th IHIW, as the DNA sequence assembly algorithms in HLA genotyping software are continually being improved.

The haplotype tables that were built in this project can be used as a reference for reviewing and predicting haplotypes from clinical HLA genotyping and a search guide for unrelated hematopoietic transplantation donors. As we move forward to the 18th IHIW, we propose to analyze more families using existing, perhaps improved or newer DNA sequencing technologies, and improved bioinformatics tools. It would be beneficial to re-analyze HLA haplotypes at both the HLA allele name, and also the consensus DNA sequence level.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgements

We thank the Stanford Blood Center for the support and promotion of the 17th IHIW endeavor, and Susan Twietmeyer for their tremendous administrative support of the 17th IHIW efforts. We also thank members of Histocompatibility, Immunogenetics & Disease Profiling Laboratory at the Stanford Blood Center, and A. Karl Kornel (Research Computing, Stanford University) for their technical support. We acknowledge the histocompatibility and immunogenetics community and the International HLA and Immunogenetics Workshop Council for their continued dedication to and support of the International Workshops. The work described here was supported by National Institutes of Health (NIH) National Institute of Allergy and Infectious Disease (NIAID) grant R01AI128775 (SM) and National Institute of Neurological Disorders and Stroke (NINDS) grant and U19NS095774 (GMM, LEC and MFV). The content is solely the responsibility of the authors and does not necessarily reflect the official views of the NIAID, NINDS, NIH or United States Government.

Abbreviations:

CDS	Coding Sequence
CSV	Comma-Separated Values
EM	Expectation-Maximization
GL	Genotype List

HLA	Human Leukocyte Antigen
HML	Histoimmunogenetics Markup Language
HWE	Hardy-Weinberg Equilibrium
IHIW	International HLA and Immunogenetics Workshop
IMGT	ImMunoGeneTics
IPD	ImmunoPolymorphism Database
LD	Linkage Disequilibrium
NGS	Next Generation Sequencing
NMDP	National Marrow Donor Program
SNP	Single Nucleotide Polymorphism
SSOP	Sequence-Specific Oligonucleotide Probe
UTR	Untranslated Region
XML	eXtensible Markup Language

Reference

[1]. Stewart CA, Horton R, Allcock RJ, Ashurst JL, Atrazhev AM, Coggill P et al. : Complete MHC haplotype sequencing for common disease gene mapping. Genome Res 2004;14:1176. [PubMed: 15140828]

[2]. Mungall AJ, Palmer SA, Sims SK, Edwards CA, Ashurst JL, Wilming L et al. : The DNA sequence and analysis of human chromosome 6. Nature 2003;425:805. [PubMed: 14574404]

[3]. Adamek M, Klages C, Bauer M, Kudlek E, Drechsler A, Leuser B et al. : Seven novel HLA alleles reflect different mechanisms involved in the evolution of HLA diversity: description of the new alleles and review of the literature. Hum Immunol 2015;76:30. [PubMed: 25500251]

[4]. Martinez-Laso J, Herraiz MA, Vidart JA, Penaloza J, Barbolla ML, Jurado ML et al. : Polymorphism of the HLA-B*15 group of alleles is generated following 5 lineages of evolution. Hum Immunol 2011;72:412. [PubMed: 21376098]

[5]. von Salome J, Gyllensten U, Bergstrom TF: Full-length sequence analysis of the HLA-DRB1 locus suggests a recent origin of alleles. Immunogenetics 2007;59:261. [PubMed: 17345114]

[6]. Marsh SG, Albert ED, Bodmer WF, Bontrop RE, Dupont B, Erlich HA et al. : Nomenclature for factors of the HLA system, 2010. Tissue Antigens 2010;75:291. [PubMed: 20356336]

[7]. Robinson J, Halliwell JA, Hayhurst JD, Flicek P, Parham P, Marsh SG: The IPD and IMGT/HLA database: allele variant databases. Nucleic Acids Res 2015;43:D423. [PubMed: 25414341]

[8]. Slatkin M: Linkage disequilibrium--understanding the evolutionary past and mapping the medical future. Nat Rev Genet 2008;9:477. [PubMed: 18427557]

[9]. Ceppellini R, Curtoni ES, Mattiuz PL, Miggiano V, Scudeller G, Serra A: Genetics of leukocyte antigens: a family study of segregation and linkage. Copenhagen: Munksgaard; 1967.

[10]. 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;110:4576. [PubMed: 17785583]

[11]. Flomenberg N, Baxter-Lowe LA, Confer D, Fernandez-Vina M, Filipovich A, Horowitz M et al. : Impact of HLA class I and class II high-resolution matching on outcomes of unrelated donor

bone marrow transplantation: HLA-C mismatching is associated with a strong adverse effect on transplantation outcome. Blood 2004;104:1923. [PubMed: 15191952]

[12]. Petersdorf EW, Malkki M, Gooley TA, Martin PJ, Guo Z: MHC haplotype matching for unrelated hematopoietic cell transplantation. PLoS Med 2007;4:e8. [PubMed: 17378697]

[13]. Trowsdale J, Knight JC: Major histocompatibility complex genomics and human disease. Annu Rev Genomics Hum Genet 2013;14:301. [PubMed: 23875801]

[14]. Maiers M, Gragert L, Klitz W: High-resolution HLA alleles and haplotypes in the United States population. Hum Immunol 2007;68:779. [PubMed: 17869653]

[15]. Klitz W, Gragert L, Maiers M, Fernandez-Vina M, Ben-Naeh Y, Benedek G et al. : Genetic differentiation of Jewish populations. Tissue Antigens 2010;76:442. [PubMed: 20860586]

[16]. Schmidt AH, Baier D, Solloch UV, Stahr A, Cereb N, Wassmuth R et al. : Estimation of high-resolution HLA-A, -B, -C, -DRB1 allele and haplotype frequencies based on 8862 German stem cell donors and implications for strategic donor registry planning. Hum Immunol 2009;70:895. [PubMed: 19683023]

[17]. Qin Qin P, Su F, Xiao Yan W, Xing Z, Meng P, Chengya W et al. : Distribution of human leucocyte antigen-A, -B and -DR alleles and haplotypes at high resolution in the population from Jiangsu province of China. Int J Immunogenet 2011;38:475. [PubMed: 21816002]

[18]. 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. Hum Immunol 2013;74:1313. [PubMed: 23806270]

[19]. Dehn J, Setterholm M, Buck K, Kempenich J, Beduhn B, Gragert L et al. : HapLogic: A Predictive Human Leukocyte Antigen-Matching Algorithm to Enhance Rapid Identification of the Optimal Unrelated Hematopoietic Stem Cell Sources for Transplantation. Biol Blood Marrow Transplant 2016;22:2038. [PubMed: 27496216]

[20]. Gragert L, Eapen M, Williams E, Freeman J, Spellman S, Baitty R et al. : HLA match likelihoods for hematopoietic stem-cell grafts in the U.S. registry. N Engl J Med 2014;371:339. [PubMed: 25054717]

[21]. 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 Comput Biol 2015;11:e1004204. [PubMed: 25901749]

[22]. Excoffier L, Slatkin M: Maximum-likelihood estimation of molecular haplotype frequencies in a diploid population. Mol Biol Evol 1995;12:921. [PubMed: 7476138]

[23]. Pappas DJ, Tomich A, Garnier F, Marry E, Gourraud PA: Comparison of high-resolution human leukocyte antigen haplotype frequencies in different ethnic groups: Consequences of sampling fluctuation and haplotype frequency distribution tail truncation. Hum Immunol 2015;76:374. [PubMed: 25637668]

[24]. Askar M, Daghstani J, Thomas D, Leahy N, Dunn P, Claas F et al. : 16(th) IHIW: global distribution of extended HLA haplotypes. Int J Immunogenet 2013;40:31. [PubMed: 23302097]

[25]. Osoegawa K, Mack SJ, Prestegaard M, Fernandez-Vina MA: Tools for Building, Analyzing and Evaluating HLA Haplotypes from Families. Hum Immunol 2019.

