UC San Diego

UC San Diego Previously Published Works

Title

High-resolution HLA allele and haplotype frequencies in several unrelated populations determined by next generation sequencing: 17th International HLA and Immunogenetics Workshop joint report

Permalink

https://escholarship.org/uc/item/1cc00183

Journal

Human Immunology, 82(7)

ISSN

0198-8859

Authors

Creary, Lisa E
Sacchi, Nicoletta
Mazzocco, Michela
et al.

Publication Date

2021-07-01

DOI

10.1016/j.humimm.2021.04.007

Peer reviewed

eScholarship.org Powered by the California Digital Library
University of California

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

HHS Public Access

Author manuscript

Hum Immunol. Author manuscript; available in PMC 2022 July 01.

Published in final edited form as:

Hum Immunol. 2021 July ; 82(7): 505–522. doi:10.1016/j.humimm.2021.04.007.

High-resolution HLA allele and haplotype frequencies in several unrelated populations determined by next generation

[*]**Corresponding authors:** Lisa E. Creary, Ph.D., Department of Pathology, Stanford University School of Medicine, 3155 Porter Drive, Palo Alto, CA 94304, USA, Tel: +1 650 498 1248, Fax: +1 650 724 0294, lcreary@stanford.edu, Marcelo A. Fernández-Viña, Ph.D., Department of Pathology, Stanford University School of Medicine, Stanford Blood Center, 3155 Porter Drive, Palo Alto, CA 94304, USA, Tel: +1 650 723 7968, Fax: +1 650 724 0294, marcelof@stanford.edu.

Author contributions

LEC was involved in the conception of the study, performed the statistical analyses, interpreted the results, and wrote the manuscript. MFV was involved in the conception of the study, and assisted in the interpretation of the results. The remaining authors performed NGS genotyping assays. All authors read and approved the final manuscript.

Publisher's Disclaimer: This is a PDF file of an unedited manuscript that has been accepted for publication. As a service to our customers we are providing this early version of the manuscript. The manuscript will undergo copyediting, typesetting, and review of the resulting proof before it is published in its final form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

Note added in proof

As an example of convergent evolution we would like to point to groups of alleles that have distinctive features in both population distributions, and associations with alleles of neighboring loci or haplotypes.

In datasets not included in the 17th IHIW we examined the *HLA-B*52:01* 3 and 4-field variants; these are particularly interesting because of the distinct associations with alleles of the neighboring locus *HLA-C* that is downstream of *HLA-B*. The common alleles of the groups differing at the 3rd field *HLA-B*52:01:01* and *HLA-B*52:01:02* differ by a single silent substitution at the 3rd nucleotide of codon 23 (see nucleotide 270 of the genomic alignment, Supplementary Figure 1); the alleles of the *HLA-B*52:01:01* group are common in Asians and Europeans and associate tightly principally with *HLA-C*12:02:02:01* (*HLA-B*52:01:01:01* and *HLA-B*52:01:01:02*) and occasionally with *HLA-C*07* alleles. *HLA-B*52:01:01:03* associates tightly with *HLA-C*07:01* (Barsakis and Fernandez Vina, unpublished observations).

On the other hand the alleles of the *HLA-B*52:01:02* group are found almost only in subjects with African ancestry associate with *HLA-C*16:01:01:01* and in Natives from North, Central and South America. For simplicity we would like to focus on the alleles *HLA-B*52:01:02* in subjects with African ancestry; in these populations alleles, *HLA-B*78:01:01:02* and *HLA-B*51:01:01:01* associate also tightly with *HLA-C*16:01:01:01*. The genomic alignment of *HLA-B*52:01:02:02* compared to *HLA-B*78:01:01:02* (highlighted in yellow or yellow and grey) shows differences only by 10 substitutions in a short segment spanning codons 63–83 (segment spanning nucleotides 388–438 of the genomic alignments), and share otherwise the full gene (including the above mentioned nucleotide 270 plus nucleotide 3702). The distinguishing sequence present in *HLA-B*52:01:02:02* spanning the short segment is highlighted in grey in the figure and is also present in *HLA-B*49:01:01:01* as well as several other Bw4 positive alleles. We postulate that because of the high sequence homology and the associations with one allele of the downstream locus the allele *HLA-B*52:01:02:02* may have been generated through a gene conversion event in which one Bw4 positive allele, for example *HLA-B*49:01:01*, may have donated a short segment that was inserted and replaced the sequence in a backbone recipient haplotype carrying *HLA-B*78:01:01:02* and *HLA-C*16:01:01:01*.

In a similar fashion the allele *HLA-B*52:01:02:03* (highlighted in blue and green) may have arisen from a similar gene conversion even involving the *HLA-B*51:01:01:01* (highlighted in blue) as recipient allele in which an allele like *HLA-B*40:01:02:01* (or many other alleles) may have donated a short sequence spanning codons 63–67 (segment spanning nucleotides 388–401, highlighted in green). With exception of this short segment, alleles *HLA-B*52:01:02:03* and *HLA-B*51:01:01:01* share identical sequences in the full gene and associate with the same HLA-C allele in Africans. The likely origin of *HLA-B*52:01:02:02* and *HLA-B*52:01:02:03* may be explained by a single recombination event. In contrast if one was to postulate their origin from mutations in *HLA-B*52:01:01* alleles (or vice-versa), it would require at least two genetic events to explain the current haplotype constitution. In addition the fact that *HLA-B*52:01:01* alleles are absent in pure African and Native American populations makes it less likely that these alleles are the recipient alleles of mutations, since it would require that the same silent substitution mutation takes place in isolated populations. The possible origin of *HLA-B*52:01:01* alleles coming from Native Americans and Africans is less likely.

Given the fact that these *HLA-B*52:01* alleles are common in different populations, one can suggest selective forces for keeping these alleles at significant frequencies. The fact that the common alleles of *HLA-B*52:01* may have independent origins, convergent evolution may be invoked. In support of selection of alleles with the same protein sequence is that the alleles *HLA-B*52:01:01:01* and *HLA-B*52:01:02:05* may have arisen independently in North/Central American Natives and South American Natives independently. *HLA-B*52:01:02:01* and *HLA-B*52:01:02:05* appear to associate with *HLA-C*03:03* and *HLA-C*15:02:01:01*, respectively.

We identified several alleles with identical protein sequences presenting nucleotide differences resulting from silent substitutions or replacements in non-coding regions (e.g. *HLA-DQB1*05:01* bearing haplotypes). The common occurrence of multiple alleles with the same protein sequences suggests that convergent evolution events may be more prevalent phenomena than initially thought.

Conflict of interest disclosure

The authors declare no competing financial or other interests.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

sequencing: 17th International HLA and Immunogenetics Workshop joint report

Lisa E. Creary[1,2,*], **Nicoletta Sacchi**[3], **Michela Mazzocco**[3], **Gerald P. Morris**[4], **Gonzalo Montero-Martin**[2], **Winnie Chong**[5], **Colin J. Brown**[6,19], **Amalia Dinou**[7], **Catherine Stavropoulos-Giokas**[7], **Clara Gorodezky**[8], **Saranya Narayan**[9], **Srinivasan Periathiruvadi**[9], **Rasmi Thomas**[10], **Dianne De Santis**[11], **Jennifer Pepperall**[12], **Gehad E. ElGhazali**[13], **Zain Al Yafei**[13], **Medhat Askar**[14], **Shweta Tyagi**[15], **Uma Kanga**[15], **Susana R. Marino**[16], **Dolores Planelles**[17,20], **Chia-Jung Chang**[18], **Marcelo A. Fernández-Viña**[1,2]

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

[2]Histocompatibility and Immunogenetics Laboratory, Stanford Blood Center, Palo Alto CA, USA

[3]Italian Bone Marrow Donor Registry Tissue Typing Laboratory, E.O. Ospedali Galliera, Genova, Italy

[4]Department of Pathology, University of California San Diego, La Jolla, CA, USA

[5]Histocompatibility and Immunogenetics Service Development Laboratory, NHS Blood and Transplant, London, UK

[6]Department of Histocompatibility and Immunogenetics, NHS Blood and Transplant, London, UK

[7]Biomedical Research Foundation Academy of Athens, Hellenic Cord Blood Bank, Athens, Greece

[8]Laboratory of Immunology and Immunogenetics, Fundación Comparte Vida, A.C. Mexico City, Mexico

[9]HLA Laboratory, Jeevan Stem Cell Foundation, Chennai, India

[10]US Military HIV Research Program, Walter Reed Army Institute of Research, Silver Spring, USA

[11]Department of Clinical Immunology, PathWest, Perth, Australia

[12]Welsh Transplant and Immunogenetics Laboratory, Welsh Blood Service, Pontyclun, United Kingdom

[13]Sheikh Khalifa Medical City-Union 71, Abu Dhabi and the Department of Immunology, College of Medicine and Health Sciences, UAE University, Al Ain, United Arab Emirates

[14]Department of Pathology and Laboratory Medicine, Baylor University Medical center, Dallas, USA

[15]Department of Transplant Immunology and Immunogenetics, All India Institute of Medical Sciences, New Delhi, India

[16]Department of Pathology, The University of Chicago Medicine, Chicago, IL, USA

[17]Histocompatibility, Centro de Transfusión de la Comunidad Valenciana, Valencia, Spain

[18]Stanford Genome Technology Center, Palo Alto, CA, USA.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

[19]Faculty of Life Sciences and Medicine, King's College London, University of London, England, UK

[20]Grupo Español de Trabajo en Histocompatibilidad e Inmunología del Trasplante (GETHIT), Spanish Society for Immunology, Madrid, Spain

Abstract

The primary goal of the unrelated population HLA diversity (UPHD) component of the 17th International HLA and Immunogenetics Workshop was to characterize HLA alleles at maximum allelic-resolution in worldwide populations and re-evaluate patterns of HLA diversity across populations. The UPHD project included HLA genotype and sequence data, generated by various next-generation sequencing methods, from 4,240 individuals collated from 12 different countries. Population data included well-defined large datasets from the USA and smaller samples from Europe, Australia, and Western Asia. Allele and haplotype frequencies varied across populations from distant geographical regions. HLA genetic diversity estimated at 2- and 4-field allelic resolution revealed that diversity at the majority of loci, particularly for European-descent populations, was lower at the 2-field resolution.

Several common alleles with identical protein sequences differing only by intronic substitutions were found in distinct haplotypes, revealing a more detailed characterization of linkage between variants within the HLA region. The examination of coding and non-coding nucleotide variation revealed many examples in which almost complete biunivocal relations between common alleles at different loci were observed resulting in higher linkage disequilibrium. Our reference data of HLA profiles characterized at maximum resolution from many populations is useful for anthropological studies, unrelated donor searches, transplantation, and disease association studies.

Keywords

Human leukocyte antigen; Next-generation sequencing; International HLA and Immunogenetics Workshop; Allele frequency; Haplotype frequency; HLA diversity; Balancing selection; Population genetics

1. Introduction

The human leukocyte antigen (HLA) genes located at 6p21.3 are amongst the most polymorphic genes in the human genome. To date, over 28,300 HLA alleles and sequences have been deposited in the IPD-IMGT/HLA Database release 3.42.0, and the number of new alleles identified is constantly increasing in an unpredictable manner [1]. This high level of allelic polymorphism, as well as the heterozygosity of HLA molecules, allows the immune system to combat the broad array of pathogens individuals may encounter [2]. On the other hand, the extreme allelic diversity can be detrimental to the success of solid-organ and hematopoietic stem cell transplantation (HSCT) between unrelated individuals; HLA allelic disparities between recipient and donor pairs increases the risk of graft rejection and graft-versus-host disease [3–5]. In addition, HLA genetic variants have been shown to associate with both susceptibility and protection to the development of many human diseases and syndromes, including various autoimmune disorders [6–8],drug-induced hypersensitivities

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

[9,10], and cancer [11]. HLA alleles and zygosity also associate with progression and outcomes of infectious diseases [12,13]

HLA genes typically exhibit distinct allele frequencies between the major population groups originating from different geographical regions of the world [14–19]. It is now generally thought that HLA diversity between populations is maintained by balancing selection [20]. Although, it is speculated that pathogen-driven selection also may have contributed significantly to HLA variation [21]. HLA heterozygosity allows the individual to combat pathogens but diversity influences the ability of the population to combat pathogens. The signature patterns of global HLA allelic variation can be used to infer ancestry of an individual as well as to estimate ancestry proportions in admixed groups [22]. In addition, HLA allele and haplotype distributions together with linkage disequilibrium (LD) between HLA genes, which also show distinct patterns of variation between regions and diverse populations, have been exploited to examine human evolutionary processes such as migration (gene flow) and natural selection [20,23].

The study of unrelated subjects and genetic diversity by next-generation sequencing (NGS) HLA project (NGS of full-length HLA genes) of the 17th International HLA and Immunogenetics Workshop (IHIW) was established to examine worldwide population diversity of HLA characterized at the full-gene level. Anthropological studies based on the global variation of HLA alleles have been a common theme in workshops. However, it wasn't until the Tenth International Histocompatibility Workshop and Conference (IHWC) in 1987 [24,25], which focused on the molecular genetic basis for HLA polymorphism, that HLA population studies became a key fixture of future workshops [23,26,27]. The HLA international workshops provide an ideal opportunity to collate a vast amount of HLA data from unique populations to study human genetic variability. Although, the rapid growth of international migration in recent years has ensured that many diverse populations can be found in just a few countries [28].

During the 11–16th workshops, HLA and other immunogenetic markers were defined using molecular technologies available at that time, such as PCR sequence-specific primer (PCR-SSP) [29], PCR reverse sequence-specific oligonucleotide probe (PCR-rSSOP) [30], PCR Sanger sequencing based typing (PCR-SSBT) [31–33]. These methods typically had limited sequence coverage because clinical laboratories only focused on the interrogation of the most polymorphic exons of the HLA gene that constituted the antigen-recognition site. Since then genomic typing methods have evolved tremendously, giving rise to the now popular next-generation sequencing (NGS) technology.

NGS was an important innovation for the molecular characterization of HLA genes since it allowed the examination of long clonal sequencing reads [34]. NGS approaches that either sequence multiple over-lapping reads encompassing partial gene segments [6,18,35–37], or sequencing multiple long reads that include whole gene amplicons [38,39] provides phase information and diminishes genotype ambiguities. In addition, the other benefits of NGS such as its amenability to unbiased variant discovery, increased accuracy, and efficiency as well as decreased costs, make NGS an attractive option for HLA typing in both the clinical and research settings. Moreover, NGS applied to population studies allows the exploration of

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

HLA genomic population diversity on a more detailed level and provides a more accurate picture of inter- and intra-population differentiation, population admixture, and demographic history.

Here we present the findings of the analyses of HLA variation in 4,240 individuals from twelve unrelated population samples submitted to the unrelated diversity component of the 17th IHIW. We find important differences in the distribution of allele frequencies and LD among different populations. We describe results on detailed haplotype associations observed at the maximum allelic resolution and LD measurements for 2- and 3-locus haplotypes.

2. Materials and Methods

2.1. Participating laboratories

HLA tissue typing laboratories were invited to participate by completing the '*Study of Unrelated subjects by NGS HLA'* questionnaire available on the http://17ihiw.org/ website. The questionnaire gathered general information about the NGS protocol and instrumentation, HLA genes the laboratory could type, software packages used to analyze HLA sequencing data, and the size of the population sample(s) they planned to submit to the project. The '*Study of Unrelated subjects by NGS HLA*' component includes large HLA genotype datasets submitted by unrelated donor bone marrow registries from Argentina (INCUCAI, >36,000 donor genotypes submitted) and Germany (DKMS, > 2 million donor genotypes submitted) as well as smaller datasets from non-registry institutions. The data from registries submitted to this component resulted from partial gene coverage in which introns were not tested. In the present study, we only included samples tested for complete class I sequences and complete or extended coverage class II loci. This study is designated the Unrelated Population HLA diversity project (UPHD), where only the non-registry data is considered. For the UPHD study, recruitment began in late 2015 and by August 2017 a total of 13 accredited histocompatibility testing laboratories from 10 different countries contributed HLA genotype and sequence data generated by various NGS protocols and instrumentations to the project. Table 1 lists the participating laboratories, principal investigators and the number of population samples submitted to the UPHD project.

