Human Immunology 80 (2019) 807–822

Contents lists available at ScienceDirect

Human Immunology

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

Next-generation sequencing reveals new information about HLA allele and haplotype diversity in a large European American population

Lisa E. Creary[a,b,*], Sridevi Gangavarapu[b], Kalyan C. Mallempati[b], Gonzalo Montero-Martín[a,b], Stacy J. Caillier[c], Adam Santaniello[c], Jill A. Hollenbach[c], Jorge R. Oksenberg[c], Marcelo A. Fernández-Viña[a,b]

[a] Department of Pathology, Stanford University School of Medicine, Palo Alto, CA, USA
[b] Histocompatibility, Immunogenetics and Disease Profiling Laboratory, Stanford Blood Center, Palo Alto, CA, USA
[c] Department of Neurology, School of Medicine, University of California San Francisco, San Francisco, CA, USA

ARTICLE INFO

Keywords:
Human leukocyte antigen
Next-generation sequencing
European Americans
Allele frequency
Haplotype frequency
Linkage disequilibrium
Haplotype blocks
Population genetics

ABSTRACT

The human leukocyte antigen (HLA) genes are extremely polymorphic and are useful molecular markers to make inferences about human population history. However, the accuracy of the estimation of genetic diversity at HLA loci very much depends on the technology used to characterize HLA alleles; high-resolution genotyping of long-range HLA gene products improves the assessment of HLA population diversity as well as other population parameters compared to lower resolution typing methods. In this study we examined allelic and haplotype HLA diversity in a large healthy European American population sourced from the UCSF-DNA bank. A high-resolution next-generation sequencing method was applied to define non-ambiguous 3- and 4-field alleles at the *HLA-A*, *HLA-C*, *HLA-B*, *HLA-DRB1*, *HLA-DRB3/4/5*, *HLA-DQA1*, *HLA-DQB1*, *HLA-DPA1*, and *HLA-DPB1* loci in samples provided by 2248 unrelated individuals. A number of population parameters were examined including balancing selection and various measurements of linkage disequilibrium were calculated. There were no detectable deviations from Hardy-Weinberg proportions at *HLA-A*, *HLA-DRB1*, *HLA-DQA1* and *HLA-DQB1*. For the remaining loci moderate and significant deviations were detected at *HLA-C*, *HLA-B*, *HLA-DRB3/4/5*, *HLA-DPA1* and *HLA-DPB1* loci mostly from population substructures. Unique 4-field associations were observed among alleles at 2 loci and haplotypes extending large intervals that were not apparent in results obtained using testing methodologies with limited sequence coverage and phasing. The high diversity at *HLA-DPA1* results from detection of intron variants of otherwise well conserved protein sequences. It may be speculated that divergence in exon sequences may be negatively selected. Our data provides a valuable reference source for future population studies that may allow for precise fine mapping of coding and non-coding sequences determining disease susceptibility and allo-immunogenicity.

1. Introduction

The human leukocyte antigen (HLA) complex, spanning ~3.6 Mb on the short arm of chromosome 6, includes genes coding for essential cell surface proteins mediating adaptive immune responses. Nucleotide variants in these genes play a critical role in transplantation [1], and have been associated with risk or protection to many human diseases and syndromes, including autoimmune disorders [2–4], cancers [5,6], viral infections [7], and drug hypersensitivities [8,9]. Antigen presenting HLA genes are typically extremely polymorphic; currently over 20,000 class I and class II HLA alleles are catalogued in the IPD-IMGT/HLA Database release 3.34.0 [10]. It has been speculated that this extreme allelic diversity has provided human populations with survival advantage against pathogens [11], and the relative frequency of different HLA alleles in different populations is maintained by balancing selection [12]. Although the high allelic diversity of HLA genes poses a challenge in finding matched donors for hematopoietic stem cell transplantation (HSCT), this feature can be exploited to examine genetic diversity and migration patterns. HLA allele and haplotype frequencies as well as linkage disequilibrium (LD) patterns tend to differ

Abbreviations: HLA, human leukocyte antigen; NGS, next-generation sequencing; AF, allele frequency; HF, haplotype frequency; LD, linkage disequilibrium; CEH, conserved extended haplotypes

* Corresponding author at: Department of Pathology, Stanford University School of Medicine, 3155 Porter Drive, Palo Alto, CA 94304, USA.
E-mail address: lcreary@stanford.edu (L.E. Creary).

https://doi.org/10.1016/j.humimm.2019.07.275
Received 24 December 2018; Received in revised form 21 June 2019; Accepted 6 July 2019
Available online 22 July 2019

significantly across populations as shown by the numerous HLA population studies reported over the years [13–25]. However, the type and accuracy of the genotyping method used to characterize genetic variants greatly impacts the findings of the population parameters examined. HLA can be defined at the serological (which often corresponds to the first field allele group) or allelic resolution level which is comprised of three fields; 2-field denotes the specific HLA protein encoded, 3-field denotes synonymous nucleotide substitutions within the coding sequence, 4-field refers to variant differences in noncoding regions. Maximum 4-field resolution is optimal as it captures allelic variants located within introns and the 5′ and 3′ untranslated regions, which have largely remained unexplored but may play an additional role in protein expression and structural variation [26,27]. These markers provide an additional level of diversity that could increase the efficiency of unrelated donor selection and further enhance our understanding of global patterns of HLA variation as well as the relationship between HLA and pathogen-mediated selection. Advancements in high-throughput and high-resolution genotyping technology led to the development of next-generation sequencing (NGS). The inherent attributes of NGS namely massive parallel sequencing of extended clonal fragments, typically reduces most of the allele and phase ambiguities often encountered with traditional HLA molecular typing methods [28–32]. Consequently, NGS is fast becoming the optimal method for achieving accurate, cost-efficient, unambiguous 4-field HLA typing in the clinical laboratory, as well as for disease association and anthropological studies.