[26]. Chang CJ, Osoegawa K, Milius RP, Maiers M, Xiao W, Fernandez-Vina M et al. : Collection and storage of HLA NGS genotyping data for the 17th International HLA and Immunogenetics Workshop. Hum Immunol 2018;79:77. [PubMed: 29247682]

[27]. Milius RP, Heuer M, Valiga D, Doroschak KJ, Kennedy CJ, Bolon YT et al. : Histoimmunogenetics Markup Language 1.0: Reporting next generation sequencing-based HLA and KIR genotyping. Hum Immunol 2015;76:963. [PubMed: 26319908]

[28]. Choo SY: The HLA system: genetics, immunology, clinical testing, and clinical implications. Yonsei Med J 2007;48:11. [PubMed: 17326240]

[29]. Ikeda N, Kojima H, Nishikawa M, Hayashi K, Futagami T, Tsujino T et al. : Determination of HLA-A, -C, -B, -DRB1 allele and haplotype frequency in Japanese population based on family study. Tissue Antigens 2015;85:252. [PubMed: 25789826]

[30]. Milius RP, Mack SJ, Hollenbach JA, Pollack J, Heuer ML, Gragert L et al. : Genotype List String: a grammar for describing HLA and KIR genotyping results in a text string. Tissue Antigens 2013;82:106. [PubMed: 23849068]

[31]. Hedrick PW: Gametic disequilibrium measures: proceed with caution. Genetics 1987;117:331. [PubMed: 3666445]

[32]. Lewontin RC: On measures of gametic disequilibrium. Genetics 1988;120:849. [PubMed: 3224810]

[33]. Single RM, Strayer N, Thomson G, Paunic V, Albrecht M, Maiers M: Asymmetric linkage disequilibrium: Tools for assessing multiallelic LD. Hum Immunol 2016;77:288. [PubMed: 26359129]

[34]. Thomson G, Single RM: Conditional asymmetric linkage disequilibrium (ALD): extending the biallelic r2 measure. Genetics 2014;198:321. [PubMed: 25023400]

[35]. Lancaster AK, Single RM, Solberg OD, Nelson MP, Thomson G: PyPop update--a software pipeline for large-scale multilocus population genomics. Tissue Antigens 2007;69 Suppl 1:192. [PubMed: 17445199]

[36]. Guo SW, Thompson EA: Performing the exact test of Hardy-Weinberg proportion for multiple alleles. Biometrics 1992;48:361. [PubMed: 1637966]

[37]. Slatkin M: An exact test for neutrality based on the Ewens sampling distribution. Genet Res 1994;64:71. [PubMed: 7958833]

[38]. Slatkin M: A correction to the exact test based on the Ewens sampling distribution. Genet Res 1996;68:259. [PubMed: 9062082]

[39]. Holland AT, Palaniappan LP: Problems with the collection and interpretation of Asian-American health data: omission, aggregation, and extrapolation. Ann Epidemiol 2012;22:397. [PubMed: 22625997]

[40]. Cullen M, Perfetto SP, Klitz W, Nelson G, Carrington M: High-resolution patterns of meiotic recombination across the human major histocompatibility complex. Am J Hum Genet 2002;71:759. [PubMed: 12297984]

[41]. Hollenbach JA, Thomson G, Cao K, Fernandez-Vina M, Erlich HA, Bugawan TL et al. : HLA diversity, differentiation, and haplotype evolution in Mesoamerican Natives. Hum Immunol 2001;62:378. [PubMed: 11295471]

[42]. Cerna M, Falco M, Friedman H, Raimondi E, Maccagno A, Fernandez-Vina M et al. : Differences in HLA class II alleles of isolated South American Indian populations from Brazil and Argentina. Hum Immunol 1993;37:213. [PubMed: 8300406]

[43]. Horton R, Gibson R, Coggill P, Miretti M, Allcock RJ, Almeida J et al. : Variation analysis and gene annotation of eight MHC haplotypes: the MHC Haplotype Project. Immunogenetics 2008;60:1. [PubMed: 18193213]

[44]. Norman PJ, Norberg SJ, Guethlein LA, Nemat-Gorgani N, Royce T, Wroblewski EE et al. : Sequences of 95 human MHC haplotypes reveal extreme coding variation in genes other than highly polymorphic HLA class I and II. Genome Res 2017;27:813. [PubMed: 28360230]

[45]. Kent WJ: BLAT--the BLAST-like alignment tool. Genome Res 2002;12:656. [PubMed: 11932250]

[46]. Osoegawa K, Mack SJ, Udell J, Noonan DA, Ozanne S, Trachtenberg E et al. : HLA Haplotype Validator for quality assessments of HLA typing. Hum Immunol 2015.

[47]. Cullen M, Noble J, Erlich H, Thorpe K, Beck S, Klitz W et al. : Characterization of recombination in the HLA class II region. Am J Hum Genet 1997;60:397. [PubMed: 9012413]

[48]. Mack SJ, Tu B, Lazaro A, Yang R, Lancaster AK, Cao K et al. : HLA-A, -B, -C, and -DRB1 allele and haplotype frequencies distinguish Eastern European Americans from the general European American population. Tissue Antigens 2009;73:17. [PubMed: 19000140]

[49]. Thorstenson YR, Creary LE, Huang H, Rozot V, Nguyen TT, Babrzadeh F et al. : Allelic resolution NGS HLA typing of Class I and Class II loci and haplotypes in Cape Town, South Africa. Hum Immunol 2018;79:839. [PubMed: 30240896]

[50]. Braud VM, Allan DS, O'Callaghan CA, Soderstrom K, D'Andrea A, Ogg GS et al. : HLA-E binds to natural killer cell receptors CD94/NKG2A, B and C. Nature 1998;391:795. [PubMed: 9486650]

[51]. Lee N, Llano M, Carretero M, Ishitani A, Navarro F, Lopez-Botet M et al. : HLA-E is a major ligand for the natural killer inhibitory receptor CD94/NKG2A. Proc Natl Acad Sci U S A 1998;95:5199. [PubMed: 9560253]

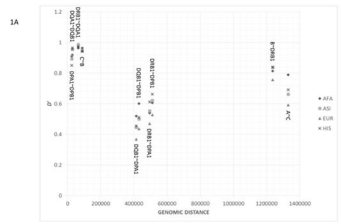
[52]. Dunn PP, Hammond L, Coates E, Street J, Griner L, Darke C: A “silent” nucleotide substitution in exon 4 is responsible for the “alternative expression” of HLA-A*01:01:38L through aberrant splicing. Hum Immunol 2011;72:717. [PubMed: 21741421]