2.2. Population samples

The total number of population samples contributed to the UPHD project and the HLA loci typed is shown in Table 2. Initially, a total of 30 population samples typed at various loci were contributed by 13 laboratories, we excluded 1 population sample (Maori n = 7) because the sample size was less than 20 individuals. The remaining 29 population datasets were assigned to 12 worldwide ethnic groups, which are diagrammatically displayed in Figure 1; European American, African American, USA Hispanic, Mexican, Spanish, Italian, Greek, European, Arab, Asians- Pacific Islanders (API), Thai, and Indian. The final workshop dataset used for the UPHD study consists of 4,240 healthy individuals. It is important to highlight that not all populations represent the aboriginal origins of the associated geographical regions. For instance, the population groups from the USA represent self-reported ethnicity. The analyses of HLA genotype data with the double-blinded sample IDs

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

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).

2.3. Submission of HLA data into the 17th IHIW database

The 17th IHIW database https://ihiws17.stanford.edu was the central storage point for all workshop data [40]. In the 17th IHIW database, participating laboratories completed a more detailed picture about each sample submitted to the project, including geographical, demographic and linguistic information, as well as detailed parameters of the NGS reagents, instrumentation and software used. Laboratories submitted HLA data by uploading Histoimmunogenetics Markup Language (HML) or eXtensible Markup Language (XML) typing report files to the 17th IHIW database. The IPD-IMGT/HLA Database version 3.25.0 (released July 2016) [1] was used as the only reference source for the 17th IHIW. If the alleles submitted by laboratories were analyzed using other reference database versions, the workshop database converted those allele names to correspond to their most similar lowest digit allele present in version 3.25.0. All HLA genotype data was converted to the genotype list (GL) string format [41,42] in the workshop database in order to standardize input data and facilitate downstream analyses.

2.4. HLA typing

All samples included in the UPHD project were typed for HLA loci using the established NGS methods used routinely by the participating laboratories. Two core laboratories (labcodes ussta1 and utxask) were identified to perform NGS HLA typing on samples submitted by three laboratories; labcodes mexgor, areelg, and gbrpep. For the remaining laboratories, all operations from PCR amplification through to generation of HLA data were performed at the individual participant sites. Various NGS methods were employed, but most of the participating laboratories shared common NGS reagents and protocols, NGS software, and hardware. All groups used NGS commercial reagents and protocols for typing: NGS Engine (GenDx, Utrecht, Netherlands); TypeStream (One Lambda/Thermo Fisher Scientific Inc., CA, USA); Holotype HLA & HLA Twin (Omixon, Budapest, Hungary); MIA FORA NGS (Immucor, Inc., Norcross, GA, USA); TruSight HLA Assign (Illumina Inc. CA, USA). The number of HLA loci typed at the full-gene level for class I loci and full-gene or wide coverage for class II genes ranged from 5 to 11 across the groups. A summary of the NGS software and hardware, as well as HLA loci typed by each group is shown in Table 3.

2.5. Data pre-processing and checking

2.5.1. HLA allele ambiguity assignments—Population allele data were systematically reviewed and pre-processed or ‘cleaned’ to assign common ambiguities found across groups. The majority of the ambiguities were observed in class II alleles which were due to the presence of short tandem repeat (STR) enriched regions located within intronic gene segments of some genes. These low complexity regions, consist of ~1–6 bp nucleotide units that are repeated numerous times and cannot be enumerated with high accuracy by the NGS method. In an attempt to standardize alleles that could not be discriminated due to

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

STRs, alleles were assigned to groups and were given the suffix SG (STR Group) to the lowest numbered allele in that group [6,37]. For instance, *HLA-DRB1*15:01:01:01SG* denotes the *HLA-DRB1*15:01:01:01*/*HLA-DRB1*15:01:01:02*/*HLA-DRB1*15:01:01:03* STR ambiguous group. Characteristics of ambiguities due to STRs and non-sequenced regions identified in the UPHD dataset are shown in Supplementary Table 1.

2.5.2. *HLA-DRB1~HLA-DRB3/4/5* haplotypes—The occurrence of functional *HLA-DRB* (*HLA-DRB3*, *HLA-DRB4*, *HLA-DRB5*) genes on haplotypes varies according to the presence of specific *HLA-DRB1* alleles [43]. *HLA-DRB3* occurs whenever the *HLA-DRB1*03, 11, 12, 13, 14* alleles are present. Similarly, *HLA-DRB4* is present with *HLA-DRB1*04, 07, 09* alleles, whilst *HLA-DRB5* is found on haplotypes bearing *HLA-DRB1*15, 16* alleles. Typically, if the *HLA-DRB1*01, 08, 10* alleles are present all of the functional *HLA-DRB* genes are absent. Each individual may contain zero, one, or two copies of a *HLA-DRB3*, *HLA-DRB4*, *and HLA-DRB5* allele. The absence of *HLA-DRB3/4/5* genes on *HLA-DRB1*01, 08, 10* bearing haplotypes were assigned *HLA-DRB*00:00* and were counted when calculating allele frequencies; if blank alleles are not considered the frequencies of alleles at the *HLA-DRB3/4/5* loci will be over-estimated and deviations from HWE may be observed under the assumption that the presence of a single allele corresponds to a homozygous genotype at this combined locus. *HLA-DRB* haplotypes have been well-defined across many populations, however, there are reported exceptions such as *HLA-DRB1*08~HLA-DRB3* found in Asian populations [44] and *HLA-DRB1*15~HLA-DRB5*absent* haplotypes in African descent populations [18]. In the present study *HLA-DRB* data that deviated from the general pattern of *HLA-DRB3/4/5~HLA-DRB1* associations mostly resulted from allele drop-out or non-typing of a particular *HLA-DRB* gene. For instance, a group may consistently type *HLA-DRB3* and *HLA-DRB4* but not *HLA-DRB5* in this case genotype data were excluded from any subsequent haplotype analyses involving the *HLA-DRB3/4/5* genes. In this study, the *HLA-DRB3*, *HLA-DRB4*, *HLA-DRB5* loci were evaluated as a single combined locus, denoted *HLA-DRB3/4/5*, in order to increase the accuracy of the low frequency haplotypes estimated by the EM programs.

2.6. Population statistical analyses

The *Py*thon for Population Genomics (PyPop) v0.7.0 software package [45] was used to perform the majority of the population analyses. Allele carrier frequencies were determined by direct counting and were calculated as the number of individuals carrying a specific allele (either at the homozygous or heterozygous state) divided by the total number of individuals. Allele frequencies at each locus were tested for deviations from Hardy-Weinberg equilibrium (HWE) proportions using the exact test of Guo and Thompson [46] and by the chi-square test when expected counts were equal or greater than 5.

An implementation of the iterative Expectation-Maximization (EM) algorithm in the PyPop software was used to estimate the frequencies of two loci allelic haplotypes. Linkage disequilibrium (LD) between pairs of alleles at different loci and overall ‘global’ LD between pairs of loci were calculated. Overall LD was computed using two formulae’s normalized to range from −1 to +1; Hendricks D_{ij} ’statistic [47] which weights the

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

contribution to LD of specific allele pairs by the product of their allele frequencies, and Cramér's V statistic [48] also described as W_n that calculates a chi-squared statistic for deviations between observed and expected haplotype frequencies, and conditional asymmetric LD (cALD) measures. cALD between two HLA loci describes the level of allele variation at locus 1 given the presence of specific alleles at locus 2. For example, it is possible that a specific allele at locus 1 is in complete LD with a specific allele at locus 2, such as *HLA-DQA1*05:01:01:02* (at locus 1) associates exclusively with *HLA-DQB1*02:01:01* (at locus 2). However, *HLA-DQB1*02:01:01* also associates with *HLA-DQA1*05:01:01:01* and *HLA-DQA1*05:01:01:03* therefore LD for both of these haplotypes is less than 1. This complimentary pair of cALD measures is denoted $W_{loc1/loc2}$ and $W_{loc2/loc1}$, when both these measures are equal to 1, meaning a complete correlation between both loci, the ALD measure is equal to W_n. cALD was computed using the formula described by Thomson and Single [49] implemented in the Phased Or Unphased Linkage Disequilibrium (POULD) v0.9.1.9000 package.

Extended haplotype (encompassing more than two loci) frequencies were estimated using the R 'haplo.stats' package to run EM in the Bridging Immuno Genomic Data Analysis Workflow Gaps (BIGDAWG) v1.8 package [50]. To reduce the uncertainty of extended haplotypes estimated, particularly those encompassing more than six loci, we limited our analyses to individuals that had complete allelic genotype data.

3. Results

For each ethnic group submitted to the UPHD project, allele frequencies estimated at the maximum allelic resolution, LD measurements between pairs of loci, and 2-, 3-, 4-, 5-, 9-locus haplotype frequencies are available on the http://17ihiw.org website.

3.1. Hardy-Weinberg equilibrium (HWE) tests

The results of HWE exact tests performed at each locus in the 12 population samples are summarized in Table 4. The maximum number of loci tested was 3 class I and 6 class II loci, since *HLA-DRB3*, *HLA-DRB4*, and *HLA-DRB5*, were evaluated as a combined locus denoted *HLA-DRB3/4/5*. All data was tested at the original maximum allelic resolution and 2-field resolution. At the maximum resolution, significant HWE deviations ($P<0.05$) were observed for at least one locus amongst 11 population samples. In the Greek and Thai populations, no significant differences from HWE expectations were detected at any of the genotyped loci.

For the European American sample, all loci tested at maximum resolution exhibited deviations from HWE. Analyses using 2-field allelic resolution data only eliminated HWE deviation at the *HLA-DQA1* locus. Most of the HWE deviations could be attributed to an excess of specific homozygous and heterozygous genotypes at 5 of these loci (*HLA-A*, *HLA-C*, *HLA-DPA1*, *HLA-DPB1*, *HLA-DRB3/4/5*) and mostly an excess of heterozygous genotypes at the other 4 loci (*HLA-B*, *HLA-DQA1*, *HLA-DQB1*, *HLA-DRB1*). It is noteworthy that low frequency or rare alleles were detected in the European American samples that contributed to some of the HWE deviant genotypes. Importantly, according to the allele frequencies database (ww.allelefrequencies.net) and other sources these infrequent

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

or rare alleles have been observed in elevated frequencies in other ethnic groups such as Ashkenazi Jews (*HLA-A*03:02*, *HLA-A*26:08*), and Natives from the American continents (*HLA-B*15:15*, and *HLA-B*39:05:01*) [51]. The presence of these alleles strongly suggests subpopulation structure also known as the Wahlund effect [52], which occurs due to admixture of two or more populations with distinct allele frequencies resulting in overall decreased heterozygosity of the admixed population. It is also possible that genotyping errors may have also contributed to HWE deviations. Another factor to be considered is that the European American or 'USA white' is a loosely based ethnic category defined by the USA 2010 census since it encompasses individuals having origins from vast and distant geographic regions including Europe, North Africa, and the Middle East [53].

Similarly, in the African American population, extreme and moderate deviations from HWE proportions at *HLA-DPA1*, *HLA-DQA1*, and *HLA-DRB3/4/5* were the results of roughly equal proportions of excess observed homozygous and heterozygous genotypes compared to the expected genotypes. For the remaining loci that deviated from HWE a significant excess of heterozygous genotypes were observed. Again, population substructure was evident due to the presence of rare alleles that are common in other populations.

An excess of specific heterozygous genotypes accounted for all of the loci that deviated from HWE proportions in the USA Hispanic (*HLA-A*, *HLA-C*, *HLA-DRB3/4/5*, *HLA-DPA1*, *HLA-DPB1*), Mexican (*HLA-C*, *HLA-DQA1*, *HLA-DQB1*), API (all loci except *HLA-DPA1* and *HLA-DPB1*), and Indian (*HLA-DPA1*, *HLA-DPB1*) population samples. Whereas, in European's significant deviations at *HLA-DPA1*, *HLA-DQB1*, and *HLA-DRB3/4/5* were due to both excess homozygous and heterozygous genotypes, and excess heterozygous genotypes explained the deviation observed at *HLA-DQA1* and *HLA-DRB1*.

In general, for all populations, HWE testing using 2-Field allelic data did little to correct the majority of deviations observed using maximum resolution data. For this reason, other population parameters presented are based on 3/4-Field allelic resolution data.

3.2. Genetic Diversity

The genetic diversity within each population was assessed using the expected heterozygosity index values per locus, the results of which are displayed in Figure 2. Supplementary Table 17 lists the observed and expected heterozygosity values. For all class I loci the average heterozygosity index was highest, at near maximal, for *HLA-B* ranging from 0.94 to 0.97 across all populations; the largest value was observed in USA Hispanics.

Overall, heterozygosity was greater at *HLA-C* (0.88 to 0.94) than *HLA-A* (0.85 to 0.94). From the *HLA-A* heterozygosity values, we see clear differences between populations with the highest values detected in African American and USA Hispanics, intermediate values observed in the API, Spain, and Italy and the lowest values seen in Europeans, European Americans, and the remaining Asian populations. On average, the USA Hispanics exhibited the highest heterozygosity values at class I loci, followed by African Americans, and Arabs. It is noteworthy that the USA Hispanic population represents an amalgam of populations from North, Central and South America. Also, the high level of diversity observed in USA Hispanics is reflective of their multi-ethnic composition which includes contributions from

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

European, Native American, and African populations. Population substructures within this multi-ethnic group may also explain the HWE deviations detected.

It is well known that African populations are highly genetically heterogeneous; this is likely due to both an ancient complex demographic history, and pathogen-driven selection acting in response to unusual pathogens [54,55]. Geographic separation of tribes along with common pathogens also play a role in HLA diversification on the continent. These features, as well as admixture with non-African populations, may account for the relatively high HLA diversity observed in the African American population.

At the class II loci, expected heterozygosity was highest at the *HLA-DRB1* locus; the highest value of 0.95 was observed in USA Hispanics, followed by 0.94 in African Americans and Arabs. The lowest value of 0.90 was observed in the API sample. Due to the fact that only 8 groups typed the *HLA-DRB3/4/5* loci, it was difficult to observe trends in heterozygosity values across populations. Interestingly at the *HLA-DQA1* locus, the lowest value of 0.87 was observed in both the USA Hispanic and Mexican populations, whereas the highest value of 0.92 was found in API's. In the USA Hispanic and Mexican populations the relatively low values are due to predominance of just a few alleles in the relatively small (n ~20) *HLA-DQA1* allele repertoire such as *HLA-DQA1*03:01:01*, *HLA-DQA1*04:01:01*, and *HLA-DQA1*05:051:01:01.* In comparison, the API group consisted of several *HLA-DQA1* alleles of intermediate frequencies, accounting for the overall high heterozygous frequencies. The *HLA-DQB1* heterozygosity values across groups displayed a similar pattern to the *HLA-DQA1* values. At the *HLA-DPA1* gene the lowest value was seen in Mexicans (0.69), due to the high frequency of the *HLA-DPA1*01:03:01:05* (allele frequency, AF = 0.52) and the highest values were seen in the African American (0.88) and the Asian Indian (0.86) groups. A similar trend was observed at the *HLA-DPB1* locus.

In summary, USA Hispanics and African American populations are more genetically diverse at the majority of the HLA loci, than European origin (intermediate levels of diversity) and Asian populations (lowest level) populations.