The European American population describes individuals who are residents of the United States of America (USA) who can trace their ancestry back to Europe. Migration from Europe to the USA occurred in three major waves during the period from the 16th century to the mid-1920s. The majority of the initial settlers (16th~18th centuries) were largely from the British Isles [33]. The years between 1840s and 1850s witnessed the second wave of European immigrants mainly from Ireland, Germany and Scandinavia. Whilst during the third wave from the mid-1890s to mid-1920s European immigrants arrived primarily from Eastern and Southern Europe [33]. Since then Europeans from the aforementioned countries as well as other regions such as Western Europe, continued to arrive in the USA over the years in ebbs and flows. According to the US Census b 2010 [34], 72.4% of Americans self-identified themselves as white, a loosely based category since it refers to individuals having origins from Europe, North Africa, and the Middle East. The rest of the USA is comprised of individuals with ancestry from Africa (12.6%), Native America and Alaska Native origin (0.9%), Asia (4.8%), Native Hawaiian and Other Pacific Islander (0.2%), two or more races (2.9%) and the 'some other race' category (6.2%). Hispanics, which refers to individuals with heritage from Spanish and Portuguese speaking countries in North, Central and South America, comprise 16.3% of the US population and could belong to any of the broad race groups. Recent genetic studies have shown that American individuals typically carry DNA inherited from multiple continental populations, due to the mating of individuals from previously isolated and genetically differentiated populations, referred to as population admixture. In genetic studies European Americans are often used as a proxy for Europeans, however genetic admixture studies have revealed that the ancestry of European Americans consists of a mixture of genetically distinct European populations (estimated to range from 84 to 98% European ancestry) as well as lesser contributions from non-European populations including Indigenous Americans (1–10%), Africans (2–6%), and Asians (~1.4%) [35–37]. Therefore, from a HLA perspective, it is important to fully characterize HLA genetic profiles of European Americans, as well as other ethnic groups in the USA, to ensure optimal patient and donor matching for HSCT and better control traits-association studies.

In this study, we analyzed the HLA genetic diversity of 2248 self-described European Americans in the *HLA-A*, *HLA-C*, *HLA-B*, *HLA-DRB1, HLA-DRB3/4/5*, *HLA-DQA1*, *HLA-DQB1*, *HLA-DPA1* and *HLA-DPB1* loci defined using a high-resolution NGS typing method [38]. We also describe novel alleles and haplotypes observed at the 4-field level of resolution. In addition, we present data on various linkage disequilibrium measurements for two-locus haplotypes, extended haplotype frequencies, and we draw inferences about balancing selection acting on HLA loci.

2. Materials and methods

2.1. Study population

The study population consisted of 2248 healthy unrelated individuals from the United States of America of self-reported European ancestry. All individuals were non-Hispanic who reported no self-history of personal chronic diseases or in their nuclear family. DNA samples from study participants were obtained at random from among samples housed at the University of California San Francisco (UCSF) DNA bank. The samples were originally acquired from the following sources; UCSF-DNA bank (n = 761), the Coriell Institute for Medical Research (n = 1,408), and the Michael J Fox Foundation (n = 79). Of the total 2248 individuals 1275 were female (56.7%) and 973 male (43.3%). The overall median age of the cohort was 52.0 years (range 15–95 years). The UCSF Institutional Review Board approved the study and all participants provided written informed consent.