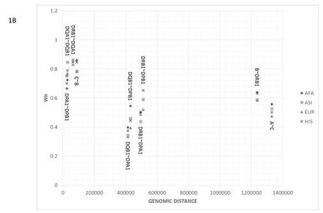
[53]. Lee KW, Shin JY, Lee JY: Expression defect of an HLA-A*24 allele associated with DNA methylation in a normal individual. Tissue Antigens 2003;61:325. [PubMed: 12753672]

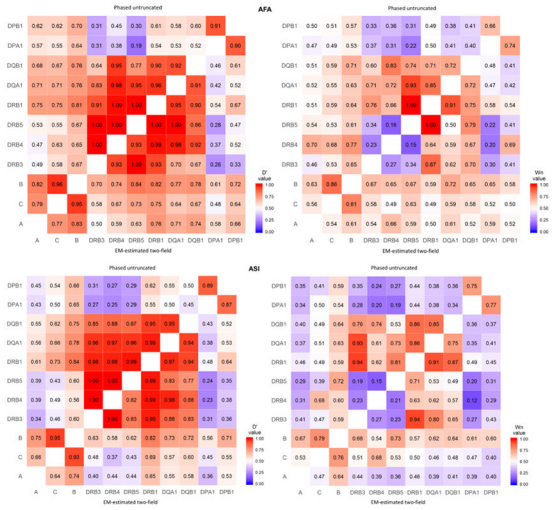
[54]. Sutton VR, Knowles RW: An aberrant DRB4 null gene transcript is found that could encode a novel HLA-DR beta chain. Immunogenetics 1990;31:112. [PubMed: 2303277]

[55]. Shimizu M, Kuroda Y, Nakajima F, Nagai T, Satake M: A novel HLA-C allele, HLA-C*07:02:01:17N, with an alternative splice site. HLA 2018.

[56]. Petersdorf EW, Malkki M, O’HUigin C, Carrington M, Gooley T, Haagenson MD et al. : High HLA-DP Expression and Graft-versus-Host Disease. N Engl J Med 2015;373:599. [PubMed: 26267621]







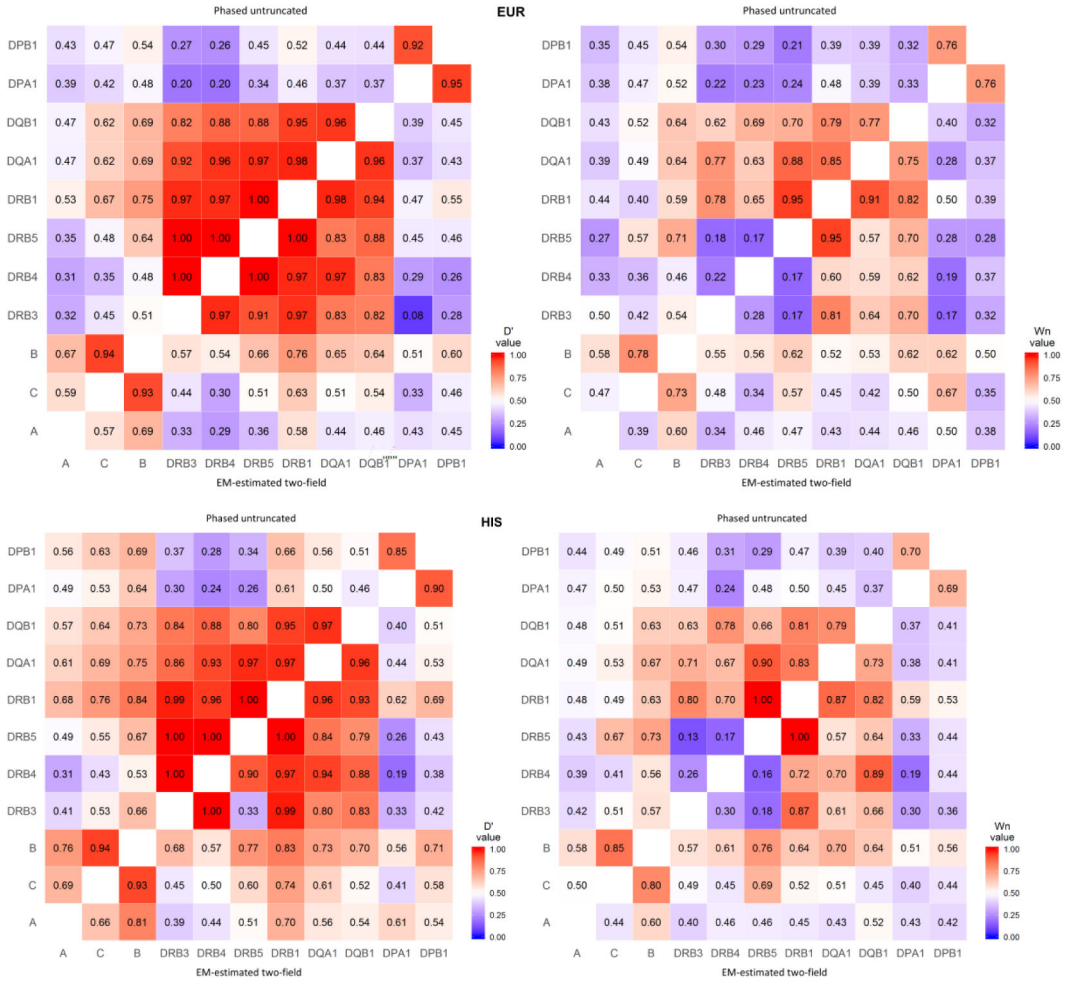


Figure 1.

The graphs A and B show the differences between *D'* (Figure 1A) and *Wn* (Figure 1B) values for the 10 two-locus haplotypes (*HLA-A~HLA-C, HLA-C~HLA-B, HLA-B~HLA-DRB1, HLA-DRB1~HLA-DQA1, HLA-DRB1~HLA-DPA1, HLA-DRB1~HLA-DPB1, HLA-DQA1~HLA-DQB1, HLA-DQB1~HLA-DPA1, HLA-DQB1~HLA-DPB1* and *HLA-DPA1~HLA-DPB1).* The pre-fix "HLA-" was omitted for the label of these haplotypes in the plots. The X-axis represents the genomic distance of the two loci, the Y-axis shows *D'* (A) or *Wn* (B) values. Figures 1C-1F show heatmaps of LD measures [*D'* (left panel) and *Wn* (right panel)] for AFA, ASI, EUR and HIS. The top-half (above the diagonal) represents the heatmap for the phased untruncated allele name haplotypes, and the bottom half represents the heatmap for the EM-estimated two-field allele name haplotypes.