3.3. Population differences in allele frequencies

The number of unique HLA alleles identified in each population is shown in Table 5. For all populations, the largest number of unique alleles was observed at *HLA-B* and the lowest at *HLA-DPA1*. For populations typed at 4-field resolution, there was a direct correlation with the sample size and the number of unique alleles identified, with European Americans having the largest number of unique alleles and the smallest number found in the Thai group. We observed different levels of variation across populations in regards to HLA class I and class II allele frequencies. In general, class II alleles exhibited lower levels of diversity compared to class I alleles. The African American and USA Hispanic populations had the most variation, at both class I and II loci; the least variation was observed in the Asian populations. Overall, 743 unique alleles were identified in the total 12 populations that characterized alleles at maximum resolution; *HLA-A* (118), *HLA-C* (109), *HLA-B* (197), *HLA-DPA1* (23), *HLA-DPB1* (68), *HLA-DQA1* (43), *HLA-DQB1* (57), *HLA-DPB1* (97), *HLA-DRB3/4/5* (29). Class I and class II allele frequencies estimated per population are presented in Supplementary Table 2.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

3.3.1. Class I allele frequencies

3.3.1.1. HLA-A: *HLA-A*02:01:01:01* was the most common allele identified in the entire cohort, with an overall allele frequency (AF) of 0.233. This allele was the top ranking in 9 populations, with frequencies ranging from over 0.316 in Europeans and 0.127 in USA Hispanics. Three other 4-field variants of *HLA-A*02:01:01* allele were identified, however they occurred at much lower frequencies amongst selected groups, *HLA-02:01:01:03* (European American, AF = 0.0006), *HLA-02:01:01:04* (European American, AF = 0.0002), and *HLA-02:01:01:05* (European American, AF = 0.0008; Europeans AF = 0.0024). In comparison, in the API, Thai, and Indian populations, *HLA-A*11:01:01:01* (API AF = 0.174, Thai AF = 0.286), and *HLA-A*01:01:01:01* (India AF = 0.165) were the most common. Other common alleles identified in populations originating from European regions such as Northern European, South-West Europe (Spain), Southern Europe (Italy), Southern Eastern Europe (Greece), as well as European Americans were *HLA-A*01:01:01:01* (AF = 0.106 – 0.165), *HLA-A*03:01:01:01* (AF = 0.067 – 0.133), *HLA-A*24:02:01:01* (AF = 0.040 – 0.137), and *HLA-A*11:01:01:01* (AF = 0.040 – 0.085). However, the USA Hispanics and Mexican populations also shared some of the common alleles identified in the populations originating from European regions, for instance, allele *HLA-A*24:02:01:01* was the second most common in these groups. In the African American population, allele *HLA-A*30:01:01* was common (AF = 0.084) but was found at lower frequencies (AF = 0.012 – 0.029) in all other groups. *HLA-A*74:01:01* was only detected in both the African (AF = 0.049) and Arab (AF = 0.038) populations. The Arab samples were from the indigenous United Arab Emirates (UAE). The UAE population structure is diverse due to admixture of the indigenous population with immigrants from Yemen, Oman, North Africa, India, Europe, Australia, North America, and Latin America. Another region-specific allele found at an elevated frequency (~0.05) was *HLA-A*02:11:01* found only in Asian Indians (AF = 0.071).

3.3.1.2. HLA-B: Overall, the most common *HLA-B* allele identified was *HLA-B*07:02:01*, occurring at frequencies ranging from 0.041 to 0.127. However, this allele was found at a lower frequency in Arabs (AF = 0.0096) and was not detected in the Thai group. The other common alleles *HLA-B*08:01:01:01*, *HLA-B*44:02:01:01*, *HLA-B*35:01:01:02*, and *HLA-B*15:01:01:01* were found predominantly in individuals of European ancestry, as well as the Mexican and Indian populations. *HLA-B*53:01:01* was the top-ranked allele in African Americans, occurring at frequencies of 0.114, in comparison this allele was observed at either very low frequencies, ≤0.01, or not at all in the other groups. This finding suggests that *HLA-B*53:01:01* can be used as a marker to distinguish African populations from other worldwide populations. Similarly, *HLA-B*39:05:01* was found at high frequencies in Mexicans (AF = 0.111) and USA Hispanics (0.089), but was observed at very low frequencies (<0.0025) in 3 groups (European American, African American, Italian), and not detected in the remaining groups. *HLA-B*39:05:01* is an allele frequently observed in individuals of Native American ancestry. Noteworthy is the allele *HLA-B*39:02:02* which was found solely in the Mexican group at a relatively high frequency of 0.082 but was not detected in USA Hispanics as well as the other groups. The Mexicans in this study correspond to a relatively isolated group and may not be representative of the general Mexican population or Hispanic group [51]. The API and Thai group shared some common

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

HLA-B alleles, however, there were distinct differences in frequencies. For example, the allele frequency of *HLA-B*46:01:01* was 0.070 in the API group, but 0.143 in the Thai group. Also, *HLA-B*40:06:01:01* AF = 0.105% in API, AF = 0.024 in Thai's. In contrast *HLA-B*38:02:01* is lower in APIs (AF = 0.023) but higher in Thai's (AF = 0.071). Two alleles, *HLA-B*15:12* (AF = 0.058), *HLA-B*15:21* (AF = 0.035) were exclusive to the API group. In general populations with the largest number of alleles had a larger number of rare alleles.

3.3.1.3. HLA-C: Globally, *HLA-C*07:01:01:01* was the most frequent allele observed at 0.121 and was detected across all groups with the highest frequency occurring in Europeans (AF = 0.149) and lowest in the Asian and Arab groups, AF = ~0.009 – 0.023. The 4-field alternative, *HLA-C*07:01:01:04*, was only observed in the Indian group, albeit at a very low frequency; 0.0093. Similarly, *HLA-C*07:02:01:03* was also relatively common across all groups (global AF = 0.095), with the exception of the Thai group where it was not observed, and the lowest frequencies found in the African ancestry groups; African Americans (AF = 0.0203), and Arabs (AF = 0.010). In comparison, the 4-field counterpart, *HLA-C*07:02:01:01* was observed at higher frequencies in African Americans (0.046), Arabs (0.125), API (0.163), Thai (0.131), and an extremely elevated frequency of 0.260 in Mexicans. Countries from North and South Europe had lower frequencies of around 0.01.

Three 4-field variants of *HLA-C*04:01:01* were detected in the entire cohort. *HLA-C*04:01:01:01* was the most frequent in the African American group at (AF = 0.193), followed closely by the Mexicans (AF = 0.183), and ~ 0.070 to 0.120 in the remaining groups. The highest frequency of *HLA-C*04:01:01:06* was found in Italians (AF = 0.079), and approximately 0.040 in Spanish, Greeks, and Arabs. Overall, *HLA-C*04:01:01:05* was less common with frequencies ranging from 0.005 to 0.015 across eight groups; this allele was not detected in USA Hispanics, APIs, Thais, and Arabs.

The *HLA-C*06:02* allele group was also relatively common in the entire cohort. *HLA-C*06:02:01:01* was present in all groups; AF ranging from 0.018 in USA Hispanics to 0.170 in Indians. *HLA-C*06:02:01:02* was the top-ranking allele in the Arab group (AF = 0.135), but occurred at either low frequencies across the other groups or were not detected at all. Likewise, allele *HLA-C*06:02:01:03* was either absent or not very common amongst the population samples.

3.3.2. Class II allele frequencies

3.3.2.1. *HLA-DPA1* and *HLA-DPB1*: The five intronic variants of *HLA-DPA1*01:03:01* were the most common, collectively, accounting for 76% of the total *HLA-DPA1* allelic diversity observed across all groups. Overall, *HLA-DPA1*01:03:01:02* was the most common variant across groups (AF = 0.154 – 0.307), with the exception of the following groups: USA Hispanics, where *HLA-DPA1*01:03:01:05*, was highly frequent (AF = 0.40); *HLA-DPA1*01:03:01:01* was the most common in India (AF = 0.202) and APIs (AF = 0.293); *HLA-DPA1*01:03:01:04* Arabs (AF = 0.240); *HLA-DPA1*01:03:01:05* in Mexicans (AF = 0.524).

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

The *HLA-DPA1*02:01:01:01* and *HLA-DPA1*02:01:01:02* occurred at relatively high frequencies globally, 0.050 and 0.065 respectively. Strikingly, *HLA-DPA1*02:01:01:01* was most frequent in African Americans, USA Hispanics, Europe, Spanish (AF = ~0.08 – 0.112), whereas *HLA-DPA1*02:01:01:02* was common amongst Mexicans, APIs, and Arabs (AF = 0.05 – 0.135).

The most frequent *HLA-DPB1* allele detected was the 4-field ambiguous pair *HLA-DPB1*04:01:01:01/HLA-DPB1*04:01:01:02*, representing over 0.32 in countries from Europe, 0.375 in Indians and 0.356 in Arabs. In African Americans *HLA-DPB1*01:01:01* was most common (AF = 0.237), whereas *HLA-DPB1*04:02:01:02* ranked number 1 in USA Hispanics (AF = 0.366) and Mexicans (AF = 0.510). *HLA-DPB1*04:01:01:01/HLA-DPB1*04:01:01:02* and *HLA-DPB1*02:01:02/HLA-DPB1*02:01:19* was equally common in the API group (AF = 0.174). Also *HLA-DPB1*03:01:01* was relatively common across groups, ~0.02– 0.08. *HLA-DPB1*18:01* was found at a relatively high frequency in African Americans (AF = 0.074), observed on one occasion in the USA Hispanic and Arab populations, but absent in the other groups; this allele occurs at ~ 5% in West, Central, and South African populations (www.allelefrequencies.net).

3.3.2.2. *HLA-DQA1* and *HLA-DQB1*: For *HLA-DQA1*, the ambiguous intronic SG-level alleles were very common and accounted for 57.5% of the total allele frequency.

The *HLA-DQA1*01:02:01:01SG* allele was found in all groups, at AF around 0.04 to 0.208. Whereas the *HLA-DQA1*01:02:01:04SG* and *HLA-DQA1*01:02:01:02* intronic counterparts were less common. Similarly, *HLA-DQA1*01:03:01:02SG* (AF = 0.053) was more common than *HLA-DQA1*01:03:01:01* (AF = 0.013) and *HLA-DQA1*01:03:01:03SG* (AF = 0.0006). *HLA-DQA1*02:01:01:01SG* was common in European Americans (AF = 0.126), Europeans (AF = 0.170), Spanish (AF = 0.162), Indians (AF = 0.277), and Arabs (AF = 0.144). Whereas, *HLA-DQA1*01:02:01:01SG* was most common in African Americans (AF = 0.208), *HLA-DQA1*05:05:01:01SG* in Greeks (AF = 0.254) and Arabs (AF = 0.164). Whilst *HLA-DQA1*03:01:01* was most frequent in USA Hispanics (AF = 0.31) and Mexicans (AF = 0.274).

Of the total 57 *HLA-DQB1* alleles characterized globally *HLA-DQB1*06:02:01* was the most common occurring at a frequency of 11.1%. This allele was also the most frequent in European Americans (AF = 0.123), African Americans (AF = 0.219), and Europeans (AF = 0.144).

Globally *HLA-DQB1*02:01:01* was the second most frequent allele (AF = 0.1047) closely followed by *HLA-DQB1*03:01:01:03* (AF = 0.1029). The latter allele was prevalent in the Southern European countries; Italy AF = 0.236 and Greece AF = 0.2477. In comparison, the 4-field alternative *HLA-DQB1*03:01:01:01* was found at elevated frequencies in the API and Thai populations (0.214). Whereas, *HLA-DQB1*03:01:01:02* was more prevalent in Arab's (AF = 0.135).

*HLA-DQB1*06:01:01* was most frequent in API and India, AF of 0.131 and 0.156 respectively. *HLA-DQB1*03:19:01* was the second most frequent in African Americans at

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

0.093, and was detected at a moderate frequency in Arabs (AF = 0.039), and very low frequencies in European Americans (AF = 0.002), USA Hispanics (AF = 0.018), Spanish (AF = 0.0074), Mexicans (AF = 0.005), and Italians (AF = 0.005). *HLA-DQB1*03:19:01* is a well-known HLA allele of African ancestry, suggesting African gene flow in the aforementioned populations.

3.3.2.3. *HLA-DRB1, HLA-DRB3, HLA-DRB4, and HLA-DRB5*: At the *HLA-DRB1* locus, overall alleles *HLA-DRB1*07:01:01:01SG*, *HLA-DRB1*03:01:01:01SG*, and *HLA-DRB1*15:01:01:01SG* were the most common occurring at allele frequencies above 0.10. These alleles were relatively common among all groups from Europe, Thai, India, and Arab. Whereas, *HLA-DRB1*15:03:01:01SG* is common in African Americans at 0.149, and were found on single occasions in European Americans, Mexicans, and Arabs. *HLA-DRB1*04:07:01* was common to USA Hispanics (AF = 0.116) and Mexicans (0.149), as well as *HLA-DRB1*14:06:01*; USA Hispanics, 0.036 and Mexican 0.063 but was found at lower frequencies in other groups. *HLA-DRB1*14:05:01* was exclusive to the API group, AF = 0.047.

There are eight different alleles in the *HLA-DRB1*08* family, and they all exhibit varying frequencies across the major population groups. For instance, *HLA-DRB1*08:01:01* is present at ~ 0.02 in European Americans, Europeans, Spanish, and Italians. Whilst *HLA-DRB1*08:04:01* was most common in African Americans (AF = 0.063) and Arabs (AF = 0.019). *HLA-DRB1*08:02:01* was frequent in Mexicans (AF = 0.130) and USA Hispanics (AF = 0.063). In Asians, *HLA-DRB1*08:03:02* was most moderately common; API (AF = 0.035), Thai (AF = 0.024) but was either absent or found at lower frequencies in the other groups.

At the *HLA-DRB3* locus, the most frequent allele, globally, was *HLA-DRB3*02:02:01:02* (0.140), the 4-field alternative form *HLA-DRB3*02:02:01:01* occurred less frequently at 0.063. At *HLA-DRB4*, the *HLA-DRB4*01:03:01:01/HLA-DRB4*01:03:01:03* ambiguous allele pair was the most frequent across all groups at ~0.113 to 0.275. The most frequent *HLA-DRB5* allele detected was *HLA-DRB5*01:01:01* (global AF = 0.127). Due to the limited allelic diversity at the *HLA-DRB3*, *HLA-DRB4*, *HLA-DRB5* loci, there were no obvious distribution patterns observed amongst the populations.

3.4. Population differences in the distribution of Linkage disequilibrium and 2-locus haplotypes

3.4.1 Global (locus-level) linkage disequilibrium—The results of global LD tests to assess the strength of association between neighboring and non-neighboring loci pairs are summarized in Supplementary Table 3. Due to deviation from HWE, the LD and haplotype data presented in this manuscript should be used with caution. Three measures of LD were used for alleles characterized at maximum and 2-field resolution; D', W_n, $W_{loc1/loc2}$, $W_{loc2/loc1}$. The standard D' and W_n values range from 0 to 1, 0 refers to linkage equilibrium, and positive values indicate association; values >0.8 suggests a very strong association. There is a strong negative relationship between the measure of LD, as well as the recombination fraction, and the physical distance separating the loci on the chromosome;

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

generally longer inter-marker distances are associated with smaller LD values. LD plots displaying *D*' and W_n measurements for class I and class II haplotypes are depicted in Figures 4 and 5 respectively.

Overall, strong associations (*D*' values greater than 0.5) were observed for all loci pairs except for non-neighboring loci pairs that included either the *HLA-A* or *HLA-DPB1* loci that showed weak LD. However, unexpectedly the USA Hispanic group exhibited very strong associations for *HLA-A~HLA-B* (D'= 0.90) and *HLA-A~HLA-C* (D'= 0.84) haplotypes. Due to the well-characterized recombination hotspot that exists between the *HLA-DP* and *HLA-DQ* loci, LD is typically weak (*D*' ~0.3) between *HLA-DP* and other class II genes.

We have conducted an additional analyses shown in Supplementary Table 19 that examines *HLA-A* to *HLA-DQB1* block associations with *HLA-DPB1* alleles; in this table it can be observed that some haplotypes extend to *HLA-DPB1* with statistically significant chi-squared values however, the *D*' values are relatively low. In general *HLA-A~HLA-DQB1* blocks show no significant associations with *HLA-DPB1* alleles resulting in an overall weak association with the centromeric more distant loci.

Across the majority of populations, which typed the *HLA-DQA1* gene, the strength of association was greatest for the *HLA-DRB1~HLA-DQA1* haplotype, with *D*' ranging from 0.930 in African Americans to maximal LD observed in Arabs. At the *HLA-C~HLA-B* haplotype *D*' was highest in USA Hispanics (D' = 0.983), API (D' = 0.972), and the Thai (D' = 0.978) groups, whilst in the Italian and Greek groups association was strongest for the *HLA-DRB1~HLA-DQB1* and *HLA-DQA1~HLA-DQB1* haplotypes, respectively. These results reveal that there was no obvious pattern between the strongest LD measures for particular HLA haplotypes and the geographical regions the samples originated from. However, we observed, that African Americans tended to have lower LD values for the majority of the haplotypes compared to the other populations, reflective of the higher gene diversity found in African descent populations.