2.2. HLA genotyping

HLA typing was performed on genomic DNA extracted from blood samples using local protocols such as the salting-out method or the standard phenol/chloroform method. DNA samples were retrospectively typed at high-resolution for class I (*HLA-A*, *HLA-C*, *HLA-B*) and class II (*HLA-DRB1*, *HLA-DRB3*, *HLA-DRB4*, *HLA-DRB5*, *HLA-DQA1*, *HLA-DQB1, HLA-DPA1, HLA-DPB1*) HLA loci using the MIA FORA NGS high-throughput (HT) semi-automated typing protocol (Immucor, Inc., Norcross, GA, USA) and performed as previously described [21]. Biomek-NX^{P} and Biomek-FX^{P} liquid handling workstations (Beckman Coulter, Inc., Indianapolis, IN, USA) were used to perform all the PCR and the majority of the post-PCR library operations respectively. Briefly, long-range PCR was used to amplify the entire genes of all class I loci (> 200 bp 5′UTR to 3′UTR ~200–400 bp), *HLA-DQA1* (~200 bp of the 5′UTR to ~200 bp of the 3′UTR), and *HLA-DQB1* (~70 bp of the 5′UTR to ~100 bp of the 3′UTR) genes. For the remaining class II loci key extended regions of the gene were amplified. For *HLA-DPA1* coverage ranged from exon 1 through to exon 4 and for *HLA-DPB1* from exon 2 to exon 4. All *HLA-DRB1/3/4/5* genes were co-amplified in two separate reactions. The coverage for *HLA-DRB1/3/4* loci included ~300–500 bp of the 5′UTR to the first ~270 bp of intron 1 and from the end of intron 1 (~250 bp) to exon 6. For the *HLA-DRB5* gene exon 2 to exon 6 were amplified. Each PCR contained 100 ng genomic DNA and a solution of PCR master mix consisting of a cocktail of enzymes, buffers, and primers specific for each HLA locus. The thermal cycling parameters for all genes were as follows; initial denaturation 94 °C/30 s, followed by 15 cycles at 94 °C/1 min 15 s, 60 °C/30 s, 66 °C/7 min 30 s, followed by 20 cycles at 94 °C/30 s, 60 °C/30 s, 66 °C/7 min 30 s, ending with a final extension step at 66 °C for 10 min. PCR's were performed using Veriti Thermal Cyclers (Applied Biosystems/Thermo Fisher Scientific, Waltham, MA, USA). Amplicons were quantified using a Pico-Green assay (Invitrogen/Thermo Fisher Scientific, Waltham, MA, USA) using a Victor X plate reader (Perkin Elmer, Waltham, MA, USA). Amplicons for all genes were pooled in optimal molar amounts, and purified using AMPure XP beads (Beckman Coulter, Fullerton, CA, USA). Barcoded samples libraries were prepared as follows: enzymatic cleavage into 300–500 bp fragments, followed by enzymatic end repair to remove dNTPs overhangs and incorporation of deoxynucleotide dAMP to blunt ended 3′ ends, purification using AMPure XP, followed by ligation of a unique adaptor indices to each pooled sample. All adaptor

ligated samples were pooled into a single tube, purified using AMPure XP and DNA fragments were size selected for 400–500 bp fragments using the Blue Pippin system (Sage Science, Inc., Beverly, MA, USA). The eluted sample was enriched by a short PCR cycle using Illumina primers that contain adaptor sequences required for binding to the surface of the Illumina flow cell, purified using AMPure XP and quality checked using the Agilent 2200 TapeStation instrument (Agilent Technologies, Inc., Santa Clara, CA, USA). The sample was quantified using the Qubit™ dsDNA BR Assay Kit (ThermoFisher Scientific, Waltham, MA, USA) with the Qubit Fluorometer. The sample was denatured with sodium hydroxide, and sequenced at a final concentration of 1.3 pM spiked with 0.2% PhiX Control v3 (Illumina, Inc., San Diego, CA, USA) on the Illumina NextSeq 500 or MiniSeq instruments using 150 cycle paired-end kits (Illumina, Inc., San Diego, CA, USA).

2.3. HLA sequence data analysis and genotype assignment

NGS sequence data stored as FASTQ files were uploaded into the MIA FORA FLEX v3.0 alignment software (Immucor, Norcross, GA, USA) for analyses and assignment of HLA genotypes. The MIA FORA software demultiplexes FASTQ files according to each unique index and uses two complementary informatic strategies; competitive mapping of paired-end sequence reads and *de novo* assembly of paired-end reads to construct one or two phased consensus sequences. Paired-end reads and consensus sequences were compared with three sources of HLA reference sequences included in the MIA FORA software; (i) the IPD-IMGT/HLA Database release 3.25.0 [10] (https://www.ebi.ac.uk/ipd/imgt/hla/), (ii) internal MIA FORA HLA references generated by cloning and sequencing and (iii) internal MIA FORA *in silico* HLA sequences, which describes IPD-IMGT/HLA reference sequences containing partial exon sequences that have been filled with the closest complete exon sequence. These additional reference sequences were included in the MIA FORA database to further evaluate areas with low reads due to lack of coverage since the reference sequence for the segment would be missing for some alleles. HLA alleles were assigned with the aid of internal reference sequences, but the final calls were made using reference sequences from the IPD-IMGT/HLA Database. Final HLA genotypes were assigned after manual review of automatic genotype calls and sequence data. HLA genotype data was submitted to the 17th International HLA and Immunogenetics (IHIW) 'Study of unrelated subjects by NGS HLA' component for population comparative analyses.

2.4. Coding of allele calls: HLA-DRB1~HLA-DRB3/4/5 haplotypes

The *HLA-DRB3*, *HLA-DRB4*, and *HLA-DRB5* loci exist on specific *HLA-DRB1* haplotypes that display structural variation [39]. *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, and*HLA-DRB5* is found on haplotypes bearing *HLA-DRB1*15, 16* alleles. In general, if the *HLA-DRB1*01, 08, and 10* alleles are present *HLA-DRB3/4/5* loci are absent. Each individual may contain zero, one, or two copies of a *HLA-DRB3/4/5* allele. These associations have been well-characterized particularly in individuals of European ancestry. However, exceptions do exist in non-European populations such as *HLA-DRB5*01~HLA-DRB1*01* and *HLA-DRB5*absent~HLA-DRB1*15,* haplotypes in African descent groups [21]. A haplotype bearing *HLA-DRB1*08:01~HLA-DRB3*02:02* has been observed in a donor of Northern European ancestry at the Stanford Blood Center Histocompatibility and Immunogenetics clinical laboratory (unpublished observation). This haplotype was identified using the NGS method described in this manuscript, and confirmed using Luminex-based reverse sequence-specific oligonucleotide probes (rSSOP) and Sanger sequence-based typing (SBT).