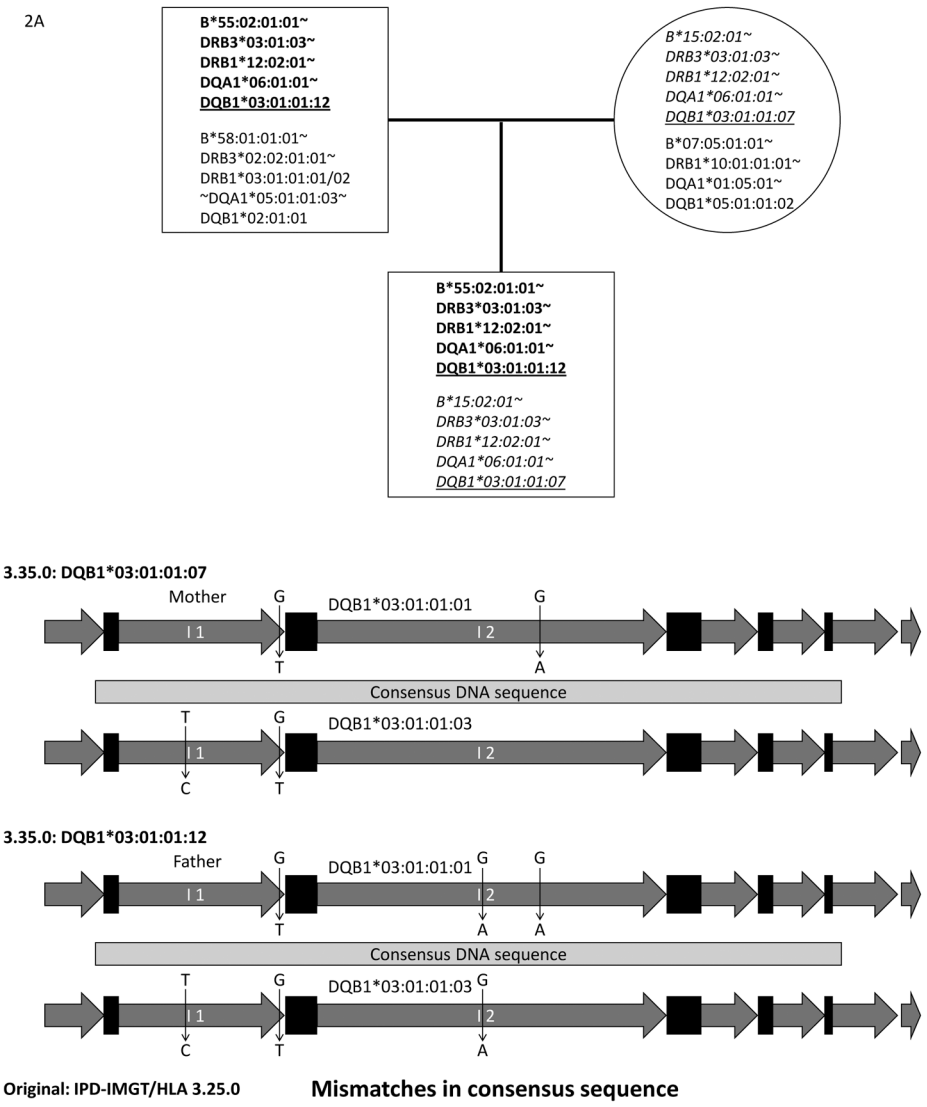


Figure 2.

Figure 2A shows a pedigree for an Asian family: father (top-left square), mother (top-right circle) and child (bottom square). *HLA-B~HLA-DRB3~HLA-DRB1~HLA-DQA1~HLA-DQB1* haplotypes are shown in each square and circle, with the "HLA-" prefix removed from the allele names. The paternal haplotypes with bold letters and maternal haplotypes with italicized letters were transmitted to the child. Originally *HLA-DQB1*03:01:01:03* and *HLA-DQB1*03:01:01:01* were assigned as the closest alleles for father and mother, respectively, using IPD-IMGT/HLA Database version 3.25.0 reference sequences. For the same family, Figure 2B shows that maternal *HLA-DQB1* consensus sequence (Top) and paternal sequence (Bottom) are compared to *HLA-DQB1*03:01:01:01* and *HLA-DQB1*03:01:01:03* DNA sequences. Light gray bars indicate the consensus DNA sequences, black rectangles show exons of *HLA-DQB1*, and dark gray arrows show introns. The novel variants are shown in vertical arrows: nucleotide in reference sequence (top) -> nucleotide in consensus sequence (bottom). Two and three novel variants were identified by hlaPoly from maternal and paternal consensus DNA sequences, respectively. These novel

sequences have been named *HLA-DQB1*03:01:01:07* and *HLA-DQB1*03:01:01:12*, respectively (Table 6B). The applied HLA genotyping software generated only one consensus sequence that corresponds to *HLA-DQB1*03:01:01:07* for the child (Figure 2A), and thus failed to identify *HLA-DQB1*03:01:01:12* by 2-step filtering strategy. *HLA-DQB1*03:01:01:12* was, however, found in father. We confirmed the presence of *HLA-DQB1*03:01:01:12* allele in the child by careful review of the raw sequence alignments in the HLA genotyping software.

Table 1:

Study subjects

Ethnic/racial background	Subject	Parent
African American (AFA)	139	72
Asian American (ASI)	235	115
European American (EUR)	423	210
Hispanic American (HIS)	237	118
OTHER	21	11
Total	1055	526

Table 1 shows the number of subjects and parents that were categorized in each broad ethnic/racial origin group. The number of parents includes inferred parents.

Table 2:

Imputing inferred parental genotypes

A				
Sample	**Relationship**	***HLA-A***	***HLA-B***	***HLA-C***
45C8	mother	*HLA-A*02:01:01:01*+	*HLA-B*40:01:02*+	*HLA-C*03:04:01:01*+
		*HLA-A*02:01:01:01*	*HLA-B*51:01:01:01*	*HLA-C*14:02:01:01*
45C9	child	***HLA-A*01:01:01:01***+	*HLA-B*40:01:02*+	*HLA-C*03:04:01:01*+
		*HLA-A*02:01:01:01*	***HLA-B*51:08:01***	***HLA-C*16:02:01***
45CA	child	*HLA-A*02:01:01:01*+	*HLA-B*51:01:01:01*+	***HLA-C*12:03:01:01***+
		HLA-A*11:01:01:01	***HLA-B*51:01:01:01***	*HLA-C*14:02:01:01*
5EDD	father	***HLA-A*01:01:01:01***+	***HLA-B*51:01:01:01***+	***HLA-C*12:03:01:01***+
	(inferred)	***HLA-A*11:01:01:01***	***HLA-B*51:08:01***	***HLA-C*16:02:01***
B				
Sample	**Relationship**	***HLA-A***	***HLA-B***	***HLA-C***
		*HLA-A*24:02:01:01*+	*HLA-B*35:01:01:01/HLA-B*35:01:01:02*+	*HLA-C*03:03:01:01*+
45BC	mother	*HLA-A*31:01:02:01*	*HLA-B*55:01:01*	*HLA-C*04:01:01:01*
		*HLA-A*31:01:02:01*+	*HLA-B*55:01:01*+	***HLA-C*03:02:02:01***+
45BD	child	***HLA-A*33:03:01***	***HLA-B*58:01:01:01***	*HLA-C*03:03:01:01*
		*HLA-A*24:02:01:01*+	*HLA-B*35:01:01:01/HLA-B*35:01:01:02*+	***HLA-C*03:02:02:01***+
45BE	child	***HLA-A*33:03:01***	***HLA-B*58:01:01:01***	*HLA-C*04:01:01:01*
	father	***HLA-A*33:03:01***+	***HLA-B*58:01:01:01***+	***HLA-C*03:02:02:01***+
5ED9	(inferred)	***HLA-A*NT***	***HLA-B*NT***	***HLA-C*NT***