In general, quantitative estimates of LD values were greater at 4-field than 2-field, irrespective of the distance between the loci, which reflects the greater allelic diversity and the specific allelic haplotypic associations observed for some allele combinations at 4-field. For example, in European Americans *HLA-C*08:02:01:01* associates strongly with *HLA-B*14:02:01:01* (D' = 0.98), and *HLA-B*14:02:01:02* (D' = 1), whereas *HLA-C*08:02:01:02* associates strongly with *HLA-B*14:01:01* (D' = 0.98) and weakly with *HLA-B*14:02:01:01* (D' = 0.12). When these alleles are reduced to 2-fields, *HLA-C*08:02* associates with *HLA-B*14:01* (D' = 0.98) or *HLA-B*14:02* (D' = 0.99) leading to lower global LD values at 2-field (D' = 0.93) compared to 4-field (D' = 0.94). Conversely, for common 4-field alleles that associate with many alleles with the same 2-field group global LD values are higher at 2-field compared to 4-field allele data. For instance, the *HLA-DPA1~HLA-DPB1* haplotypes consist of predominantly *HLA-DPB1*04:01:01:01* that associates tightly with specific low diversity *HLA-DPA1* alleles such as *HLA-DPA1*01:03:01:02* (D' = 0.80) and *HLA-DPA1*01:03:01:04* (D' = 0.83). Collapsing these

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

allele names to 2-field results in an equal ratio between *HLA-DPA1* and *HLA-DPB1* alleles leading to an overall higher LD measure; *HLA-DPA1*01:03~HLA-DPB1*04:01*, D' = 0.91.

cALD measures allowed for a more detailed analysis of LD for locus pairs. Noteworthy is the complementary pair of cALD values for the *HLA-C~HLA-B* haplotype ($W_{HLA-C/HLA-B}$ and $W_{HLA-B/HLA-C}$) observed across all population groups. The $W_{HLA-C/HLA-B}$ measures were larger than $W_{HLA-B/HLA-C}$ indicating more diversity of *HLA-B* alleles compared to *HLA-C* alleles. Similar cALD patterns were observed for haplotypes *HLA-A~HLA-B* and *HLA-A~HLA-C*. At the *HLA-DQA1~HLA-DQB1* haplotype, $W_{HLA-DQB1/HLA-DQA1}$ measures were marginally larger than $W_{HLA-DQA1/HLA-DQB1}$ suggesting more allelic diversity of *HLA-DQB1* than *HLA-DQA1*. At the *HLA-DPA1~HLA-DPB1* haplotype, the complementary cALD values tended to be similar in European Americans, and Europeans, but $W_{HLA-DPA1/HLA-DPB1}$ was moderately more elevated than $W_{HLA-DPB1/HLA-DPA1}$ in African Americans, USA Hispanics, Mexicans, and Spain. These findings tie into the allele distributions observed in the populations.

3.4.2 Allelic haplotype linkage disequilibrium—Supplementary Tables 4 to 10 shows 2-locus haplotypes (*HLA-A~HLA-B*, *HLA-A~HLA-C*, *HLA-C~HLA-B*, *HLA-DRB1~HLA-DQA1*, *HLA-DRB1~HLA-DQB1*, *HLA-DQA1~HLA-DQB1*, *HLA-DPA1~HLA-DPB1*) that occurred at least twice for various loci pairs. Distinctive 4-field allele associations were observed across all populations, some were shared across all groups whereas others were ethnic/region-specific.

The *HLA-C*07:02:01:03~HLA-B*07:02:01* alleles were either in complete LD or near-complete LD, all at relatively high haplotype frequencies (HF~ 8 – 12%), in populations originating from Europe and USA Hispanics. However this haplotype was also detected in African Americans, Mexicans, APIs, and Indians all with maximal LD but at lower frequencies. In comparison, the alternative 4/3-field *HLA-C*07~HLA-B*07* haplotype, *HLA-C*07:02:01:01~HLA-B*07:02:01* was more prevalent and LD was stronger in African Americans (HF = 0.034, *D*' = 0.73) than European Americans (HF = 0.038, *D*' = 0.16); this haplotype was not observed in other groups. Other notable haplotypes common and with significant LD in populations with European ancestry were *HLA-C*07:01:01:01~HLA-B*08:01:01:01* (global HF = 0.077) and *HLA-C*05:01:01:02~HLA-B*44:02:01:01* (global HF = 0.048). Haplotype *HLA-C*03:04:02~HLA-B*15:10:01* was ethnic-specific and found only in African Americans (HF = 0.019, *D*' = 0.83, rank 12). A similar example was observed in the API and Thai groups, where haplotype *HLA-C*01:02:01~HLA-B*46:01:01* was specific to these groups at maximal LD and occurred at relatively high frequencies of 0.070 and 0.143, respectively.

Due to the lower allelic diversity at the *HLA-DQ* loci common haplotypes were shared amongst populations, with high LD. Such as haplotypes *HLA-DQA1*01:02:01:01SG~HLA-DQB1*06:02:01* (global HF = 0.118, present in all populations except India); *HLA-DQA1*02:01:01:01SG ~HLA-DQB1*02:02:01:01* (global HF = 0.093, all populations); *HLA-DQA1*03:01:01~HLA-DQB1*03:02:01* (global HF = 0.089, all populations, except India, with the lowest HF seen in African Americans 0.025 highest in European Americans (HF = 0.095). Four haplotypes bearing *HLA-*

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

*DQB1*03:19:01* were observed in African Americans. *HLA-DQA1*05:05:01:01SG~HLA-DQB1*03:19:01* was the most frequent in African Americans (HF = 0.041, *D*' = 3.9) and Arab's (HF = 0.039, *D*' = 1). This haplotype was observed at lower frequencies in European Americans (HF = 0.001, *D*' = 0.66), and Spanish individuals (HF = 0.006, *D*' = 0.71), evident of African gene flow in these two latter populations. Noteworthy, is haplotype *HLA-DQA1*04:01:02:02~HLA-DQB1*03:19:01* that was only observed in African Americans at a moderate frequency (HF = 0.040, *D*' = 0.77).

Interesting allelic haplotypes were observed at the *HLA-DP* loci. As alluded to earlier, 4-field intronic variants of *HLA-DPA1*01:03:01* are relatively common across all groups, and they tend to associate with specific *HLA-DPB1* alleles with strong LD. Such as *HLA-DPA1*01:03:01:01~HLA-DPB1*02:01:02/HLA-DPB1*02:01:19*; *HLA-DPA1*01:03:01:02~HLA-DPB1*04:01:01:01/02*; *HLA-DPA1*01:03:01:03~HLA-DPB1*03:01:01*; *HLA-DPA1*01:03:01:04~HLA-DPB1*04:01:01:01/02*; *HLA-DPA1*01:03:01:05~HLA-DPB1*04:02:01:02* are frequent in European Americans, African Americans, USA Hispanic, Mexicans, and Spaniards. However we also observed ethnic specific associations in non-European groups such as *HLA-DPA1*01:03:01:02~HLA-DPB1*18:01*, *HLA-DPA1*02:01:08~HLA-DPB1*01:01:01*, *HLA-DPA1*03:01~HLA-DPB1*105:01* observed more frequently (HF = 0.07 – 0.12, *D*'>0.8) in African Americans. In APIs haplotypes *HLA-DPA1*02:02:05~HLA-DPB1*05:01:01* (HF = 0.049, *D*' = 1), *HLA-DPA1*02:02:02~HLA-DPB1*135:01* (HF = 0.037, *D*' = 1), *HLA-DPA1*02:01:01:02~HLA-DPB1*09:01:01* (HF = 0.024, *D*' = 0.5), and *HLA-DPA1*02:02:02~HLA-DPB1*02:02* (HF = 0.024, *D*' = 1), were common. Whilst in Indians, haplotypes *HLA-DPA1*02:01:01:01/02~HLA-DPB1*26: 01:02*, and *HLA-DPA1*01:03:03~HLA-DPB1*04:01:01:01/02*, *HLA-DPA1*02:01:02~HLA-DPA1*17:01* appeared to be specific to this group. These haplotypes are prime examples demonstrating the power of NGS to define ethnic specific groups.

3.5. Extended haplotypes comprising of 4–11 loci

Extended haplotypes, that is haplotypes composed of more than 2 loci are detailed in Supplementary Tables 11 to 16 corresponding to *A~C~B~DRB1~DQB1*, *A~C~B~DRB3/4/5~DRB1~DQB1*, *A~C~B~DRB1~DQB1~DPB1*, *DRB3/4/5~DRB1~DQA1~DQB1*, *A~C~B~DRB3/4/5~DRB1~DQA1~DQB1~DPA1~DPB1*, and *A~C~B~DRB3/4/5~DRB1~DQA1~DQB1* haplotypes respectively. Haplotype counts of 3 estimated by the EM algorithm are generally considered reliable. However, haplotypes counted twice may not be highly accurate but may be informative for populations with small sample sizes. Although haplotypes where n = 2 have been checked for known allele associations we advise the user to err on the side of caution when interpreting or using these haplotypes. Complete extended haplotype datasets, which includes haplotypes with counts less than 2, are available from the 17th IHIW website upon request.

3.5.1. Four locus haplotypes

Silent mutations in *HLA-DQ* and *HLA-DRB3* loci show mutational events that led to present-day haplotype diversity of the *HLA-DRB1*03:01:01:01* group. In total nine

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

haplotypes all bearing *HLA-DRB1*03:01:01:01SG* were estimated (Supplementary Table 14). Of this total, three haplotypes all carried *HLA-DRB3*01:01:02:01* but two haplotypes differed at the 4-field for *HLA-DQA1*05:01:01*, but both carried *HLA-DQB1*02:01:01*, and one haplotype carried *HLA-DQA1*01:02:01:01SG*. The rarer haplotype *HLA-DRB3*01:01:02:01~HLA-DRB1*03:01:01:01SG~HLA-DQA1*01:02:01:01SG~HLA-DQB1*06:02:01*, was detected only in European Americans. Three haplotypes all carried *HLA-DQB1*02:01:01* and *HLA-DRB3*02:02:01:01* but each carried different intronic *HLA-DQA1*05:01* variants; *HLA-DQA1*05:01:01:01*, *HLA-DQA1*05:01:01:02*, *HLA-DQA1*05:01:01:03*. The *HLA-DQA1*05:01:01:01* was the most common (global HF = 0.016) across all groups except for USA Hispanics and Arabs where they are absent. These haplotypes frequently exist on extended haplotypes encompassing *HLA-C*05:01:01:01* and *HLA-B*18:01:01:01* alleles. In comparison two haplotypes all had *HLA-DRB3*02:02:01:02* and *HLA-DQB1*02:01:01* but either carried *HLA-DQA1*05:01:01:01* or *HLA-DQA1*05:01:01:02*. We see a similar pattern of high-level diversity for the *HLA-DRB1*01* bearing haplotypes, where both antigenic and intronic differences at *HLA-DQA1* and *HLA-DQB1* account for distinct haplotypes.

It is noteworthy that there are different eight variations of the *HLA-DRB1*07:01:01:01SG~HLA-DQA1*02:01:01:01SG* haplotype each bearing different *HLA-DQB1* and *HLA-DRB4* alleles. The *HLA-DRB4~HLA-DRB1*07:01:01:01SG* may bear *HLA-DQA1* alleles *HLA-DQA1*02:01:01* (most common across all groups) or *HLA-DQA1*03:03:01:01* (relatively infrequent and only found in African Americans), and *HLA-DQB1* alleles *HLA-DQB1*02:02:01:01*, *HLA-DQB1*02:02:01:02*, *HLA-DQB1*03:01:01:03* and *HLA-DQB1*03:03:02:01. The HLA-C*16:01:01:01~HLA-B*44:03:01:01* block commonly exists on the *HLA-DRB4*01:01:01:01~HLA-DRB1*07:01:01:01SG~HLA-DQA1*02:01:01:01SG~ HLA-DQB1*02:02:01:01* haplotype. *HLA-DRB4*01:03:01:01~HLA-DRB1*07:01:01:01SG~ HLA-DQA1*02:01:01:01SG~HLA-DQB1*02:02:01:01* extended haplotypes bear either *HLA-C*06:02:01:01~HLA-B*13:02:01*, *HLA-C*06:02:01:02~HLA-B*50:01:01*, or *HLA-C*16:01:01:01~HLA-B*44:03:01:01*. Also *HLA-DRB4*01:03:01:02N~HLA-DRB1*07:01:01:01SG~HLA-DQA1*02:01:01:01SG~HLA-DQB1*03:03:02:01* exists on haplotypes bearing *HLA-C*06:02:01:01~HLA-B*57:01:01*.

Asians have distinctive *HLA-DRB1*15* haplotype blocks compared to other groups. They typically carry *HLA-DRB5*01:02~HLA-DRB1*15:02:01:01* with different *HLA-DQ* alleles differing at the antigenic level; *HLA-DQA1*01:03:01:01*, *HLA-DQA1*01:02:01:01SG*, *HLA-DQA1*01:01:01:02SG*, *HLA-DQB1*06:01:01*, *HLA-DQB1*05:02:01*, and *HLA-DQB1*05:01:01:01*.

3.5.2. Nine locus haplotypes based on eleven locus genotyping data

We estimated extended haplotypes both in the presence (Supplementary Table 15) and the absence of *HLA-DP* loci (Supplementary Table 16) due to the weaker association of *HLA-DP* alleles with alleles at other HLA loci. Overall, the most frequent extended haplotypes predicted by the EM-algorithm reflect the allele distribution found in individual groups and are essentially composed of the most frequent 2-locus haplotypes. Extended ancestral

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

haplotypes mediated by strong LD between distant class I and class II alleles have been well-described in the literature [55–57]. The most common 9 loci haplotype was *HLA-A*01:01:01~HLA-C*07:01:01:01~B*08:01:01:01~HLA-DRB3*01:01:02:01~HLA-DRB1*03:01:01:01SG~HLA-DQA1*05:01:01:02~HLA-DQB1*02:01:01~HLA-DPA1*01:03:01:02~HLA-DPB1*04:01:01:01/HLA-DPB1*04:01:01:02* (global HF = 0.021) which corresponds to the most frequent ancestral haplotype (AH8.1) observed in ~5–12% of Northern European populations [56,57]. Interestingly, this haplotype was the most common in European Americans (HF = 0.028) but not in the European population where it was ranked number 3 and haplotype *HLA-A*03:01:01~HLA-C*07:02:01:03~HLA-B*07:02:01~HLA-DRB5*01:01:01~HLA-DRB1*15:01:01:01SG~HLA-DQA1*01:02:01:01SG~HLA-DQB1*06:02:01~HLA-DPA1*01:03:01:02~HLA-DPB1*04:01:01:01/HLA-DPB1*04:01:01:02* was the most frequent (HF = 0.024); the low resolution equivalent of this haplotype was the 4th most frequent in a Norwegian population. The discrepancy in ranking may be due to the small sample size of the Europeans relative to the European Americans.

In African Americans, the most prevalent haplotype was *HLA-A*68:01:01:02~HLA-C*06:02:01:01~HLA-B*58:02:01~HLA-DRB3*02:02:01:02~HLA-DRB1*12:01:01:03~HLA-DQA1*01:05:01~HLA-DQB1*05:01:01:02~HLA-DPA1*01:03:01:02~HLA-DPB1*18:01* (HF = 0.011); this haplotype was not observed in any other population. The haplotype consists of alleles that are prevalent to Africans such as *HLA-B*58:02:01* and *HLA-DPB1*18:01*, also the 2-locus haplotype *HLA-DQA1*01:05:01~HLA-DQB1*05:01:01:02* is most frequent in African groups.

Although 9 loci were estimated in all populations, which typed these loci, the numbers were generally too small in other groups to make definite conclusions, for example in USA Hispanics the two most frequent haplotypes were found at only 4 copies corresponding to a frequency of 0.042.

4. Discussion

In this report, we detailed the major findings of the collaborative efforts initiated under the auspices of the 17th IHIW unrelated HLA diversity component. This study involved many laboratories worldwide and focused on the collection of diverse populations, for detailed characterization of HLA allelic variants determined using high-resolution NGS methods. The HLA typing data was then used to infer haplotype frequencies, and together with allele frequency data were utilized to conduct a thorough analysis of HLA diversity in worldwide populations. In addition, we quantitated LD between loci and alleles in order to gain fresh insights into the demographic history of populations. Due to the power of NGS, the unrelated population data submitted to the 17th IHIW has extended and provided new HLA population data compared to previous workshops [23,26,27]. In this study we showed that there are distinct differences in 4-field allele, haplotype, and LD patterns in populations originating from different geographical regions of the world.