The absence of *HLA-DRB3/4/5* loci, due to sequence variants, on chromosomes were assigned *HLA-DRB*00:00* and were counted when calculating allele frequencies; if blank alleles are not considered the frequencies of alleles at *HLA-DRB3/4/5* loci will be over-estimated.

2.5. Ambiguity group assignments

The MIA FORA software permits detailed examination of all sequence segments, as well as unambiguous allele assignment, with the exception of short tandem repeat (STR) enriched regions located within introns of some class II genes. These STR regions consists of mononucleotides (homopolymer tracts) and/or dinucleotides that are repeated typically ~10 to ≥20 times and cannot be assessed accurately by the sequencing methodology. In order to standardize alleles that are indistinguishable due to STRs, alleles were assigned to groups and were given the suffix SG (denoting STR Group) to the lowest numbered allele in that group. For example *HLA-DQA1*01:02:01:01SG* denotes the *HLA-DQA1*01:02:01:01/HLA-DQA1*01:02:01:03/HLA-DQA1*01:02: 01:05 STR* ambiguous group. SG alleles have recently been reported by Creary and colleagues [2]; this manuscript extends those findings. Detailed characteristics of STRs for each SG and ambiguous allele pairs due to unsequenced regions are described in Supplementary Table 1.

2.6. Population statistical analyses

Our main objectives were to utilize multi-locus HLA genotype data to estimate allele and haplotype (2-locus and extended) frequencies, identify deviation from Hardy-Weinberg equilibrium (HWE), localize patterns of locus (global) and allele level pairwise linkage disequilibrium (LD), and analyze selective pressures acting on loci. The majority of these tasks were accomplished using the *Py*thon for Population Genomics (PyPop) v0.7.0 software package (http://www.pypop.org) [40]. Overall deviation from HWE proportions for allele frequencies at each HLA locus were evaluated using Guo and Thompson's exact method [41] and by $\chi 2$ test when expected values were equal or greater than 5. In addition, $\chi 2$ tests were used to examine HWE deviations for various classes of genotypes; common genotypes, all homozygotes, all heterozygotes, 'common' heterozygotes for a specific allele.

The Ewens-Watterson's (EW) homozygosity test of neutrality [42,43] was applied to each HLA locus. The EW test compares whether the observed homozygosity (F, the sum of the squared allele frequencies) is significantly different from the mean value of homozygosity expected for a population of the same sample size (2n) with the same number of unique alleles (k) undergoing neutral evolution. For each locus the normalized deviate of F (F_{nd}, the difference between the observed homozygosity and the expected homozygosity divided by the square root of the variance of the expected homozygosity) was calculated. Significant negative F_{nd} values correspond to lower observed F values than expected mean F values, suggestive of balancing selection. Significant positive F_{nd} values indicate a skewed distribution towards specific alleles and are interpreted as the signature of directional selection or as an extreme demographic effect e.g. population bottleneck [43]. F_{nd} values equal to zero correspond to the mean value of homozygosity predicted by the EW model. The null hypotheses of the Ewens-Watterson test is neutral evolution ($F_{nd} = 0$). For a one-tailed test of the null hypothesis probability (p) values less than 0.05 are considered significant at the 5% level. For a two-tailed test against the alternatives of balancing or directional selection, p-values less than 0.025 or greater than 0.975 are considered significant at the 5% level.

Overall normalized measures of LD for locus-pairs were computed using D' [44], and Cramér's V statistic (also referred to as W_n) [45]. The standard D' and W_n values range from 0 to 1, a zero value refers to linkage equilibrium (the association of alleles is random), and positive values indicate association; values > 0.8 suggests very strong association. The significance of overall LD was computed based on 1000 permutations of the log-likelihood ratio test. The asymLD v0.1 R package was used to compute conditional asymmetric LD (cALD) between two

HLA loci. The cALD parameter describes the level of allele variation at locus 1 given the presence of specific alleles at locus 2. cALD is computed using the formula given by Thomson and Single [46]:

$$W_{loc1/loc2}{}^2 = (F_{loc1/loc2} - F_{loc1})/(1 - F_{loc1})$$

Where $W_{loc1/loc2}$ is the correlation coefficient (ALD statistic) for alleles at locus 1 conditioned on alleles at locus 2, inversing the locus subscripts in the formula would give $W_{loc2/loc1}$ that is a measure for alleles at locus 2 conditioned on variation at locus 1. When the relationship is symmetry i.e. $W_{loc1/loc2}$ and $W_{loc2/loc1}$ are equal to 1 meaning complete correlation between both loci, the ALD measures are equal to W_n.