Table 2A and 2B show two families that lacked paternal HLA genotypes. Both families consist of a mother and two children. Two alleles per class I locus are shown in the tables. The maternal HLA alleles are shown in thin letters, and two alleles were distinguished with single and double underlines. In Table 2A, two different paternal HLA alleles (bold and underlined) per locus are inferred from two children after subtracting maternal HLA alleles (thin letters). It was, therefore, possible to impute complete paternal genotypes (5EDD). Contrarily, for the family in Table 2B, the two children share the same paternal alleles, indicated with bold letters (haplo-identical), although two different maternal HLA alleles were transmitted to these children. Only one paternal allele at each locus could be imputed, because the second paternal was not found in the child genotypes, leaving the second allele unknown (5ED9). In these cases, “NT” (with the corresponding locus prefix, e.g., *HLA-A*NT*) was used to represent the unknown allele.

Table 3:

Common *HLA-A~HLA-C~HLA-B~HLA-DRB1~HLA-DQB1* haplotypes

Haplotype	AFA	ASI	EUR	HIS
*A*01:01:01:01~C*07:01:01:01~B*08:01:01:01~DRB1*03:01:01:01~DQB1*02:01:01*	Y		Y	Y
*A*29:02:01:01~C*16:01:01:01~B*44:03:01:01~DRB1*07:01:01:01~DQB1*02:02:01:01*			Y	Y
*A*03:01:01:01~C*07:02:01:03~B*07:02:01~DRB1*15:01:01:01~DQB1*06:02:01*	Y		Y	
*A*03:01:01:01~C*04:01:01:01~B*35:01:01:02~DRB1*01:01:01~DQB1*05:01:01:03*	Y		Y	
*A*02:01:01:01~C*05:01:01:02~B*44:02:01:01~DRB1*04:01:01:01~DQB1*03:01:01:01*		Y	Y	
*A*11:01:01:01~C*08:01:01~B*15:02:01~DRB1*12:02:01~DQB1*03:01:01:01*		Y		
*A*33:03:01~C*03:02:02:01~B*58:01:01:01~DRB1*03:01:01:01~DQB1*02:01:01*		Y		
*A*30:01:01~C*06:02:01:01~B*13:02:01~DRB1*07:01:01:01~DQB1*02:02:01:01*		Y	Y	Y
*A*33:03:01~C*07:06~B*44:03:02~DRB1*07:01:01:01~DQB1*02:02:01:01*		Y		
*A*33:03:01~C*04:01:01:01~B*53:01:01~DRB1*08:04:01~DQB1*03:19:01*	Y			
*A*24:02:01:01~C*07:02:01:01~B*39:06:02~DRB1*14:06:01~DQB1*03:01:01:01*				Y
*A*02:06:01:01~C*07:02:01:01~B*39:05:01~DRB1*04:07:01~DQB1*03:02:01*				Y
*A*01:01:01:01~C*06:02:01:01~B*57:01:01~DRB1*07:01:01:01~DQB1*03:03:02:01*		Y	Y	Y

The “Haplotype” column shows *HLA-A~HLA-C~HLA-B~HLA-DRB1~HLA-DQB1* haplotypes. The prefix “HLA-“ is omitted from the allele names in the “Haplotype” column. The “Y” indicates that a specific haplotype is found in AFA, ASI, EUR and/or HIS from this study. The following ambiguous alleles are shown using the lowest-digit allele names: *HLA-DRB1*03:01:01:01/HLA-DRB1*03:01:01:02; HLA-DRB1*07:01:01:01/HLA-DRB1*07:01:01:02; HLA-DRB1*15:01:01:01/HLA-DRB1*15:01:01:02/HLA-DRB1*15:01:01:03.*

Table 4:

Haplotype frequencies

A: African American					
Haplotype	**Count**	**Frequency**	**NMDP_Hap**	**NMDP_Freq**	**NMDP_Rank**
*C*04:01:01:01~B*53:01:01*	12	0.085106	*C*04:01g~B*53:01*	0.107896	1
*C*02:10:01:01~B*15:03:01:02*	10	0.070922	*C*02:02g~B*15:03g*	0.060763	2
*C*04:01:01:01~B*35:01:01:02*	7	0.049645	*C*04:01g~B*35:01g*	0.054819	4
*C*06:02:01:01~B*58:02:01*	7	0.049645	*C*06:02g~B*58:02*	0.040085	6
*C*16:01:01:01~B*45:01:01*	5	0.035461	*C*16:01~B*45:01g*	0.037349	7
*C*17:01:01:02~B*42:01:01*	5	0.035461	*C*17:01g~B*42:01*	0.052617	5
*DRB5*01:01:01~DRB1*15:03:01:01~DQA1*01:02:01:01/DQA1*01:02:01:03/ DQA1*01:02:01:05~DQB1*06:02:01*	13	0.093525	*DRB5*01:01~DRB1*15:03~DQB1*06:02*	0.116721	1
*DRB1*01:02:01~DQA1*01:01:02~DQB1*05:01:01:01*	9	0.064748	*DRB1*01:02~DQB1*05:01*	0.040139	6
*DRB3*03:01:01~DRB1*13:02:01~DQA1*01:02:01:04/DQA1*01:02:01:06/ DQA1*01:02:01:07~DQB1*06:09:01*	8	0.057554	*DRB3*03:01~DRB1*13:02~DQB1*06:09*	0.037567	8
*DRB3*02:02:01:01~DRB1*03:01:01:01/ DRB1*03:01:01:02~DQA1*05:01:01:01~DQB1*02:01:01*	7	0.05036	*DRB3*02:02g~DRB1*03:01~DQB1*02:01g*	0.051723	5
*DRB4*01:03:01:01/DRB4*01:03:01:03~DRB1*07:01:01:01/ DRB1*07:01:01:02~DQA1*02:01:01:01/DQA1*02:01:01:02~DQB1*02:02:01:01*	6	0.043165	*DRB4*01:01g~DRB1*07:01~DQB1*02:01g*	0.097462	2
*DPA1*02:02:02~DPB1*01:01:01*	19	0.143939			
*DPA1*03:01~DPB1*105:01*	15	0.113636			
*DPA1*02:01:08~DPB1*01:01:01*	13	0.098485			
*DPA1*01:03:01:02~DPB1*04:01:01:01/DPB1*04:01:01:02*	13	0.098485			
*DPA1*01:03:01:01~DPB1*02:01:02*	10	0.075758			
B: Asian American					
Haplotype	**Count**	**Frequency**	**NMDP_Hap**	**NMDP_Freq**	**NMDP_Rank**
*C*01:02:01~B*46:01:01*	18	0.078603	*C*01:02g~B*46:01g*	0.051956	2
*C*03:02:02:01~B*58:01:01:01*	15	0.065502	*C*03:02g~B*58:01g*	0.053895	1
*C*08:01:01~B*15:02:01*	11	0.048035	*C*08:01g~B*15:02g*	0.041805	3
*C*07:02:01:03~B*07:02:01*	8	0.034934	*C*07:02g~B*07:02g*	0.029234	12
*C*07:06~B*44:03:02*	8	0.034934	*C*07:01g~B*44:03g*	0.038588	6

Hum Immunol. Author manuscript; available in PMC 2020 September 01.