In regards to allele frequencies, although a relatively high number of class I and class II alleles were shared among populations they differed in their frequencies. For instance, *HLA-*

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

*A*02:01:01:01* was found in all groups, but on average was found at almost twice the frequency in Europeans (32%) compared to African Americans (15%) and USA Hispanics (13%), and the lowest frequencies were found in groups from Asia. Another example is given by *HLA-B*08:01:01:01* which was moderately high in populations of Eurasian ancestry (~10%), lower in African Americans (3.5%), and even lower in the Asian groups (~1%). Although admixture was evident in the USA populations unique alleles were also observed in these groups. These trends are also observed for some alleles at the class II loci, such as *HLA-DRB1*12:01:01:03* which is relatively common in Eurasian ancestry groups, in contrast its 4-field counterpart *HLA-DRB1*12:01:01:01* is more prevalent in Asia [36]. Alleles that were detected exclusively in a single population were also observed. We can refer to such alleles as 'region specific' eloquently described by Meyer and colleagues [23], since they are present in populations derived from distinct geographical regions. In our study, region specific alleles (RSA) accounted for a remarkable 42% of the total unique alleles identified, and the majority of these alleles were found at relatively low frequencies, and tended to be less abundant at the least polymorphic loci, in agreement with the Meyer study. Overall, the largest proportion of RSA's was found at *HLA-DQB1* (51%) and lowest found at *HLA-DPA1* (4.4%). As expected the largest population examined, that is European Americans had the most RSA and rare alleles; rare alleles are indicative of a rapidly expanding population. A prime example is demonstrated by *HLA-A*02:60:01* and *HLA-B*07:12* which were found exclusively, both at 0.13% allele frequency, in African Americans. It is noteworthy that *HLA-B*07:12* has also been reported in Ashkenazi, Ethiopia, Libya, and Morocco Jews albeit at lower frequencies (~0.001%) according to www.allelefrequencies.net. Due to the slightly higher frequency and the fact that the African American individual who carried *HLA-B*07:12* also carried *HLA-C*15:05:02*, which is a well-known to be frequent in African populations, we speculate that *HLA-B*07:12* is an African allele and was detected in the Jewish groups due to African gene flow mediated by migration. Surprisingly, an RSA *HLA-A*02:11:01* was found at a relatively high frequency in the Indian cohort (AF = 7.1%), this allele is in significant strong LD with *HLA-B*40:06:01:02*; *HLA-B*40:06:01:02* occurs at very low frequencies in Eurasian population but is elevated in Indians at 8.6%. The majority of individuals in the Indian sample were from Southern Indian Andhra Pradesh population, and recently *HLA-A*02:11:01* has recently been reported in this population [58]. Although, in this study we define *HLA-A*02:11:01* as region specific this allele has also been observed in aboriginal Indians from South America at low frequencies [59]. The limited sample size of some of the populations included in this study has reduced the number of RSA and low-frequency alleles detected.

We observed several population specific alleles that differed at the 4-field resolution. For the *HLA-A*24:02:01* group, 7 intronic variants were identified in the entire sample, however *HLA-A*24:02:01:07* was found solely in African Americans at a low frequency (AF = 0.13%), whilst *HLA-A*24:02:01:06* was observed in USA Hispanics (AF = 1.8%); *HLA-A*24:02:01:01* was detected in all other groups at low or intermediate frequency. Similarly, at the *HLA-30:02:01* group three 4-field variants were found in our dataset, but only *HLA-A*30:02:01:03* was detected in African Americans at 1.1%. These findings illustrate two things; (i) the power of NGS for capturing variants in intron and untranslated regions resulting in more fine-scale discrimination between populations, (ii) the different natural

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

selection pressures acting on intronic segments in alleles from distant geographical regions. Since our data is fairly current, it may be too preliminary to speculate why the specific 4-field alleles examples described above are region specific. However, it is well known that introns may harbor functional polymorphisms that can directly influence gene expression, affect splicing, and may confer risk or protection to disease [6,60]; specific 4-field variants may have developed due to the different pathogen environments.

The large proportion of RSA identified in this dataset is likely to have influenced the deviations from HWE found at several loci in many populations. In addition, HWE deviations can also be explained by recent genetic admixture, inaccurate population sampling, genotyping errors, and natural selection in favor of an excess of HLA heterozygotes which could confer resistance to pathogens and cancers. For this reason, the population samples out of HWE may not be considered as true representatives of the population, and the data should be used with caution.

This study also identified several common alleles with identical protein sequences differing only by intronic substitutions that were found in distinct haplotypes suggesting that their origins were distant or independent. This observation together with the previously well-described protein sequence variability suggests that both diversification and convergent evolution have been at play in shaping the present-day variation of the HLA region.

The variation in allele frequencies leads to diverse haplotype frequencies among populations, and in turn haplotype variation gives rise to a complex pattern of LD across the MHC. LD across the MHC is known to be strong and extensive, creating extended ancestral haplotypes across the entire region. A comparison of LD distribution across populations revealed some interesting observations. Due to the relatively high level of genetic heterogeneity and ancient history of African populations, LD in African descent populations is expected to be less than observed in European as well as other populations and also less structured [54,61]. This means that LD from diverse groups usually cannot be extrapolated to other populations, highlighting the importance of characterizing LD patterns in different populations. As expected, we observed lower levels of LD in the African American population compared to others. Although Africans Americans are often used as a proxy for an African group, several genetic studies have shown that they harbor ancestry from many parts of West Africa as well as 10–25% of European admixture depending on the geographical location in the US. Unfortunately Africa is not well represented in this study, the Arab group from North Africa (the United Emirates of Arab) is too small and is known to be highly admixed with European, African, and indigenous groups to make any concrete interpretations about short and long range LD patterns. In order to better capture HLA diversity on a more detailed level in the African continent, the next 18th Workshop should focus on recruitment of population samples from several widespread African countries.

The higher levels of LD observed across the European descent groups, is indicative of low levels of recombination. Also the long-range LD with high allele frequency, observed with the so-called 8.1 or Cox haplotype, is indicative of recent positive selection, because if an allele has preserved ancestral relationships with neighboring alleles, this implies that the population is relatively young.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

Conclusion

The introduction of next-generation sequencing (NGS) based genotyping technologies created valuable opportunities for collaboration and learning for the HLA and Immunogenetics (H&I) community. With a clear goal of advancing the fields of H&I research through the application of NGS technologies, the 17th IHIW created the foundation and structure for a centralized database system the tools developed for the examination of HLA diversity in multiple world populations. The different HLA profiles in unique populations provide valuable information for studying the molecular evolution of HLA and also for association analysis with immune and autoimmune diseases. Our analyses was somewhat hindered by the relatively small sample size for some groups such as Asians and Thais. As a consequence allele and haplotype frequencies may not reflect those observed in larger or entire populations. Larger samples sizes are required for adequate analyses particularly for the more heterogenetic groups such as USA Hispanics, African Americans and Arabs. The 17th UPHD study highlights the importance of high resolution typing, including intronic and flanking segments for analyzing population differences. Our data provides valuable 2-locus haplotype maximum resolution data that are useful and more informative than allele frequencies to determine population differences. Our study presents analyses of associations of alleles at the under-studied *HLA-DRB3*, *HLA-DRB4*, *HLA-DRB5*, *HLA-DQA1*, and *HLA-DPA1* loci with alleles from flanking loci and their presence in extended haplotypes and corresponding blocks. In spite of the sample size limitations of this study and the lower abundance of non-admixed populations, the results provided by this study illustrate the even larger haplotypic diversity defined by variations in the non-coding regions. These observations may lend information leading to plan and conduct new studies focused in further understanding the evolution of the human MHC. In addition, the information provided here may help with the design of novel disease association studies in selected populations that may allow for finer mapping of the disease determining HLA associated factors. In this regard the rich variation of associations between alleles at *HLA-DRB1* and *HLA-DRB5* loci in some Asian populations allows us to speculate that the primary and secondary roles of *HLA-DRB1* and *HLA-DRB5* alleles may be dissected in diseases associated with haplotypes bearing alleles at the *HLA-DRB5* locus. The present study identified fully extended haplotypes that included alleles at loci not evaluated extensively in the past. These findings taken together with the inclusion of distinct populations may provide significant information about the chances of finding closely HLA matched and/or antibody compatible unrelated donors for sensitized HSCT patients with unique, understudied or admixed ancestry. With the addition of new reference sequences and alleles to the IMGT/HLA Database, we anticipate that more accurate allele calling will be achieved. It should be noted that after the 17th IHIW the sequence intervals considered for naming of HLA alleles has been expanded for the class I loci resulting in the addition of new names very often with the preservation of the initial allele name; therefore, some of the allele assignments could change with some possible changes in the associations described in the present study. The impact of improvements in HLA assignments and changes in HLA Nomenclature will be examined in some projects of the 18th IHIW. The tables presented in this manuscript are available in the 17th IHIW website (http://17ihiw.org/17th-ihiw-ngs-hla-data/).

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgements

We gratefully acknowledge the Stanford Blood center for their financial and general administrative support for the 17th IHIW. We thank members of the Histocompatibility and Immunogenetics Laboratory at the Stanford Blood Center, and Mr. Alfred K. Kornel (Research Computing, Stanford University) for their technical and laboratory support. We are also immensely grateful to the many vendors that provided reagents and software free of charge to some of the investigators. We are indebted to the Histocompatibility and Immunogenetics community and the IHIW Council for their continued dedication and support of the International Workshops. For the collection of samples in the Spanish population cohort, we thank all participating members of the Spanish Working Group in Histocompatibility and Transplant Immunology (GETHIT) of the Spanish Society for Immunology (SEI).

Funding

This project was supported in part by grant U19NS095774 (MFV, LEC, GMM) from the U.S. National Institutes of Health (NIH). The content is solely the responsibility of the authors and does not necessarily reflect the views of the NIH. This work was supported by a cooperative agreement (W81XWH-07–2–0067) between the Henry M. Jackson Foundation for the Advancement of Military Medicine, Inc., and the U.S. Department of Defense (DOD). This research was funded, in part, by the U.S. National Institute of Allergy and Infectious Disease. The views expressed are those of the authors and should not be construed to represent the positions of the U.S. Army or the DOD. This work was supported by grant from Indian Council of Medical Research, New Delhi- Sanction numbers: 63/7/2010-BMS.

Abbreviations:

HLA	Human leukocyte antigen
UPHD	Unrelated Population HLA Diversity project
AF	Allele frequency
HF	Haplotype frequency
F	homozygosity
LD	linkage disequilibrium
cALD	conditional asymmetric linkage disequilibrium
CEH	Conserved extended haplotypes

References

[1]. Robinson J, Halliwell JA, Hayhurst JD, Flicek P, Parham P, Marsh SGE, The IPD and IMGT/HLA database: allele variant databases., Nucleic Acids Res 43 (2015) D423–31. doi:10.1093/nar/gku1161. [PubMed: 25414341]

[2]. Trowsdale J, Knight JC, Major Histocompatibility Complex Genomics and Human Disease., Annu Rev Genomics Hum Genet. Annu Rev G (2015) 301–323. doi:10.1146/annurev-genom-091212-153455.Major.

[3]. Flomenberg N, Baxter-Lowe LA, Confer D, Fernandez-Vina M, Filipovich A, Horowitz M, Hurley C, Kollman C, Anasetti C, Noreen H, Begovich A, Hildebrand W, Petersdorf E, Schmeckpeper B, Setterholm M, Trachtenberg E, Williams T, Yunis E, Weisdorf D, Molecular Typing Shows a High Level of HLA Class I Incompatibility in Serologically Well Matched Donor/Patient Pairs: Implications for Unrelated Bone Marrow Donor Selection, Blood 92 (2004) 4864–4871. doi:10.1182/blood.v99.11.4200.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

[4]. Lee SJ, Klein J, Haagenson M, Baxter-Lowe LA, Confer DL, Eapen M, Fernandez-Vina M, Flomenberg N, Horowitz M, Hurley CK, Noreen H, Oudshoorn M, Petersdorf E, Setterholm M, Spellman S, Weisdorf D, Williams TM, Anasetti C, High-resolution donor-recipient HLA matching contributes to the success of unrelated donor marrow transplantation., Blood 110 (2007) 4576–83. doi:10.1182/blood-2007-06-097386. [PubMed: 17785583]

[5]. Fernández-Viña MA, Klein JP, Haagenson M, Spellman SR, Anasetti C, Noreen H, Baxter-Lowe LA, Cano P, Flomenberg N, Confer DL, Horowitz MM, Oudshoorn M, Petersdorf EW, Setterholm M, Champlin R, Lee SJ, de Lima M, Multiple mismatches at the low expression HLA loci DP, DQ, and DRB3/4/5 associate with adverse outcomes in hematopoietic stem cell transplantation., Blood 121 (2013) 4603–10. doi:10.1182/blood-2013-02-481945. [PubMed: 23596045]

[6]. Creary LE, Mallempati KC, Gangavarapu S, Caillier SJ, Oksenberg JR, Fernández-Viña MA, Deconstruction of HLA-DRB1*04:01:01 and HLA-DRB1*15:01:01 class II haplotypes using next-generation sequencing in European-Americans with multiple sclerosis, Mult. Scler. J (2018) 135245851877001. doi:10.1177/1352458518770019.

[7]. Hollenbach JA, Norman PJ, Creary LE, Damotte V, Montero-Martin G, Caillier S, Anderson KM, Misra MK, Nemat-Gorgani N, Osoegawa K, Santaniello A, Renschen A, Marin WM, Dandekar R, Parham P, Tanner CM, Hauser SL, Fernandez-Viña M, Oksenberg JR, A specific amino acid motif of HLA-DRB1 mediates risk and interacts with smoking history in Parkinson's disease., Proc. Natl. Acad. Sci. U. S. A 116 (2019) 7419–7424. doi:10.1073/pnas.1821778116. [PubMed: 30910980]

[8]. Busch R, Kollnberger S, Mellins ED, HLA associations in inflammatory arthritis: emerging mechanisms and clinical implications, Nat. Rev. Rheumatol 15 (2019) 364–381. doi:10.1038/s41584-019-0219-5. [PubMed: 31092910]

[9]. Schutte RJ, Sun Y, Li D, Zhang F, Ostrov DA, Human Leukocyte Antigen Associations in Drug Hypersensitivity Reactions, Clin. Lab. Med 38 (2018) 669–677. doi:10.1016/j.cll.2018.08.002. [PubMed: 30420060]

[10]. Hu K, Xiang Q, Wang Z, Mu G, Zhang Z, Ma L, Xie Q, Chen S, Zhou S, Zhang X, Cui Y, Associations between human leukocyte antigen polymorphisms and hypersensitivity to antiretroviral therapy in patients with human immunodeficiency virus: a meta-analysis, BMC Infect. Dis 19 (2019) 583. doi:10.1186/s12879-019-4227-5. [PubMed: 31277607]

[11]. Chowell D, Morris LGT, Grigg CM, Weber JK, Samstein RM, Makarov V, Kuo F, Kendall SM, Requena D, Riaz N, Greenbaum B, Carroll J, Garon E, Hyman DM, Zehir A, Solit D, Berger M, Zhou R, Rizvi NA, Chan TA, Patient HLA class I genotype influences cancer response to checkpoint blockade immunotherapy, Science (80-.) 359 (2018) 582–587. doi:10.1126/science.aao4572.