In order to improve the integrity of the extended haplotypes estimated, we omitted any individuals from the dataset who had more than three loci with missing allele data. Extended haplotypes encompassing 6, 9, and 11 HLA loci were estimated using the expectation-maximization (EM) algorithm implemented in the Bridging Immuno Genomic Data Analysis Workflow Gaps (BIGDAWG) v1.8 package [47].

3. Results

3.1. Genotype frequencies and Hardy-Weinberg equilibrium

Hardy-Weinberg exact tests were performed on each of the three class I and six class II HLA loci: *HLA-DRB3*, *HLA-DRB4*, and *HLA-DRB5* were treated as a single locus denoted as *HLA-DRB3/4/5*. The distribution of genotype frequencies at, *HLA-A, HLA-DRB1*, *HLA-DQA1*, and *HLA-DQB1* loci met HWE expectations (Table 1).

A moderate deviation from HWE expectations was detected at *HLA-DPA1* ($P = 0.03$), and significant deviations were observed at *HLA-C* ($P = 0.02$), *HLA-B* ($P = 0.01$), *HLA-DRB3/4/5* ($P < 0.0001$), and *HLA-DPB1* ($P = 0.01$). HWE deviations were mostly attributable to an excess of specific heterozygous and rare genotypes (*HLA-C,* and *HLA-B*), specific homozygous genotypes (*HLA-DPA1*), specific heterozygous genotypes (*HLA-DPB1*), and approximately equal combination of distinct homozygous and heterozygous genotypes (*HLA-DRB3/4/5*).

With exception of *HLA-DP* loci, the observed deviations from HWE expectations result from rare genotypes, therefore the estimations for the less common haplotypes may not be accurate. All HWE results generated from PyPop are shown in Supplementary materials.

For $\chi 2$ tests to evaluate HWE deviation for classes of genotypes (i.e. common, all homozygous, and all heterozygous), deviations were observed for *HLA-A*, and *HLA-C* (common genotypes, $P = 0.03$), as well as *HLA-DPA1*, and *HLA-DRB3/4/5* where both common ($P = 0.004$, 0.0002) and all homozygotes ($P = 0.004$, < 0.00001) genotypes deviated significantly from HWE. The *HLA-B* locus had the highest observed heterozygosity index (0.945) whilst the lowest level was observed at the *HLA-DPB1* locus (0.768).

The HWE exact test was calculated for observed alleles reduced to 2-field allele resolution (results not shown). In this scenario HWE deviations were observed at three loci; *HLA-A* ($P = 0.02$), *HLA-B* ($P = 0.01$), and *HLA-DPB1* ($P = 0.02$). HWE deviations at *HLA-A* and *HLA-B* were due to an excess of particular heterozygous genotypes, whilst excess homozygotes accounted for the HWE deviation at *HLA-DPB1*. However for the remaining sections of this manuscript alleles defined at the maximum resolution are reported and evaluated.

Table 1
Hardy-Weinberg equilibrium *P* values and heterozygosity of classical class I and II HLA loci in an European American population.

Locus	2n[a]	*k*[b]	H_O[c]	H_E[d]	*P* value[e]
HLA-A	4494	63	0.869	0.871	0.05
HLA-C	4438	67	0.933	0.933	**0.02**
HLA-B	4466	93	0.945	0.946	**0.01**
HLA-DRB1	4426	49	0.926	0.924	0.14
HLA-DRB3/4/5	4496	23	0.849	0.868	**0.00**
HLA-DQA1	4352	33	0.921	0.913	0.63
HLA-DQB1	4334	33	0.925	0.916	0.24
HLA-DPA1	4494	20	0.814	0.838	**0.03**
HLA-DPB1	4492	41	0.768	0.778	**0.01**

Loci that deviated significantly ($P < 0.05$) from HWE are shown in boldface type.
[a] The total chromosome count.
[b] The number of unique alleles identified.
[c] Observed heterozygosity index.
[d] Expected heterozygosity index under Hardy-Weinberg equilibrium (HWE).
[e] Probability values for Guo and Thompson HWE tests.

3.2. Allele frequencies

3.2.1. Class I loci

Overall 223 distinct class I alleles were identified during the study; 63 *HLA-A* alleles, 67 *HLA-C* alleles, and 93 *HLA-B* alleles. The frequencies of the class I alleles are summarized in Table 2.

3.2.1.1. HLA-A. At the *HLA-A* locus *HLA-A*02:01:01:01* was the most common allele observed at a frequency greater than 10%; allele frequency (AF) = 26.0%, followed by *HLA-A*01:01:01:01* (AF = 16.1%) and *HLA-A*03:01:01:01* (AF = 13.4%). The greatest diversity was observed in the *HLA-A*02* allele family, 17 unique alleles detected in 1256 chromosomes (2n), corresponding to 27.9% of the overall AF. The second greatest level of diversity was observed in the *HLA-A*24* group (8 alleles, AF = 8.8%, 2n = 394) and most of the variation could be attributed to 3 types of HLA-A*24:02 4-field alleles; *HLA-A*24:02:01:01* (AF = 7.8%), *HLA-A*24:02:01:05* (AF = 0.4%), and *HLA-A*24:02:01:04* (AF = 0.3%).