*C*07:02:01:01~B*40:01:02*	8	0.034934	*C*07:02g~B*40:01g*	0.027985	13
*DRB1*08:03:02~DQA1*01:03:01:03/DQA1*01:03:01:04~DQB1*06:01:01*	16	0.069565	*DRB1*08:03~DQB1*06:01*	0.037305	9
*DRB3*03:01:03~DRB1*12:02:01~DQA1*06:01:01~DQB1*03:01:01:01*	14	0.06087	*DRB3*03:01~DRB1*12:02~DQB1*03:01g*	0.078293	1
*DRB4*01:03:01:01/DRB4*01:03:01:03~DRB1*07:01:01:01/ DRB1*07:01:01:02~DQA1*02:01:01:01/DQA1*02:01:01:02~DQB1*02:02:01:01*	14	0.06087	*DRB4*01:01g~DRB1*07:01~DQB1*02:01g*	0.073975	3
*DRB3*02:02:01:02~DRB1*11:01:01:01~DQA1*05:05:01:01/DQA1*05:05:01:02/ DQA1*05:05:01:04~DQB1*03:01:01:03*	12	0.052174	*DRB3*02:02g~DRB1*11:01g~DQB1*03:01g*	0.053415	6
*DRB4*01:03:01:01/ DRB4*01:03:01:03~DRB1*04:03:01~DQA1*03:01:01~DQB1*03:02:01*	10	0.043478	*DRB4*01:01g~DRB1*04:03~DQB1*03:02g*	0.036714	11
*DRB3*02:02:01:01~DRB1*03:01:01:01/ DRB1*03:01:01:02~DQA1*05:01:01:03~DQB1*02:01:01*	10	0.043478	*DRB3*02:02g~DRB1*03:01~DQB1*02:01g*	0.054923	5
*DPA1*02:02:02~DPB1*05:01:01*	62	0.269565			
*DPA1*01:03:01:01~DPB1*02:01:02*	25	0.108696			
*DPA1*01:03:01:02~DPB1*04:01:01:01/DPB1*04:01:01:02*	16	0.069565			
*DPA1*01:03:01:04~DPB1*04:01:01:01/DPB1*04:01:01:02*	15	0.065217			
*DPA1*01:03:01:05~DPB1*04:02:01:02*	12	0.052174			

C: European American

Haplotype	Count	Frequency	NMDP_Hap	NMDP_Freq	NMDP_Rank
*C*07:01:01:01~B*08:01:01:01*	36	0.086957	*C*07:01g~B*08:01g*	0.103941	2
*C*07:02:01:03~B*07:02:01*	29	0.070048	*C*07:02g~B*07:02g*	0.122997g	1
*C*05:01:01:02~B*44:02:01:01*	21	0.050725	*C*05:01g~B*44:02g*	0.070644	3
*C*04:01:01:01~B*35:01:01:02*	18	0.043478	*C*04:01g~B*35:01g*	0.055878	4
*C*03:03:01:01~B*15:01:01:01*	13	0.031401	*C*03:03g~B*15:01g*	0.031774	7
*DRB5*01:01:01~DRB1*15:01:01:01/DRB1*15:01:01:02/ DRB1*15:01:01:03~DQA1*01:02:01:01/DQA1*01:02:01:03/ DQA1*01:02:01:05~DQB1*06:02:01*	39	0.094203	*DRB5*01:01~DRB1*15:01~DQB1*06:02*	0.126887	1
*DRB4*01:01:01:01~DRB1*07:01:01:01/DRB1*07:01:01:02~DQA1*02:01:01:01/ DQA1*02:01:01:02~DQB1*02:02:01:01*	33	0.07971	*DRB4*01:01g~DRB1*07:01~DQB1*02:01g*	0.095955	2
*DRB1*01:01:01~DQA1*01:01:01:02/DQA1*01:01:01:03~DQB1*05:01:01:03*	29	0.070048	*DRB1*01:01~DQB1*05:01*	0.084440	4
*DRB3*01:01:02:01~DRB1*03:01:01:01/ DRB1*03:01:01:02~DQA1*05:01:01:02~DQB1*02:01:01*	26	0.062802	*DRB3*01:01~DRB1*03:01~DQB1*02:01g*	0.094306	3
*DRB4*01:03:01:01/ DRB4*01:03:01:03~DRB1*04:01:01:01~DQA1*03:01:01~DQB1*03:02:01*	17	0.041063	*DRB4*01:01g~DRB1*04:01~DQB1*03:02g*	0.045575	6
*DRB3*02:02:01:02~DRB1*11:01:01:01~DQA1*05:05:01:01/DQA1*05:05:01:02/ DQA1*05:05:01:04~DQB1*03:01:01:03*	17	0.041063	*DRB3*02:02g~DRB1*11:01g~DQB1*03:01g*	0.063037	5

*DPA1*01:03:01:02~DPB1*04:01:01:01/DPB1*04:01:01:02*	76	0.205962			
*DPA1*01:03:01:05~DPB1*04:02:01:02*	56	0.151762			
*DPA1*01:03:01:04~DPB1*04:01:01:01/DPB1*04:01:01:02*	45	0.121951			
*DPA1*01:03:01:01~DPB1*02:01:02*	35	0.094851			
*DPA1*01:03:01:03~DPB1*03:01:01*	27	0.073171			
D: Hispanic American					
Haplotype	**Count**	**Frequency**	**NMDP_Hap**	**NMDP_Freq**	**NMDP_Rank**
*C*07:02:01:03~B*07:02:01*	13	0.055085	*C*07:02g~B*07:02g*	0.056518	2
*C*16:01:01:01~B*44:03:01:01*	10	0.042373	*C*16:01g~B*44:03*	0.038053	6
*C*07:02:01:01~B*39:06:02*	9	0.038136	*C*07:02g~B*39:06*	0.020084	11
*C*04:01:01:01~B*35:17:01*	9	0.038136	*C*04:01g~B*35:17*	0.008341	37
*C*07:02:01:01~B*39:05:01*	8	0.033898	*C*07:02g~B*39:05*	0.025750	8
*C*03:04:01:02~B*40:02:01*	8	0.033898	*C*03:04g~B*40:02g*	0.023331	10
*DRB1*08:02:01~DQA1*04:01:01~DQB1*04:02:01*	27	0.114894	*DRB1*08:02~DQB1*04:02*	0.068817	2
*DRB4*01:03:01:01/ DRB4*01:03:01:03~DRB1*04:07:01~DQA1*03:01:01~DQB1*03:02:01*	20	0.085106	*DRB4*01:01g~DRB1*04:07g~DQB1*03:02g*	0.063367	3
*DRB4*01:03:01:01/ DRB4*01:03:01:03~DRB1*04:04:01~DQA1*03:01:01~DQB1*03:02:01*	15	0.06383	*DRB4*01:01g~DRB1*04:04~DQB1*03:02g*	0.046397	5
*DRB3*01:01:02:01~DRB1*14:06:01~DQA1*05:03~DQB1*03:01:01:01*	10	0.042553	*DRB3*01:01~DRB1*14:06~DQB1*03:01g*	0.027954	6
*DRB4*01:01:01:01~DRB1*07:01:01:01/DRB1*07:01:01:02~DQA1*02:01:01:01/ DQA1*02:01:01:02~DQB1*02:02:01:01*	10	0.042553	*DRB4*01:01g~DRB1*07:01~DQB1*02:01*	0.095357	1
*DPA1*01:03:01:05~DPB1*04:02:01:02*	72	0.306383			
*DPA1*01:03:01:02~DPB1*04:01:01:01/DPB1*04:01:01:02*	39	0.165957			
*DPA1*01:03:01:04~DPB1*04:01:01:01/DPB1*04:01:01:02*	14	0.059574			
*DPA1*01:03:01:01~DPB1*02:01:02*	12	0.051064			
*DPA1*01:03:01:03~DPB1*03:01:01*	8	0.034043			