[12]. T.I.H.C. International HIV Controllers Study, Pereyra F, Jia X, McLaren PJ, Telenti A, de Bakker PIW, Walker BD, Ripke S, Brumme CJ, Pulit SL, Carrington M, Kadie CM, Carlson JM, Heckerman D, Graham RR, Plenge RM, Deeks SG, Gianniny L, Crawford G, Sullivan J, Gonzalez E, Davies L, Camargo A, Moore JM, Beattie N, Gupta S, Crenshaw A, Burtt NP, Guiducci C, Gupta N, Gao X, Qi Y, Yuki Y, Piechocka-Trocha A, Cutrell E, Rosenberg R, Moss KL, Lemay P, O'Leary J, Schaefer T, Verma P, Toth I, Block B, Baker B, Rothchild A, Lian J, Proudfoot J, Alvino DML, Vine S, Addo MM, Allen TM, Altfeld M, Henn MR, Le Gall S, Streeck H, Haas DW, Kuritzkes DR, Robbins GK, Shafer RW, Gulick RM, Shikuma CM, Haubrich R, Riddler S, Sax PE, Daar ES, Ribaudo HJ, Agan B, Agarwal S, Ahern RL, Allen BL, Altidor S, Altschuler EL, Ambardar S, Anastos K, Anderson B, Anderson V, Andrady U, Antoniskis D, Bangsberg D, Barbaro D, Barrie W, Bartczak J, Barton S, Basden P, Basgoz N, Bazner S, Bellos NC, Benson AM, Berger J, Bernard NF, Bernard AM, Birch C, Bodner SJ, Bolan RK, Boudreaux ET, Bradley M, Braun JF, Brndjar JE, Brown SJ, Brown K, Brown ST, Burack J, Bush LM, Cafaro V, Campbell O, Campbell J, Carlson RH, Carmichael JK, Casey KK, Cavacuiti C, Celestin G, Chambers ST, Chez N, Chirch LM, Cimoch PJ, Cohen D, Cohn LE, Conway B, Cooper DA, Cornelson B, Cox DT, V Cristofano M, Cuchural G, Czartoski JL, Dahman JM, Daly JS, Davis BT, Davis K, Davod SM, DeJesus E, Dietz CA, Dunham E, Dunn ME, Ellerin TB, Eron JJ, Fangman JJW, Farel CE, Ferlazzo H, Fidler S, Fleenor-Ford A, Frankel R, Freedberg KA, French NK, Fuchs JD, Fuller JD, Gaberman J, Gallant JE, Gandhi RT, Garcia E, Garmon D, Gathe JC, Gaultier CR, Gebre W, Gilman FD, Gilson I, Goepfert PA, Gottlieb MS,

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

Goulston C, Groger RK, Gurley TD, Haber S, Hardwicke R, Hardy WD, Harrigan PR, Hawkins TN, Heath S, Hecht FM, Henry WK, Hladek M, Hoffman RP, Horton JM, Hsu RK, Huhn GD, Hunt P, Hupert MJ, Illeman ML, Jaeger H, Jellinger RM, John M, Johnson JA, Johnson KL, Johnson H, Johnson K, Joly J, Jordan WC, Kauffman CA, Khanlou H, Killian RK, Kim AY, Kim DD, Kinder CA, Kirchner JT, Kogelman L, Kojic EM, Korthuis PT, Kurisu W, Kwon DS, LaMar M, Lampiris H, Lanzafame M, Lederman MM, Lee DM, Lee JML, Lee MJ, Lee ETY, Lemoine J, Levy JA, Llibre JM, Liguori MA, Little SJ, Liu AY, Lopez AJ, Loutfy MR, Loy D, Mohammed DY, Man A, Mansour MK, Marconi VC, Markowitz M, Marques R, Martin JN, Martin HL, Mayer KH, McElrath MJ, McGhee TA, McGovern BH, McGowan K, McIntyre D, Mcleod GX, Menezes P, Mesa G, Metroka CE, Meyer-Olson D, Miller AO, Montgomery K, Mounzer KC, Nagami EH, Nagin I, Nahass RG, Nelson MO, Nielsen C, Norene DL, O'Connor DH, Ojikutu BO, Okulicz J, Oladehin OO, Oldfield EC, Olender SA, Ostrowski M, Owen WF, Pae E, Parsonnet J, Pavlatos AM, Perlmutter AM, Pierce MN, Pincus JM, Pisani L, Price LJ, Proia L, Prokesch RC, Pujet HC, Ramgopal M, Rathod A, Rausch M, Ravishankar J, Rhame FS, Richards CS, Richman DD, Rodes B, Rodriguez M, Rose RC, Rosenberg ES, Rosenthal D, Ross PE, Rubin DS, Rumbaugh E, Saenz L, Salvaggio MR, Sanchez WC, Sanjana VM, Santiago S, Schmidt W, Schuitemaker H, Sestak PM, Shalit P, Shay W, Shirvani VN, Silebi VI, Sizemore JM, Skolnik PR, Sokol-Anderson M, Sosman JM, Stabile P, Stapleton JT, Starrett S, Stein F, Stellbrink H-J, Sterman FL, Stone VE, Stone DR, Tambussi G, Taplitz RA, Tedaldi EM, Telenti A, Theisen W, Torres R, Tosiello L, Tremblay C, Tribble MA, Trinh PD, Tsao A, Ueda P, Vaccaro A, Valadas E, Vanig TJ, Vecino I, Vega VM, Veikley W, Wade BH, Walworth C, Wanidworanun C, Ward DJ, Warner DA, Weber RD, Webster D, Weis S, Wheeler DA, White DJ, Wilkins E, Winston A, Wlodaver CG, van't Wout A, Wright DP, Yang OO, Yurdin DL, Zabukovic BW, Zachary KC, Zeeman B, Zhao M, Ueda P, Vaccaro A, Valadas E, Vanig TJ, Vecino I, Vega VM, Veikley W, Wade BH, Walworth C, Wanidworanun C, Ward DJ, Warner DA, Weber RD, Webster D, Weis S, Wheeler DA, White DJ, Wilkins E, Winston A, Wlodaver CG, van't Wout A, Wright DP, Yang OO, Yurdin DL, Zabukovic BW, Zachary KC, Zeeman B, Zhao M, The major genetic determinants of HIV-1 control affect HLA class I peptide presentation., Science 330 (2010) 1551–7. doi:10.1126/science.1195271. [PubMed: 21051598]

[13]. Tukwasibwe S, Nakimuli A, Traherne J, Chazara O, Jayaraman J, Trowsdale J, Moffett A, Jagannathan P, Rosenthal PJ, Cose S, Colucci F, Variations in killer-cell immunoglobulin-like receptor and human leukocyte antigen genes and immunity to malaria, Cell. Mol. Immunol (2020). doi:10.1038/s41423-020-0482-z.

[14]. Cao K, Hollenbach J, Shi X, Shi W, Chopek M, Ferna MA, Analysis of the frequencies of HLA-A, B, and C alleles and haplotypes in the five major ethnic groups of the United States reveals high levels of diversity in these loci and contrasting distribution patterns in these populations, 8859 (2001).

[15]. Cao K, Moormann AM, Lyke KE, Masaberg C, Sumba OP, Doumbo OK, Koech D, Lancaster A, Nelson M, Meyer D, Single R, Hartzman RJ, Plowe CV, Kazura J, Mann DL, Sztein MB, Thomson G, Fernandez-Vina MA, Differentiation between African populations is evidenced by the diversity of alleles and haplotypes of HLA class I loci, Tissue Antigens 63 (2004) 293–325. doi:10.1111/j.0001-2815.2004.00192.x. [PubMed: 15009803]

[16]. Mack SJ, Tu B, Lazaro A, Yang R, Lancaster AK, Cao K, Ng J, Hurley CK, HLA-A, -B, -C, and -DRB1 allele and haplotype frequencies distinguish Eastern European Americans from the general European American population., Tissue Antigens 73 (2009) 17–32. doi:10.1111/j.1399-0039.2008.01151.x. [PubMed: 19000140]

[17]. Goeury T, Creary LE, Brunet L, Galan M, Pasquier M, Kervaire B, Langaney A, Tiercy JM, Fernández-Viña MA, Nunes JM, Sanchez-Mazas A, Deciphering the fine nucleotide diversity of full HLA class I and class II genes in a well-documented population from sub-Saharan Africa., HLA 91 (2018) 36–51. doi:10.1111/tan.13180. [PubMed: 29160618]

[18]. Thorstenson YR, Creary LE, Huang H, Rozot V, Nguyen TT, Babrzadeh F, Kancharla S, Fukushima M, Kuehn R, Wang C, Li M, Krishnakumar S, Mindrinos M, Fernandez Viña MA, Scriba TJ, Davis MM, Allelic resolution NGS HLA typing of Class I and Class II loci and haplotypes in Cape Town, South Africa, Hum. Immunol 79 (2018) 839–847. doi:10.1016/J.HUMIMM.2018.09.004. [PubMed: 30240896]

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

[19]. Maiers M, Gragert L, Klitz W, High-resolution HLA alleles and haplotypes in the United States population, Hum. Immunol 68 (2007) 779–788. doi:10.1016/j.humimm.2007.04.005. [PubMed: 17869653]

[20]. Meyer D, Aguiar VRC, Bitarello BD, Brandt DYC, Nunes K, A genomic perspective on HLA evolution, Immunogenetics 70 (2018) 5–27. doi:10.1007/s00251-017-1017-3. [PubMed: 28687858]

[21]. Prugnolle F, Manica A, Charpentier M, Guégan JF, Guernier V, Balloux F, Pathogen-Driven Selection and Worldwide HLA Class I Diversity, Curr. Biol 15 (2005) 1022–1027. doi:10.1016/J.CUB.2005.04.050. [PubMed: 15936272]

[22]. Arrieta-Bolaños E, Madrigal-Sánchez JJ, Stein JE, Órlich-Pérez P, Moreira-Espinoza MJ, Paredes-Carias E, Vanegas-Padilla Y, Salazar-Sánchez L, Madrigal JA, Marsh SGE, Shaw BE, High-resolution HLA allele and haplotype frequencies in majority and minority populations of Costa Rica and Nicaragua: Differential admixture proportions in neighboring countries, HLA 91 (2018) 514–529. doi:10.1111/tan.13280. [PubMed: 29687625]

[23]. Meyer D, Single RM, Mack SJ, Erlich HA, Thomson G, Signatures of Demographic History and Natural Selection in the Human Major Histocompatibility Complex Loci, Genetics 173 (2006) 2121–2142. doi:10.1534/genetics.105.052837. [PubMed: 16702436]

[24]. Dupont B, Immunobiology of HLA: Volume I Histocompatibility Testing 1987, 1st ed., Springer-Verlag New York Inc., New York, 1989.

[25]. Dupont B, Immunobiology of HLA: Volume II Immunogenetics and Histocompatibility, 1st ed., Springer-Verlag New York Inc., New York, 1989.

[26]. Nunes JM, Riccio ME, Buhler S, Di D, Currat M, Ries F, Almada AJ, Benhamamouch S, Benitez O, Canossi A, Fadhlaoui-Zid K, Fischer G, Kervaire B, Loiseau P, de Oliveira DCM, Papasteriades C, Piancatelli D, Rahal M, Richard L, Romero M, Rousseau J, Spiroski M, Sulcebe G, Middleton D, Tiercy J-M, Sanchez-Mazas A, Analysis of the HLA population data (AHPD) submitted to the 15th International Histocompatibility/Immunogenetics Workshop by using the Gene[rate] computer tools accommodating ambiguous data (ahpd project report), Tissue Antigens 76 (2010) 18–30. doi:10.1111/j.1399-0039.2010.01469.x. [PubMed: 20331842]

[27]. Riccio ME, Buhler S, Nunes JM, Vangenot C, Cuénod M, Currat M, Di D, Andreani M, Boldyreva M, Chambers G, Chernova M, Chiaroni J, Darke C, Di Cristofaro J, Dubois V, Dunn P, Edinur HA, Elamin N, Eliaou J-F, Grubic Z, Jaatinen T, Kanga U, Kervaire B, Kolesar L, Kunachiwa W, Lokki ML, Mehra N, Nicoloso G, Paakkanen R, Voniatis DP, Papasteriades C, Poli F, Richard L, Alonso IR, Slavčev A, Sulcebe G, Suslova T, Testi M, Tiercy J-M, Varnavidou A, Vidan-Jeras B, Wennerström A, Sanchez-Mazas A, 16th IHIW: Analysis of HLA Population Data, with updated results for 1996 to 2012 workshop data (AHPD project report), Int. J. Immunogenet 40 (2012) n/a–n/a. doi:10.1111/iji.12033.

[28]. Kobler B, Lattes P, International Migration Report 2017, Dep. Econ. Soc. Aff. United Nations (2017). doi:ST/ESA/SER.A/404.

[29]. Bunce M, O'Neill CM, Barnardo MCNM, Krausa P, Browning MJ, Morris PJ, Welsh KI, Phototyping: comprehensive DNA typing for HLA-A, B, C, DRB1, DRB3, DRB4, DRB5 & DQB1 by PCR with 144 primer mixes utilizing sequence-specific primers (PCR-SSP), Tissue Antigens 46 (1995) 355–367. doi:10.1111/j.1399-0039.1995.tb03127.x. [PubMed: 8838344]

[30]. Cao K, Chopek M, Fernández-Viña MA, High and intermediate resolution DNA typing systems for class I HLA-A, B, C genes by hybridization with sequence-specific oligonucleotide probes (SSOP)., Rev. Immunogenet 1 (1999) 177–208. [PubMed: 11253946]

[31]. Cereb N, Maye P, Lee S, Kong Y, Yang SY, Locus-specific amplification of HLA class I genes from genomic DNA: locus-specific sequences in the first and third introns of HLA-A, -B, and -C alleles, Tissue Antigens 45 (1995) 1–11. doi:10.1111/j.1399-0039.1995.tb02408.x. [PubMed: 7725305]

[32]. Kotsch K, Wehllng J, Köhler S, Blasczyk R, Sequencing of HLA class I genes based on the conserved diversity of the noncoding regions: sequencing-based typing of the HLA-A gene, Tissue Antigens 50 (1997) 178–191. doi:10.1111/j.1399-0039.1997.tb02857.x. [PubMed: 9271828]

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

[33]. Kotsch K, Wehling J, Blasczyk R, Sequencing of HLA class II genes based on the conserved diversity of the non-coding regions: sequencing based typing of HLA-DRB genes, Tissue Antigens 53 (1999) 486–497. doi:10.1034/j.1399-0039.1999.530505.x. [PubMed: 10372544]

[34]. Erlich HA, HLA typing using next generation sequencing: An overview, Hum. Immunol 76 (2015) 887–890. doi:10.1016/J.HUMIMM.2015.03.001. [PubMed: 25777625]

[35]. Goeury T, Creary LE, Fernandez-Vina MA, Tiercy J-M, Nunes JM, Sanchez-Mazas A, Mandenka from Senegal, HLA 91 (2018) 148–150. doi:10.1111/tan.13197. [PubMed: 29280562]

[36]. Creary LE, Guerra SG, Chong W, Brown CJ, Turner TR, Robinson J, Bultitude WP, Mayor NP, Marsh SGE, Saito K, Lam K, Duke JL, Mosbruger TL, Ferriola D, Monos D, Willis A, Askar M, Fischer G, Saw CL, Ragoussis J, Petrek M, Serra-Pagés C, Juan M, Stavropoulos-Giokas C, Dinou A, Ameen R, Al Shemmari S, Spierings E, Gendzekhadze K, Morris GP, Zhang Q, Kashi Z, Hsu S, Gangavarapu S, Mallempati KC, Yamamoto F, Osoegawa K, Vayntrub T, Chang C-J, Hansen JA, Fernández-Viña MA, Next-generation HLA typing of 382 International Histocompatibility Working Group reference B-lymphoblastoid cell lines: Report from the 17th International HLA and Immunogenetics Workshop, Hum. Immunol (2019). doi:10.1016/J.HUMIMM.2019.03.001.

[37]. Creary LE, Gangavarapu S, Mallempati KC, Montero-Martín G, Caillier SJ, Santaniello A, Hollenbach JA, Oksenberg JR, Fernández-Viña MA, Next-generation sequencing reveals new information about HLA allele and haplotype diversity in a large European American population, Hum. Immunol (2019). doi:10.1016/J.HUMIMM.2019.07.275.