Other *HLA-A* allele families exhibiting notable levels of allelic diversity was observed in *HLA-A*03* (7 alleles, AF = 14.5%, 2n = 651), which was mostly due to *HLA-A*03:01* 4-field differences (*HLA-A*03:01:01:01*, AF = 13.4%; *HLA-A*03:01:01:05*, AF = 0.4%; *HLA-A*03:01:01:03*, AF = 0.2%. Five *HLA-A*01* alleles represented 16.2% AF detected in 728 chromosomes.

3.2.1.2. HLA-B. As expected, of all the eleven loci typed in this study *HLA-B* was the most polymorphic. At the *HLA-B* locus only two alleles with frequencies above 10% were detected; *HLA-B*07:02:01* (AF = 13.0%), and *HLA-B*08:01:01:01* (AF = 10.5%).

The *HLA-B*15* allele family was the most diverse consisting of 16 alleles, (AF = 7.0%, 2n = 313), followed by *HLA-B*35* (9 alleles, AF = 9.9%, 2n = 440), and *HLA-B*44* (8 alleles, AF = 13.5%, 2n = 601).

Unlike *HLA-A* the vast majority of diversity at *HLA-B* could be explained by the presence of both rare alleles and alleles that differed across the antigen recognition site (2-field alleles that code for specific HLA proteins), although some similar 4-field alleles were observed. For example *HLA-B*44:02:01:01* (AF = 7.0%) and *HLA-B*44:02:01:03* (AF = 0.9%), also *HLA-B*44:03:01:01* (AF = 4.6%) and *HLA-B*44:03:01:02* (AF = 0.1%). Two non-expressed *HLA-B* alleles were identified, both on single occasions, in the entire cohort; *HLA-B*51:11N* and *HLA-B*57:79N*.

3.2.1.3. HLA-C. The most frequent *HLA-C* alleles detected ≥ 10% were *HLA-C*07:01:01:01* (AF = 14.2%), and *HLA-C*07:02:01:03* (AF = 12.4%).

The *HLA-C*07* allele group was the most diverse representing 11 distinct alleles (AF = 31%), of this 6 alleles were identified in 23 (*HLA-C*07:01:02*, AF = 0.5%) to 629 (*HLA-C*07:01:01:01*, AF = 14.2%) chromosomes. The remaining *HLA-C*07* alleles were observed only once or twice.

The *HLA-C*03* allele group comprised of 7 alleles represented a total AF of 12.9%, whereas 9 *HLA-C*15* alleles were detected but they only represented 2.8% of the total *HLA-C* allele frequency. The non-

Table 2
HLA class I allele frequencies observed in 2248 European Americans.