Table 4 shows haplotype frequencies for three haplotype blocks (*HLA-C~HLA-B, HLA-DRB3/4/5~HLA-DRB1~HLA-DQA1~HLA-DQB1* and *HLA-DPA1~HLA-DPB1*) for African, Asian, European and Hispanic Americans. NMDP_Hap, NMDP_Freq and NMDP_Rank indicates corresponding NMDP g group haplotypes, their frequencies and ranking. The HLA-prefix was removed from allele names.

Table 5:

Crossover

A: Summary

Category	Families	Parents	Children	Meioses	*A~C*	*C~B*	*B~DRB1*	*DQB1~DPA1*
Trio	31	62	31	62	N/A	N/A	N/A	N/A
Quartet	208	416	416	832	1	0	0	2
Quintet or larger	24	48	81	162	4	0	5	3
Excluded	0	0	1	2	0	1	0	0

B: *HLA-A~HLA-C* crossover

Family_ID	Sample_ID	Relation	*HLA-A*	*HLA-C*	*HLA-B*	Rec
040	74A	child	***A*68:02:01:01***	***C*08:02:01:01***	***B*14:02:01:01***	N
			*A*31:01:02:01*	*C*04:01:01:01*	*B*35:17:01*	N
	74B	child	*A*02:06:01:01*	***C*08:02:01:01***	***B*14:02:01:01***	Y
			*A*31:01:02:01*	*C*04:01:01:01*	*B*35:17:01*	N
	748	child	***A*68:02:01:01***	***C*08:02:01:01***	***B*14:02:01:01***	N
			*A*31:01:02:01*	*C*04:01:01:01*	*B*35:17:01*	N
	749	father	***A*68:02:01:01***	***C*08:02:01:01***	***B*14:02:01:01***	N
			*A*02:06:01:01*	*C*15:02:01:01*	*B*40:02:01*	N
	74C	mother	*A*31:01:02:01*	*C*04:01:01:01*	*B*35:17:01*	N
			*A*24:02:01:01*	*C*03:03:01:01*	*B*15:39:01*	N

C: Rare *HLA-C~HLA-B* crossover

Family_ID	Sample_ID	Relation	*HLA-A*	*HLA-C*	*HLA-B*	Rec
077	A4	child	***A*01:01:01:01***	***C*07:01:01:01***	***B*08:01:01:01***	N
			*A*25:01:01*	*C*12:03:01:01*	*B*18:01:01:02*	N
	A3	child	*A*33:01:01*	*C*08:02:01:01*	*B*14:02:01:01*	N
			*A*29:02:01:01*	*C*16:01:01:01*	*B*44:03:01:01*	N
	A2	child	***A*01:01:01:01***	***C*07:01:01:01***	***B*08:01:01:01***	N
			*A*29:02:01:01*	*C*16:01:01:01*	*B*44:03:01:01*	N
	A5	child	*A*33:01:01*	*C*08:02:01:01*	***B*08:01:01:01***	Y
			*A*29:02:01:01*	*C*16:01:01:01*	*B*44:03:01:01*	N
	A9	father	***A*01:01:01:01***	***C*07:01:01:01***	***B*08:01:01:01***	N
			*A*33:01:01*	*C*08:02:01:01*	*B*14:02:01:01*	N
	AA	mother	*A*29:02:01:01*	*C*16:01:01:01*	*B*44:03:01:01*	N
			*A*25:01:01*	*C*12:03:01:01*	*B*18:01:01:02*	N

Table 5A shows breakdown of families included for chromosomal crossover analyses. “Trio”, “Quartet” and “Quintet or larger” represents trio, quartet and quintet or larger families, respectively. We added “Excluded” category to exclude a non-randomly identified *HLA-B~HLA-C* crossover case. “Families”, “Parents” and “Children” columns show the number of families, parents and children in each category, respectively. “Meioses” column shows the number of meiosis in each category. “*A~C*”, “*C~B*”, “*B~DRB1*” and “*DQB1~DPA1*” columns show the number of children identified in each category.

For Tables 5B and 5C, the “N” mark in Column “Rec” indicates no recombination, while the sign “Y” shows recombinant haplotype identified. Family-based haplotypes were built using HaplObserve from three children for family 040 and four children for family 077, respectively. Table 5B shows an *HLA-A~HLA-C* crossover event identified in family 040. Table 5C indicates a rare *HLA-C~HLA-B* crossover event found in family 077. Comparing three independent haplotype constructions from three trios for family 040, and four independent haplotype building from four trios for family 077 identified the recombination events.

Table 6:

Novel variants

A: CDS

Closest allele	Feature name	Ref	Con	Start	End	Effect	IPD-IMGT/HLA 3.35.0	Ethnicity	Parent
*A*32:01:01*	Exon1	T	A	66	67	Non-synonymous TGG (W) -> AGG (R)	*A*32:106*	AFA	1
*DPA1*02:02:01*	Exon 3	G	A	17	18	Non-synonymous GTG (V) -> ATG (M)	*DPA1*02:07:01:01*	ASI EUR HIS	12
		A	G	127	128	Synonymous			
		A	G	208	209	Synonymous			
*DPA1*03:01*	Exon 4	G	C	76	77	Synonymous	*DPA1*03:01:02*	AFA	1
*DPA1*04:01*	Exon 4	G	A	32	33	Non-synonymous GCG (A) -> ACG (T)	*DPA1*04:02*	AFA	2
*DQB1*05:01:01:01*	Exon 1	C	T	80	81	Synonymous	*DQB1*05:01:024:01*	ASI	3
*DQB1*06:02:01*	Exon 3	C	T	238	239	Synonymous	*DQB1*06:02:27*	AFA	1
*DPA1*01:03:01:02*	Exon 4	C	G	76	77	Synonymous	Novel	HIS	1
*DPA1*02:02:02*	Exon 4	C	G	46	47	Synonymous	Novel	ASI	2