[38]. Mayor NP, Robinson J, McWhinnie AJM, Ranade S, Eng K, Midwinter W, Bultitude WP, Chin C-S, Bowman B, Marks P, Braund H, Madrigal JA, Latham K, Marsh SGE, HLA Typing for the Next Generation., PLoS One 10 (2015) e0127153. doi:10.1371/journal.pone.0127153. [PubMed: 26018555]

[39]. Turner TR, Hayhurst JD, Hayward DR, Bultitude WP, Barker DJ, Robinson J, Madrigal JA, Mayor NP, Marsh SGE, Single molecule real-time DNA sequencing of HLA genes at ultra-high resolution from 126 International HLA and Immunogenetics Workshop cell lines, HLA 91 (2018) 88–101. doi:10.1111/tan.13184. [PubMed: 29171935]

[40]. Chang C-J, Osoegawa K, Milius RP, Maiers M, Xiao W, Fernandez-Viña M, Mack SJ, Collection and storage of HLA NGS genotyping data for the 17th International HLA and Immunogenetics Workshop, Hum. Immunol 79 (2018) 77–86. doi:10.1016/J.HUMIMM.2017.12.004. [PubMed: 29247682]

[41]. Milius RP, Heuer M, George M, Pollack J, Hollenbach JA, Mack SJ, Maiers M, The GL service: Web service to exchange GL string encoded HLA & KIR genotypes with complete and accurate allele and genotype ambiguity., Hum. Immunol 77 (2016) 249–256. doi:10.1016/j.humimm.2015.11.017. [PubMed: 26621609]

[42]. Milius RP, Heuer M, Valiga D, Doroschak KJ, Kennedy CJ, Bolon Y-T, Schneider J, Pollack J, Kim HR, Cereb N, Hollenbach JA, Mack SJ, Maiers M, Histoimmunogenetics Markup Language 1.0: Reporting next generation sequencing-based HLA and KIR genotyping., Hum. Immunol 76 (2015) 963–74. doi:10.1016/j.humimm.2015.08.001. [PubMed: 26319908]

[43]. Andersson G, Evolution of the human HLA-DR region., Front. Biosci 3 (1998) d739–45. [PubMed: 9675159]

[44]. 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 74 (2013) 1313–1320. doi:10.1016/j.humimm.2013.06.025. [PubMed: 23806270]

[45]. Lancaster AK, Single RM, Solberg OD, Nelson MP, Thomson G, PyPop update--a software pipeline for large-scale multilocus population genomics., Tissue Antigens 69 Suppl 1 (2007) 192–7. doi:10.1111/j.1399-0039.2006.00769.x. [PubMed: 17445199]

[46]. Guo SW, Thompson EA, Performing the exact test of Hardy-Weinberg proportion for multiple alleles., Biometrics 48 (1992) 361–72. [PubMed: 1637966]

[47]. Hedrick PW, Gametic disequilibrium measures: proceed with caution., Genetics 117 (1987) 331–41. [PubMed: 3666445]

[48]. Cramer H, Mathematical Methods of Statistics, I, Princeton: Princeton University, 1946.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

[49]. Thomson G, Single RM, Conditional asymmetric linkage disequilibrium (ALD): extending the biallelic r2 measure., Genetics 198 (2014) 321–31. doi:10.1534/genetics.114.165266. [PubMed: 25023400]

[50]. Pappas DJ, Marin W, Hollenbach JA, Mack SJ, Bridging ImmunoGenomic Data Analysis Workflow Gaps (BIGDAWG): An integrated case-control analysis pipeline., Hum. Immunol 77 (2016) 283–287. doi:10.1016/j.humimm.2015.12.006. [PubMed: 26708359]

[51]. González-Quezada BA, Creary LE, Munguia-Saldaña AJ, Flores-Aguilar H, Fernández-Viña MA, Gorodezky C, Exploring the ancestry and admixture of Mexican Oaxaca Mestizos from Southeast Mexico using next-generation sequencing of 11 HLA loci, Hum. Immunol 80 (2019) 157–162. doi:10.1016/J.HUMIMM.2019.01.004. [PubMed: 30708029]

[52]. WAHLUND S, ZUSAMMENSETZUNG VON POPULATIONEN UND KORRELATIONSERSCHEINUNGEN VOM STANDPUNKT DER VERERBUNGSLEHRE AUS BETRACHTET, Hereditas 11 (2010) 65–106. doi:10.1111/j.1601-5223.1928.tb02483.x.

[53]. Humes RR, Karen R; Jones Nicholas A.; Ramirez, United States Census Bureau, Overview of Race and Hispanic Origin: 2010, United States Census Bur. (2011) 1–23. https://www.census.gov/prod/cen2010/briefs/c2010br-02.pdf (accessed August 20, 2018).

[54]. Tishkoff SA, Reed FA, Friedlaender FR, Ehret C, Ranciaro A, Froment A, Hirbo JB, Awomoyi AA, Bodo J-M, Doumbo O, Ibrahim M, Juma AT, Kotze MJ, Lema G, Moore JH, Mortensen H, Nyambo TB, Omar SA, Powell K, Pretorius GS, Smith MW, Thera MA, Wambebe C, Weber JL, Williams SM, The genetic structure and history of Africans and African Americans., Science 324 (2009) 1035–44. doi:10.1126/science.1172257. [PubMed: 19407144]

[55]. Testi M, Battarra M, Lucarelli G, Isgro A, Morrone A, Akinyanju O, Wakama T, Nunes JM, Andreani M, Sanchez-Mazas A, HLA-A-B-C-DRB1-DQB1 phased haplotypes in 124 Nigerian families indicate extreme HLA diversity and low linkage disequilibrium in Central-West Africa, Tissue Antigens 86 (2015) 285–292. doi:10.1111/tan.12642. [PubMed: 26300115]

[56]. Gambino CM, Aiello A, Accardi G, Caruso C, Candore G, Autoimmune diseases and 8.1 ancestral haplotype: An update, HLA 92 (2018) 137–143. doi:10.1111/tan.13305. [PubMed: 29877054]

[57]. Lande A, Andersen I, Egeland T, Lie BA, Viken MK, HLA -A, -C, -B, -DRB1, -DQB1 and -DPB1 allele and haplotype frequencies in 4514 healthy Norwegians, Hum. Immunol 79 (2018) 527–529. doi:10.1016/J.HUMIMM.2018.04.012. [PubMed: 29684411]

[58]. Seshasubramanian V, Manisekar NK, Sathishkannan AD, Naganathan C, Narayan S, Next Generation Sequencing in HLA haplotype distribution among Telugu speaking population from Andhra Pradesh, India, Hum. Immunol 79 (2018) 583–584. doi:10.1016/j.humimm.2018.05.005. [PubMed: 29890180]

[59]. Belich MP, Madrigal JA, Hildebrand WH, Zemmour J, Williams RC, Luz R, Petzl-Erler ML, Parham P, Unusual HLA-B alleles in two tribes of Brazilian Indians, Nature 357 (1992) 326–329. doi:10.1038/357326a0. [PubMed: 1317015]

[60]. Cooper DN, Functional intronic polymorphisms: Buried treasure awaiting discovery within our genes., Hum. Genomics 4 (2010) 284–8. doi:10.1186/1479-7364-4-5-284. [PubMed: 20650817]

[61]. Lambert CA, Tishkoff SA, Genetic structure in African populations: implications for human demographic history., Cold Spring Harb. Symp. Quant. Biol 74 (2009) 395–402. doi:10.1101/sqb.2009.74.053. [PubMed: 20453204]

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

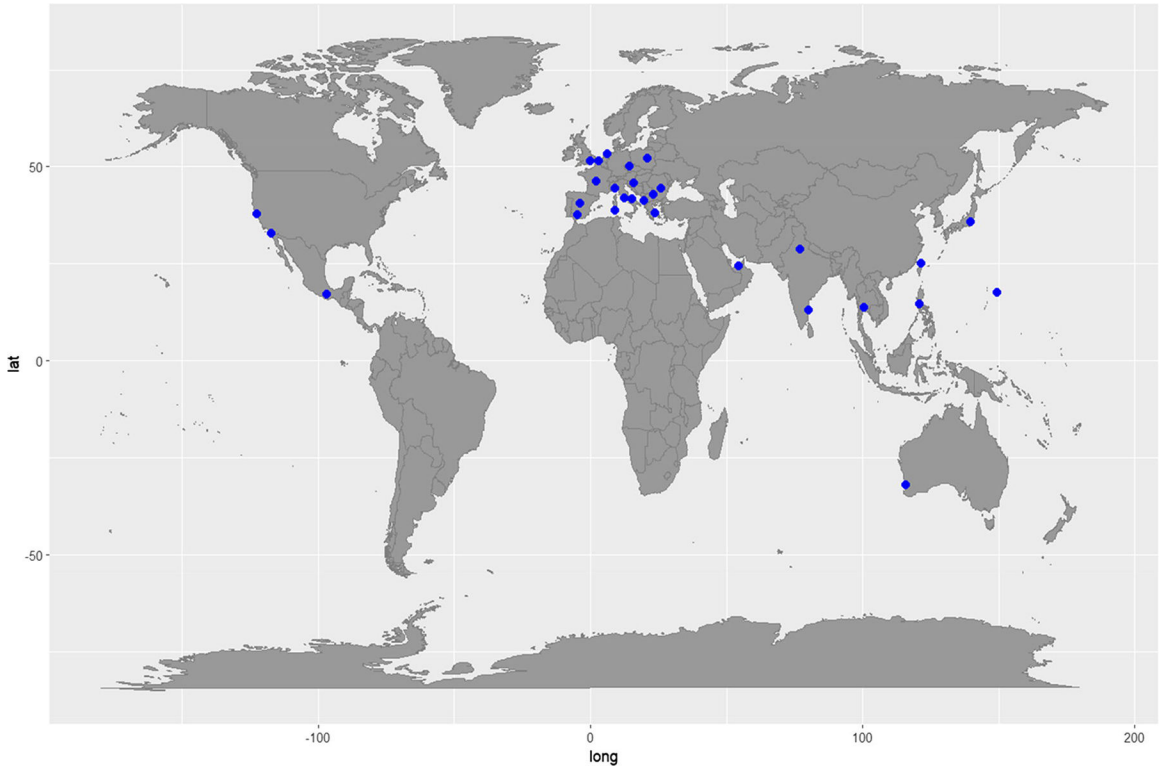

Figure 1.
Global location of population samples submitted to the 17th IHIW Unrelated Population HLA Diversity (UPHD) project.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

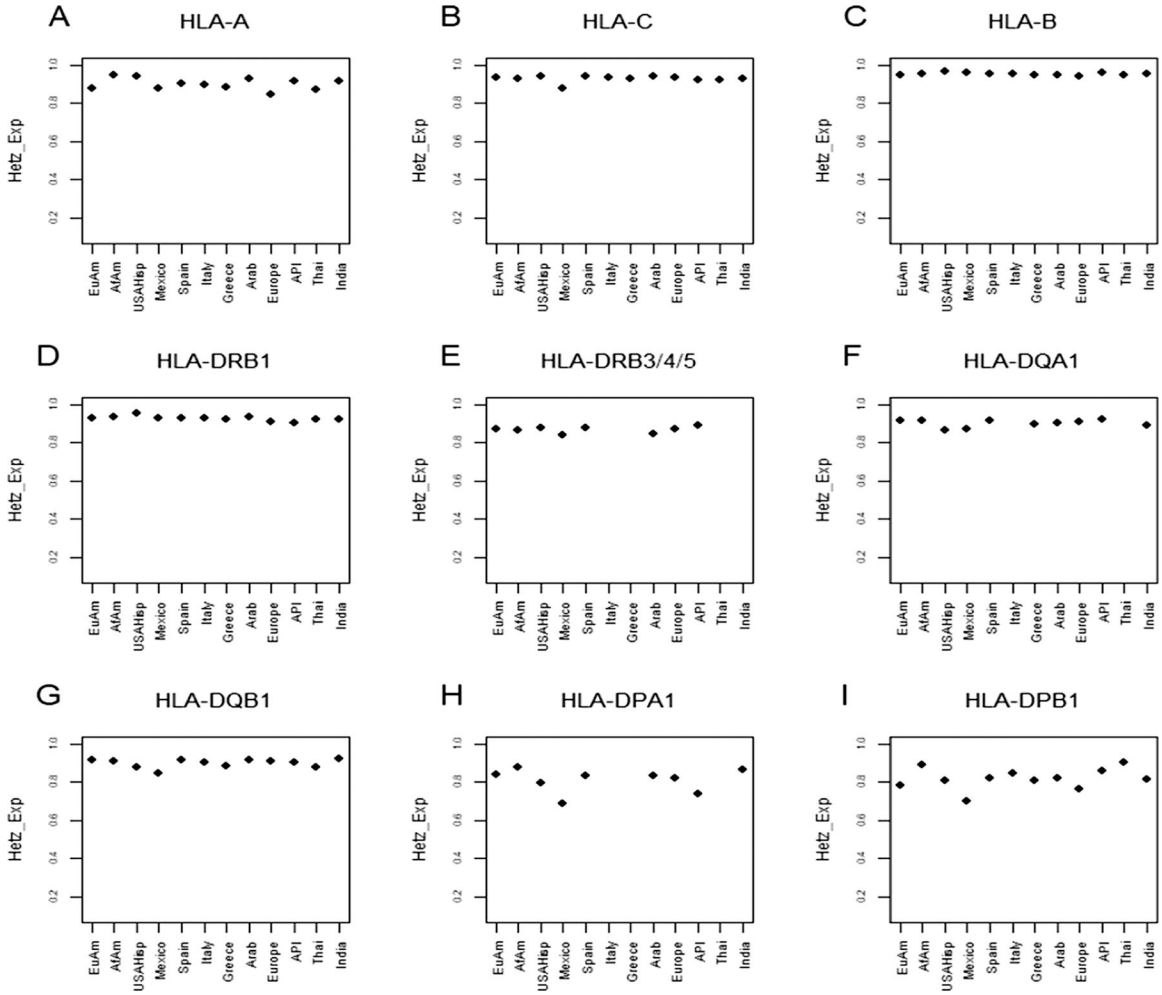

Figure 2.

Column plots of expected heterozygosity values for HLA class I and II loci (4-Field data).

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

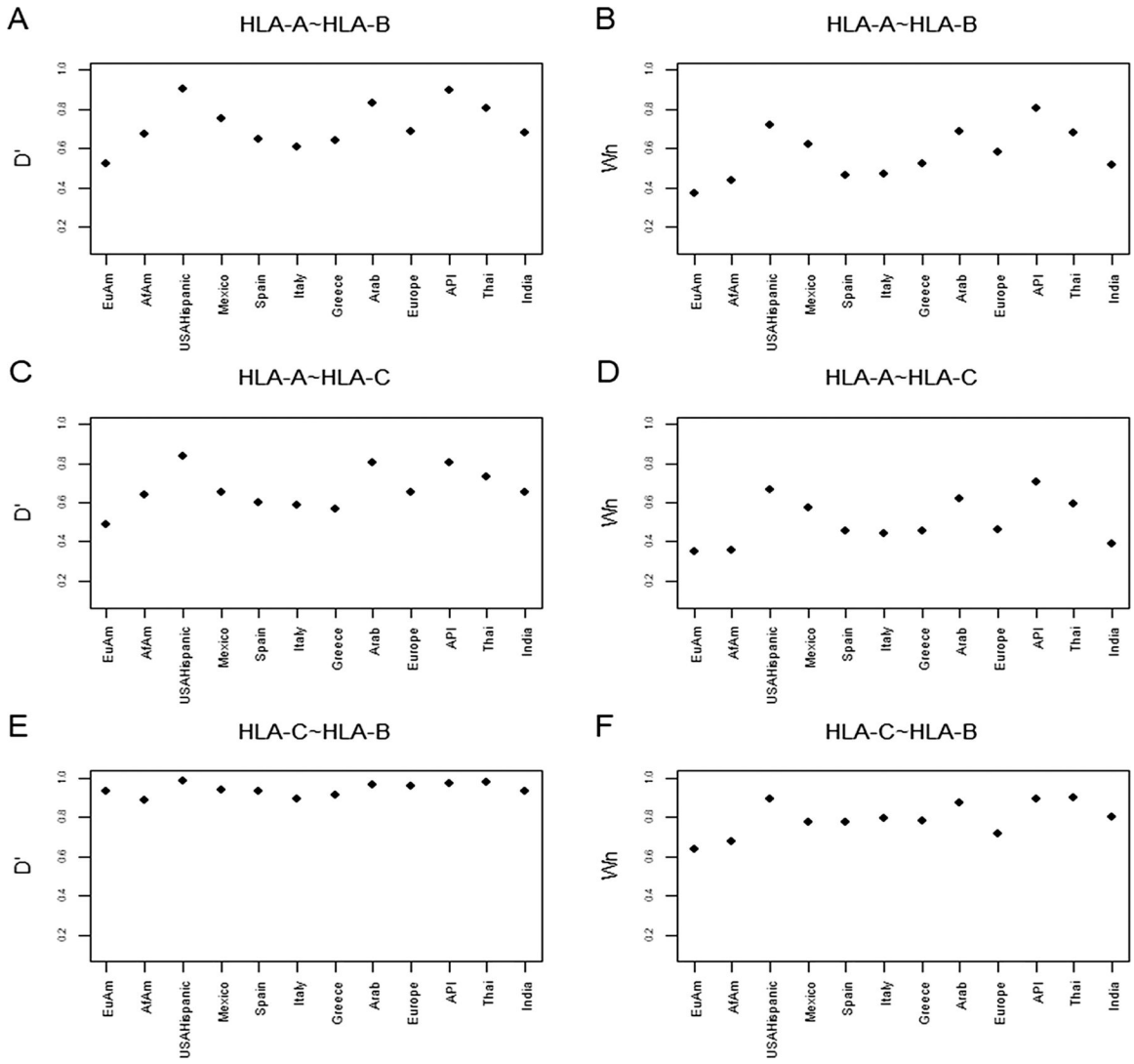

Figure 3.
Global linkage disequilibrium values of D' and W_n class I haplotypes by population.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