HLA-A	Count (2n)	Frequency	HLA-C	Count (2n)	Frequency	HLA-B	Count (2n)	Frequency
A*01:01:01:01	723	0.1609	C*01:02:01	142	0.0320	B*07:02:01	581	0.1301
A*01:01:01:03	2	0.0005	C*02:02:02:01	194	0.0437	B*07:02:45	1	0.0002
A*01:02	1	0.0002	C*02:02:02:02	7	0.0016	B*07:04	1	0.0002
A*01:06	1	0.0002	C*02:02:02:03	10	0.0023	B*07:05:01:01	15	0.0034
A*01:25	1	0.0002	C*02:07	1	0.0002	B*07:06:01	2	0.0005
A*02:01:01:01	1173	0.2610	C*02:10:01:01	3	0.0007	B*07:09	1	0.0002
A*02:01:01:05	4	0.0009	C*03:02:02:01	12	0.0027	B*08:01:01:01	467	0.1046
A*02:01:04	3	0.0007	C*03:02:02:02	2	0.0005	B*08:01:01:02	3	0.0007
A*02:01:05	2	0.0005	C*03:03:01:01	229	0.0516	B*13:02:01	104	0.0233
A*02:01:14Q	1	0.0002	C*03:03:07	1	0.0002	B*14:01:01	41	0.0092
A*02:02:01:01	2	0.0005	C*03:04:01:01	288	0.0649	B*14:02:01:01	123	0.0275
A*02:02:01:02	1	0.0002	C*03:04:01:02	39	0.0088	B*14:02:01:02	2	0.0005
A*02:05:01	53	0.0118	C*03:08	1	0.0002	B*15:01:01:01	245	0.0549
A*02:06:01:01	9	0.0020	C*04:01:01:01	307	0.0692	B*15:01:01:04	14	0.0031
A*02:07:01	1	0.0002	C*04:01:01:05	70	0.0158	B*15:03:01:02	2	0.0005
A*02:08	1	0.0002	C*04:01:01:06	132	0.0297	B*15:07:01	6	0.0013
A*02:17:02	1	0.0002	C*04:09N	7	0.0016	B*15:09	2	0.0005
A*02:24:01	1	0.0002	C*05:01:01:01	50	0.0113	B*15:16:01:02	3	0.0007
A*02:30:01	1	0.0002	C*05:01:01:02	305	0.0687	B*15:17:01:01	21	0.0047
A*02:34	1	0.0002	C*05:01:05	3	0.0007	B*15:17:02	1	0.0002
A*02:389	1	0.0002	C*05:05	1	0.0002	B*15:18:01:01	1	0.0002
A*02:66	1	0.0002	C*05:36	1	0.0002	B*15:18:01:02	11	0.0025
A*03:01:01:01	604	0.1344	C*05:37	1	0.0002	B*15:220	2	0.0005
A*03:01:01:02N	1	0.0002	C*06:02:01:01	319	0.0719	B*15:228	1	0.0002
A*03:01:01:03	8	0.0018	C*06:02:01:02	37	0.0083	B*15:29	1	0.0002
A*03:01:01:05	20	0.0045	C*06:02:01:03	23	0.0052	B*15:34	1	0.0002
A*03:01:01:06	1	0.0002	C*06:06	1	0.0002	B*15:35	1	0.0002
A*03:02:01	16	0.0036	C*06:12:70:101	1	0.0002	B*15:82	1	0.0002
A*03:26	1	0.0002	C*07:01:01:01	629	0.1417	B*18:01:01:01	54	0.0121
A*11:01:01:01	303	0.0674	C*07:01:02	23	0.0052	B*18:01:01:02	155	0.0347
A*23:01:01	91	0.0203	C*07:02:01:01	67	0.0151	B*18:03	2	0.0005
A*24:02:01:01	349	0.0777	C*07:02:01:03	551	0.1242	B*18:05	1	0.0002
A*24:02:01:02L	2	0.0005	C*07:04:01:01	76	0.0171	B*27:02:01	21	0.0047
A*24:02:01:04	12	0.0027	C*07:06	2	0.0005	B*27:05:02	153	0.0343
A*24:02:01:05	17	0.0038	C*07:15	1	0.0002	B*27:05:03	3	0.0007
A*24:02:02	1	0.0002	C*07:18	24	0.0054	B*27:07:01	2	0.0005
A*24:03:01:01	10	0.0022	C*07:38:01	1	0.0002	B*27:13	1	0.0002
A*24:22:601	1	0.0002	C*07:419	1	0.0002	B*35:01:01:01	4	0.0009
A*24:58	2	0.0005	C*07:46	1	0.0002	B*35:01:01:02	259	0.0580
A*25:01:01	95	0.0211	C*08:01:01	4	0.0009	B*35:01:01:02x1	1	0.0002
A*26:01:01:01	176	0.0392	C*08:02:01:01	117	0.0264	B*35:01:07	1	0.0002
A*26:08	9	0.0020	C*08:02:01:02	47	0.0106	B*35:02:01	57	0.0128
A*26:26	1	0.0002	C*08:03:01	6	0.0014	B*35:03:01	92	0.0206
A*29:01:01:01	19	0.0042	C*08:22	1	0.0002	B*35:08:01	24	0.0054
A*29:02:01:01	149	0.0332	C*12:02:02	52	0.0117	B*35:12:01	1	0.0002
A*29:02:01:02	9	0.0020	C*12:03:01:01	249	0.0561	B*35:41	1	0.0002
A*30:01:01	51	0.0114	C*12:05	1	0.0002	B*37:01:01	56	0.0125
A*30:02:01:01	41	0.0091	C*12:143	1	0.0002	B*38:01:01	116	0.0260
A*30:02:01:02	2	0.0005	C*14:02:01:01	58	0.0131	B*39:01:01:03	51	0.0114
A*30:04:01	3	0.0007	C*14:02:01:02	3	0.0007	B*39:06:02	27	0.0061
A*31:01:02:01	110	0.0245	C*15:02:01:01	91	0.0205	B*39:24:01	1	0.0002
A*32:01:01	172	0.0383	C*15:02:01:02	2	0.0005	B*39:35	1	0.0002
A*33:01:01	37	0.0082	C*15:04:01	1	0.0002	B*40:01:02	225	0.0504
A*33:03:01	14	0.0031	C*15:05:01	1	0.0002	B*40:02:01	60	0.0134
A*34:02:01	4	0.0009	C*15:05:02	19	0.0043	B*40:06:01:02	1	0.0002
A*66:01:01	18	0.0040	C*15:09	1	0.0002	B*41:01:01	23	0.0052
A*68:01:01:02	34	0.0076	C*15:11	1	0.0002	B*41:02:01	25	0.0056
A*68:01:02:01	14	0.0031	C*15:13	5	0.0011	B*42:02:01:02	1	0.0002
A*68:01:02:02	64	0.0142	C*15:87	1	0.0002	B*44:02:01:01	314	0.0703
A*68:01:02:03	1	0.0002	C*16:01:01:01	129	0.0291	B*44:02:01:03	39	0.0087
A*68:02:01:01	36	0.0080	C*16:01:01:02	2	0.0005	B*44:03:01:01	206	0.0461
A*69:01	11	0.0025	C*16:02:01	26	0.0059	B*44:03:01:02	3	0.0007
A*74:03	1	0.0002	C*16:04:01	7	0.0016	B*44:03:02	2	0.0005
			C*16:85	2	0.0005	B*44:04	5	0.0011
			C*17:01:01:02	1	0.0002	B*44:05:01	22	0.0049
			C*17:01:01:05	20	0.0045	B*44:27:01	10	0.0022
			C*17:03	27	0.0061	B*45:01:01	24	0.0054
						B*47:01:01:03	18	0.0040
						B*47:02	1	0.0002
						B*48:01:01	11	0.0025
						B*49:01:01	77	0.0172
						B*50:01:01	40	0.0090
						B*51:01:01:01	231	0.0517
						B*51:05	2	0.0005