B: UTRs and Introns

Closest allele	Feature name	Ref	Con	Start	End	IPD-IMGT/HLA 3.35.0	Ethnicity	Parent
*DQB1*03:01:01:03*	Intron 2	C	T	429	430	*DQB1*03:01:01:04*	Other	1
*DQB1*03:01:01:01*	Intron 3	G	T	479	480	*DQB1*03:01:01:05*	EUR/HIS	6
*DQB1*03:01:01:01*	Intron 2	AAATTTATGATTAATCAATC	-	1430	1450	*DQB1*03:01:01:06*	EUR/HIS	12
*DQB1*03:01:01:01*	Intron 1	G	T	1380	1381	*DQB1*03:01:01:07*	ASI	8
	Intron 2	G	A	1533	1534			
*DQB1*03:01:01:03*	Intron 1	T	C	632	633			
	Intron 1	G	T	1380	1381			
*DQB1*03:01:01:01*	Intron 1	G	T	1380	1381	*DQB1*03:01:01:12*	ASI	3
	Intron 2	G	A	1141	1142			
	Intron 2	G	A	1533	1534			
	Intron 1	T	C	632	633			
*DQB1*03:01:01:03*	Intron 1	G	T	1380	1381			
	Intron 2	G	A	1141	1142			

*DPA1*02:01:01:02*	Intron 1	C	T	601	602	*DPA1*02:01:01:07*	ASI	1
*DQA1*04:01:01*	Intron 1	G	A	1163	1164	Novel	EUR/HIS	27
*DQB1*03:01:01:03*	Intron 2	T	C	1390	1391	Novel	HIS	1
*DQB1*03:01:01:03*	Intron 3	C	A	86	87	Novel	ASI	1
	5’ UTR	C	T	484	485			

Table 6 shows novel variants identified in CDS (A) and UTR/introns (B). The column “Closest allele” shows allele name in IPD-IMGT/HLA Database release version 3.25.0. The novel variants were identified comparing consensus sequences against reference DNA sequences in IPD-IMGT/HLA Database 3.25.0. Prefix HLA- is removed for the allele name. *HLA-DPA1**02:02:01 existed in IPD-IMGT/HLA Database version 3.25.0, but was eliminated after IPD-IMGT/HLA Database release version 3.26.0. The “Feature name” column shows where the novel variants were identified. The “Ref” column indicates reference sequence/nucleotide, and the “Con” column indicates the consensus sequence/nucleotide. The “Start” and “End” columns indicate the nucleotide positions represented using zero-based coordinates for each reference feature. The “Effect” column in Table 6A indicates whether the novel variants are Synonymous or Non-synonymous against the reference allele. Allele names using IPD-IMGT/HLA Database 3.35.0 database corresponding to these variations are shown in the “IPD-IMGT/HLA 3.35.0” column if identified. “Novel” indicates that the variants are not included in IPD-IMGT/HLA Database version 3.35.0. The “Ethnicity” column shows the broad continental groups in which the specific variations were identified, and the “Parent” column shows the number of parents that carried the variation. *HLA-DQB1**03:01:01:07 and *HLA-DQB1**03:01:01:12 were originally assigned as *HLA-DQB1**03:01:01:01 or *HLA-DQB1**03:01:01:03, respectively, because they had equal number of mismatches; they were equally possible closest alleles (Figure 2B).

Table 7:

HLA haplotype blocks with non-coding variations

2-field allele	Untruncated allele name haplotype	Frequency
*C**05:01	*C**05:01:01:**02**~*B**44:02:01:01	0.03167
	*C**05:01:01:**01**~*B**18:01:01:01	0.011516
*C**06:02	*C**06:02:01:**01**~*B**57:01:01	0.017274
	*C**06:02:01:**01**~*B**13:02:01	0.013436
	*C**06:02:01:**01**~*B**37:01:01	0.008637
	*C**06:02:01:**01**~*B**58:02:01	0.007678
	*C**06:02:01:**01**~*B**53:01:01	0.002879
	*C**06:02:01:**02**~*B**50:01:01	0.006718
	*C**06:02:01:**03**~*B**45:01:01	0.004798
*C**07:02	*C**07:02:01:**03**~*B**07:02:01	0.051823
	*C**07:02:01:**01**~*B**39:05:01	0.014395
	*C**07:02:01:**01**~*B**39:06:02	0.014395
	*C**07:02:01:**01**~ *B**39:01:01:03	0.010557
*B**08:01	*C**07:01:01:01~*B**08:01:01:**01**	0.043186
	*C**07:02:01:01~*B**08:01:01:**02**	0.004798
*B**18:01	*C**05:01:01:01~*B**18:01:01:**01**	0.011516
	*C**07:01:01:01~*B**18:01:01:**02**	0.011516
	*C**12:03:01:01~*B**18:01:01:**02**	0.010557
*DQB1**03:01	*DRB1**11:01:01:01~*DQA1**05:05:01:01/*DQA1**05:05:01:02/*DQA1**05:05:01:04~*DQB1**03:01:01:**03**	0.032692
	*DRB1**04:01:01:01~*DQA1**03:03:01:01~*DQB1**03:01:01:**01**	0.018269
*DQB1**03:03	*DRB4**01:03:01:02N~*DRB1**07:01:01:01/*DRB1**07:01:01:02~*DQA1**02:01:01:01/ *DQA1**02:01:01:02~*DQB1**03:03:02:**01**	0.017308
	*DRB4**01:03:01:02~*DRB1**09:01:02~*DQA1**03:02~*DQB1**03:03:02:**02**/*DQB1**03:03:02:**03**	0.016346
	*DRB4**01:03:01:01/*DRB4**01:03:01:03~*DRB1**09:01:02~*DQA1**03:02~*DQB1**03:03:02:**02**/ *DQB1**03:03:02:**03**	0.008654
*DQB1**05:01	*DRB1**01:01:01~*DQA1**01:01:**01**:02/*DQA1**01:01:**01**:03~*DQB1**05:01:01:**03**	0.043269
	*DRB1**01:02:01~*DQA1**01:01:**02**~*DQB1**05:01:01:**01**	0.019231
	*DRB1**10:01:01:01~*DQA1**01:05:01~*DQB1**05:01:01:**02**	0.014423
*DPA1**01:03	*DPA1**01:03:01:**01**~*DPB1**02:01:02/DPB1*02:01:19	0.088057
	*DPA1**01:03:01:**02**~*DPB1**04:01:01:01/DPB1*04:01:01:02	0.146761
	*DPA1**01:03:01:**02**~*DPB1**104:01	0.024291
	*DPA1**01:03:01:**02**~*DPB1**02:01:02/DPB1*02:01:19	0.018219
	*DPA1**01:03:01:**03**~*DPB1**03:01:01	0.045547
	*DPA1**01:03:01:**04**~*DPB1**04:01:01:01/DPB1*04:01:01:02	0.080972
	*DPA1**01:03:01:**04**~*DPB1**02:01:02/DPB1*02:01:19	0.009109
	*DPA1**01:03:01:**05**~*DPB1**04:02:01:02	0.143725

The column “Untruncated allele name haplotype” show *HLA-C~HLA-B, HLA-DRB3/4/5~HLA-DRB1~HLA-DQA1~HLA-DQB1* or *HLA-DPA1~HLA-DPB1* haplotypes. Prefix HLA-is removed for the allele name. The haplotypes includes non-coding variants (indicated in bold underline digits) in otherwise identical HLA alleles as shown in the column “2-field allele”. In “*DQB1**05:01” row, the third-field variants are shown (indicated in bold double-underline digits). The column “frequency” indicates frequencies that combined all ethnic group which is represented as “Global frequency” in Supplemental summary tables.