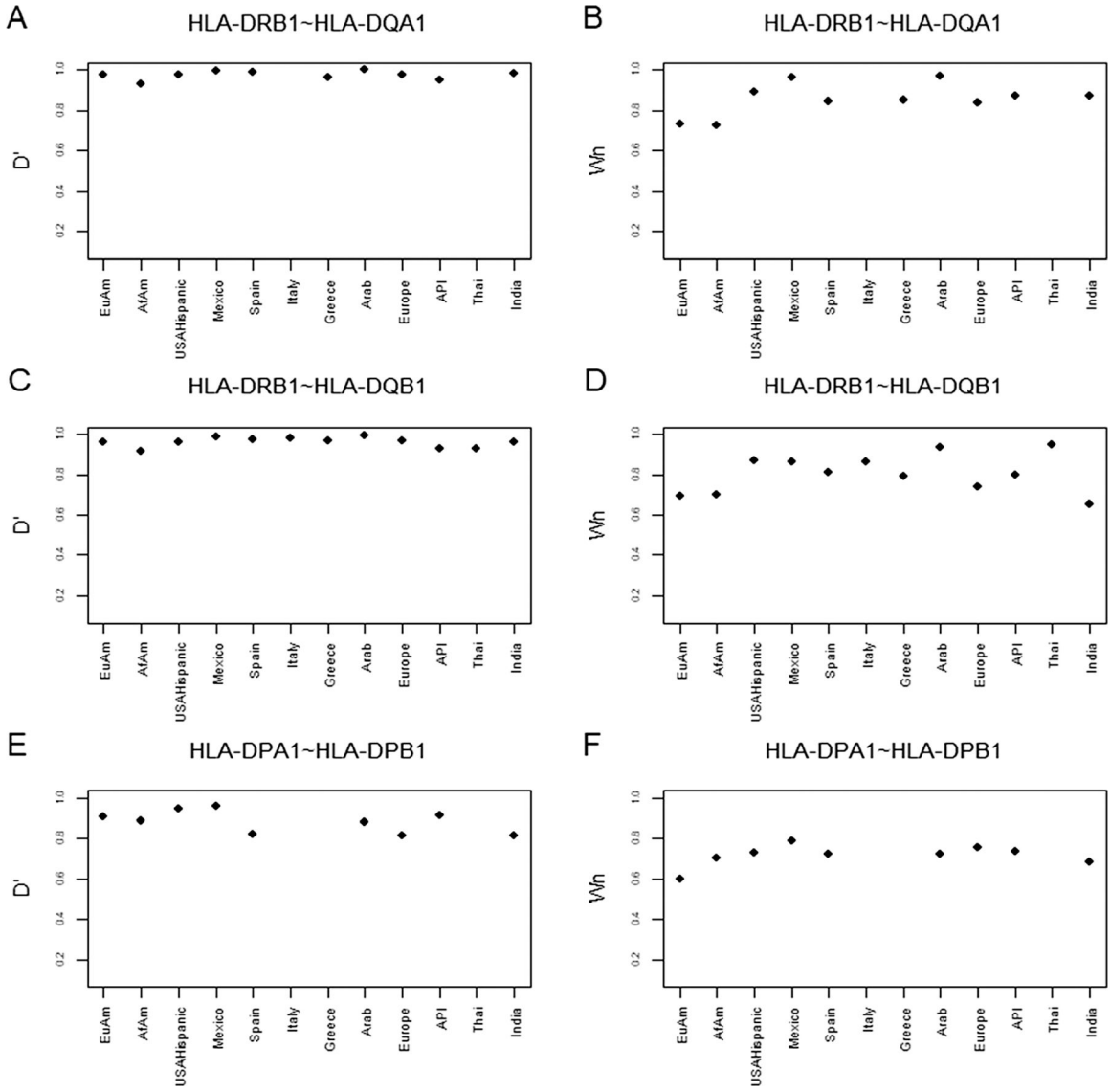

Figure 4.
Global linkage disequilibrium values of D' and W_n class II haplotypes by population.

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

Table 1

List of participants who submitted data to the 17th IHIW Unrelated Population HLA diversity (UPHD) project.

Labcode	PI(s) name	Country	No. of Populations
areelg	Gehad EIGhazali	United Arab Emirates	1
	Zain Al Yafei		
ausdes	Dianne De Santis	Australia	3
gbrcn	Colin J Brown	United Kingdom	3
gbrpep	Jennifer Pepperall	United Kingdom	1
grcsta	Catherine Stavropoulos-Giokasi	Greece	1
	Amalia Dinou		
indkan	Uma Kanga	India	1
indnar	Saranya Narayan	India	1
	Srinivasan Periathiruvadi		
itasac	Nicoletta Sacchi	Italy	1
	Michela Mazzocco		
mexgor	Clara Gorodezky	Mexico	1
ucamor	Gerald P. Morris	United States of America	4
uilmar	Susana R. Marino	United States of America	2
umdtho	Rasmi Thomas	United States of America	4
ussta1	Marcelo A. Fernandez-Vina	United States of America	3

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

Table 2

17th IHIW Unrelated Population HIA diversity (UPHD) population samples and HLA loci typed (4-Field).

	Number (2n)								
Population, n	**HLA-A**	**HLA-C**	**HLA-B**	**HLA-DRB1**	**HLA-DRB3/4/5**	**HLA-DQA1**	**HLA-DQB1**	**HLA-DPA1**	**HLA-DPB1**
European American, 2423	4846	4796	4838	4790	4836	4712	4704	4752	4826
African American, 394	788	788	788	784	750	758	786	760	788
USA Hispanic, 56	110	112	112	112	100	100	112	100	112
Mexican, 104	208	208	208	208	208	208	208	208	208
Spanish, 276	552	552	552	544	546	544	538	550	548
Italian, 292	584	584	584	584	NT	NT	584	NT	518
Greek, 184	358	366	356	368	NT	272	366	NT	268
Arab, 52	104	104	104	104	104	104	104	104	104
European[a], 212	424	424	424	424	394	424	424	424	424
Albanian. 2	4	4	4	4	4	4	4	4	4
Bulgarian, 1	2	2	2	2	2	2	2	2	2
Croatian. 2	4	4	4	4	4	4	4	4	4
Czech Republic, 1	2	2	2	2	2	2	2	2	2
English, 103	206	206	206	206	206	206	206	206	206
French, 1	2	2	2	2	2	2	2	2	2
German. 1	2	2	2	2	2	2	2	2	2
Irish, 1	2	2	2	2	2	2	2	2	2
Italian, 1	2	2	2	2	2	2	2	2	2
Mixed of NE ancestry[b] 82	164	164	164	164	164	164	164	164	164
Polish, 6	12	12	12	12	12	12	12	12	12
Portuguese, 2	4	4	4	4	4	4	4	4	4
Romanian. 9	18	18	18	18	18	18	18	18	18
API[c], 43	86	86	86	86	82	82	84	82	86
Filipino, 3	6	6	6	6	6	6	6	6	6
Japanese. 1	2	2	2	2	2	2	2	2	2
Mixed of API ancestry[d] 13	26	26	26	26	26	26	26	26	26
Polynesian. 1	2	2	2	2	2	2	2	2	2
Taiwanese Hakka. 1	2	2	2	2	2	2	2	2	2
USA Asian, 24	48	48	48	48	44	44	46	44	48
Thai, 42	84	84	84	84	NT	NT	84	NT	84
Indian, 162	322	324	324	324	NT	94	276	94	96

Abbreviations: 2n, total chromosome count: NT, HLA gene not typed: NE, Northern European: API, Asian Pacific Islander.

[a]The European population is comprised of the twelve populations from Europe shown in italicized text.

[b] Refers to individuals mixed of entirely NE ancestry or mixed of predominantly NE ancestry.

[c] The API population is comprised of the six populations of Asian ancestry shown in italicized text.

[d] Refers to individuals mixed of entirely API ancestry or mixed of predominantly API ancestry.

Author Manuscript
Author Manuscript
Author Manuscript
Author Manuscript

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

Table 3

NGS systems and HLA Loci typed by participating laboratories.

Labcode	Software	Hardware	HLA loci typed	Coverage class I/class II
areelg *	NGSengine, GenDx	MiSeq, Illumina	A, C, B, DRB1/3/4/5, DQA1, DQB1, DPA1, DPBI	Full-gene/wide-coverage
ausdes	NGSengine, GenDx	Ion S5™ system	A, C, B, DRB1/3/4/5, DQB1, DPB1	Full-gene/wide-coverage
gbrcn	TypeStream™ Visual, One lambda Thermo Fisher Scientific	Ion S5™ system	A, C, B, DRB1/3/4/5, DQA1, DQB1, DPA1, DPBI	Full-gene/full-gene (DQA1, DPA1) and wide-coverage
gbrpep *	NGSengine, GenDx	MiSeq, Illumina	A, C, B, DRB1/3/4/5, DQA1, DQB1, DPA1, DPBI	Full-gene/wide-coverage
grcsta	HLA Twin, Omixon	MiSeq, Illumina	A, C, B, DRB1, DQA1, DQB1, DPB1	Full-gene/wide-coverage
indkan	NGSengine, GenDx	MiSeq, Illumina	A, C, B, DRB1, DQA1, DQB1, DPA1, DPB1	Full-gene/wide-coverage
indnar	MIA FORA FLEX, Immucor	MiniSeq, Illumina	A, C, B, DRB1, DQB1	Full-gene/wide-coverage
itasac	NGSengine, GenDx	MiSeq, Illumina	A, C, B, DRB1, DQB1	Full-gene/wide-coverage
mexgor ‡	MIA FORA FLEX, Immucor	MiSeq, Illumina	A, C, B, DRB1/3/4/5, DQA1, DQB1, DPA1, DPB1	Full-gene/wide-coverage
ucamor	TruSight HLA Assign, Illumina	MiSeq, Illumina	A, C, B, DRB1/3/4/5, DQA1, DQB1, DPA1, DPBI	Full-gene (A, C), B (exon1-intron 6) /full-gene (DQ DPA1) and wide-coverage
uilmar	NGSengine, GenDx	MiSeq, Illumina	A, C, B, DRB1/3/4/5, DQA1, DQB1, DPA1, DPBI	Full-gene/wide-coverage
umdtho	NGSengine, GenDx	MiSeq, Illumina	A, C, B, DRB1, DQB1, DPB1	Full-gene/wide-coverage
ussta1	MIA FORA FLEX, Immucor	MiniSeq and NextSeq, Illumina	A, C, B, DRB1/3/4/5, DQA1, DQB1, DPA1, DPB1	Full-gene/wide-coverage

* Samples typed by laboratory utxask (Baylor University Medical center).

‡ Samples typed by laboratory ussta1 .

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

Table 4

Probability values for Guo and Thompson Hardy-Weinberg equilibrium (HWE) tests in the UPHD populations.

Maximum allelic resolution[a]									
Population	**HLA-A**	**HLA-C**	**HLA-B**	**HLA-DRB1**	**HLA-DRB3/4/5**	**HLA-DQA1**	**HLA-DQB1**	**HLA-DPA1**	**HLA-DPB1**
European American	**0.0036**	**0.0001**	**0.0000**	**0.0001**	**0.0000**	**0.0067**	**0.0014**	**0.0000**	**0.0000**
African American	0.3967	**0.0458**	**0.0004**	**0.0002**	**0.0045**	**0.0002**	**0.0010**	**0.0457**	0.1030
USA Hispanic	**0.0034**	**0.0011**	0.0977	0.1916	**0.0107**	0.2181	0.0879	**0.0227**	**0.0331**
Mexican	0.2713	**0.0491**	0.0629	0.2286	0.0806	**0.0375**	**0.0183**	0.2912	0.1346
Spanish	0.5230	0.9906	0.1151	0.8920	0.5300	0.3015	0.7955	**0.0129**	0.9012
Italian	0.8447	0.0985	0.8679	0.2448	NT	NT	0.1224	NT	**0.0025**
Greek	0.4691	0.5181	0.7447	0.3585	NT	0.1268	0.2305	NT	0.1233
Arab	**0.0007**	0.1866	0.1187	0.2123	0.5325	0.1640	0.1951	0.3769	0.4846
European	0.9083	0.2381	0.0852	**0.0075**	**0.0082**	**0.0012**	**0.0068**	**3.0E-05**	0.1332
API	**0.0005**	**6.0E-05**	**0.0000**	**0.0003**	**0.0009**	**0.0001**	**0.0226**	0.1848	0.2367
Thai	0.8839	0.0970	0.3526	0.2229	NT	NT	0.2942	NT	0.7104
India	0.2216	0.8316	0.2299	0.4482	NT	0.4661	0.1402	**0.0219**	**0.0197**
2-Field allele resolution[b]									
Population	HLA-A	HLA-C	H LA-B	HLA-DRB1	HLA-DRB3/4/5	HLA-DQA1	HLA-DQB1	HLA-DPA1	HLA-DPB1
European American	**0.0024**	**0.0053**	**2.0E-05**	**0.0001**	**0.0005**	0.1575	**0.0027**	**0.0007**	**0.0000**
African American	0.5170	**0.0446**	**0.0004**	**0.0003**	**0.0122**	0.0969	**0.0491**	0.5912	0.1328
USA Hispanic	**0.0041**	0.1008	0.0511	0.1217	0.3845	0.1598	0.2138	0.3784	**0.0249**
Mexican	0.3126	0.0598	0.0835	0.1564	0.2390	**0.0181**	0.0641	0.0713	0.5244
Spanish	0.3135	0.9281	0.2699	0.8588	0.2817	0.7791	0.5180	0.1513	0.8987
Italian	0.8016	0.2731	0.6705	0.2305	NT	NT	0.2882	NT	**0.0004**
Greek	0.3984	0.8091	0.8724	0.3815	NT	0.3061	0.1838	NT	0.0579
Arab	**0.0003**	0.2089	0.1479	0.1720	0.6508	0.4250	0.1559	0.7203	0.3533
European	0.7474	0.1366	0.1881	**0.0039**	0.5274	**0.0340**	**0.0016**	0.7028	0.1569
API	**0.0005**	**0.0019**	**0.0000**	**0.0004**	0.0624	**0.0112**	**0.0430**	0.5365	0.2509
Thai	0.8633	0.0970	0.3526	0.2229	NT	NT	0.2942	NT	0.6943
Indian	0.4011	0.5449	0.6403	0.3584	NT	0.5607	0.1448	1.0000	**0.0197**

HLA loci that deviated significantly ($P < 0.05$) from HWE are shown in boldface type.

[a] HWE test performed on alleles characterized at maximum (3–4-Field) allelic resolution.

[b] HWE test performed on alleles characterized at 2-Field allele resolution.

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

Table 5

The number of unique HLA alleles per population.

Number of unique alleles *(k)*									
Population	**HLA-A**	**HLA-C**	**HLA-B**	**HLA-DRB1**	**HLA-DRB3/4/5**	**HLA-DQA1**	**HLA-DQB1**	**HLA-DPA1**	**HLA-DPB1**
European American	70	72	107	59	24	36	33	23	45
African American	50	50	71	47	14	28	27	18	41
USA Hispanic	30	30	43	31	13	20	20	11	20
Mexican	27	29	46	34	13	19	21	14	16
Spanish	36	40	53	37	14	23	24	14	28
Italian	43	41	58	38	NT	NT	24	NT	32
Greek	36	37	58	33	NT	25	22	NT	22
Arab	30	26	36	28	12	19	17	11	18
European	45	43	58	38	17	25	24	13	25
API	18	21	30	23	13	19	16	9	17
Thai	17	16	29	23	NT	NT	12	NT	19
Indian	34	32	54	33	NT	19	29	11	18

Abbreviations; NT; Loci not typed.