(*continued on next page*)

Table 2 (*continued*)

HLA-A	Count (2n)	Frequency	HLA-C	Count (2n)	Frequency	HLA-B	Count (2n)	Frequency
						B*51:07:01	2	0.0005
						B*51:08:01	8	0.0018
						B*51:09:01	2	0.0005
						B*51:11N	1	0.0002
						B*51:22	1	0.0002
						B*52:01:01:01	2	0.0005
						B*52:01:01:02	51	0.0114
						B*53:01:01	15	0.0034
						B*55:01:01	86	0.0193
						B*55:01:03	1	0.0002
						B*56:01:01:02	10	0.0022
						B*56:01:01:03	21	0.0047
						B*57:01:01	145	0.0325
						B*57:02:01	1	0.0002
						B*57:03:01:02	3	0.0007
						B*57:79N	1	0.0002
						B*58:01:01:01	36	0.0081
						B*58:02:01	1	0.0002
						B*73:01	1	0.0002

Abbreviations: 2n, total chromosome count; x1, denotes novel exon variant.

expressed allele *HLA-C*04:09N* was detected at 7 chromosomes (AF = 0.2%).

3.2.2. Class II loci

The class II alleles exhibited lower levels of diversity compared to class I alleles. The allele frequencies of *HLA-DRB1/3/4/5, HLA-DQ*, and *HLA-DP* loci are listed in Tables 3–5 respectively.

3.2.2.1. HLA-DRB1, HLA-DRB3, HLA-DRB4, and HLA-DRB5. The *HLA-DRB1* locus was the most polymorphic of the class II loci with 49 distinct alleles detected. Seven alleles were ambiguous at the SG-level accounting for 52.9% of the total allele frequency. *HLA-DRB1*15:01:01:01SG* was the most frequent allele detected (AF = 12.8%), closely followed by *HLA-DRB1*07:01:01:01SG* (12.6%) and *HLA-DRB1*03:01:01:01SG* (11.7%).

At the *HLA-DRB3/4/5* combined locus 23 alleles were identified. The *HLA-DRB4*01:03:01:01/ HLA-DRB4*01:03:01:03* ambiguous allele pair was the most frequent at 20.6%, followed by *HLA-DRB3*02:02:01:02* (AF = 14.7%); the 4-field alternative form *HLA-DRB3*02:02:01:01* occurred less frequently at 4.6%. The most frequent *HLA-DRB5* allele detected was *HLA-DRB5*01:01:01* (AF = 12.8%).

3.2.2.2. HLA-DQA1 and HLA-DQB1. For *HLA-DQA1*, 33 alleles were detected; of this total 8 SG-level alleles represented 61.1% of the allele frequency. Alleles *HLA-DQA1*01:02:01:01SG*, *HLA-DQA1*02:01:01SG* and *HLA-DQA1*05:05:01:01SG* alleles were present at frequencies > 10%.

Of the total 33 *HLA-DQB1* alleles identified 4 alleles occurred at frequencies ≥10%; *HLA-DQB1*06:02:01* (AF = 12.3%), *HLA-DQB1*02:01:01* (AF = 11.5%), *HLA-DQB1*03:01:01:03* (AF = 10.7%), and *HLA-DQB1*03:02:01* (AF = 10.5%).

3.2.2.3. HLA-DPA1 and HLA-DPB1. For *HLA-DPA1*, 20 alleles were observed and 5 alleles, all 4-field variants of *HLA-DPA1*01:03:01*, contributed 81.7% of the total allele frequency; *HLA-DPA1*01:03:01:02* (28.2%), *HLA-DPA1*01:03:01:04* (18.4%), *HLA-DPA1*01:03:01:01* (13.2%), *HLA-DPA1*01:03:01:05* (11.7%), and *HLA-DPA1*01:03:01:03* (10.2%).

Forty-one *HLA-DPB1* alleles were detected, of this total 28 alleles were non-ambiguous, 12 alleles were ambiguous due to the inability to set cis–trans phase because of the low complexity region across intron 2 of the *HLA-DPB1* gene. Also 2 alleles were ambiguous due to unsequenced regions (*HLA-DPB1*02:01:02/HLA-DPB1*02:01:19* and *HLA-DPB1*13:01:01/HLA-DPB1*107:01*.

The most frequent allele detected was *HLA-DPB1*04:01:01:01* representing 42.6% of the total *HLA-DPB1* allele frequency. Interestingly this allele was also the most common in the entire cohort. A graphical representation of the cumulative allele frequencies at all class I and class II HLA loci is depicted in Fig. 1.

3.2.3. Allele frequency classification

We classified the allele frequencies found in this study according to the Common and Well-Documented catalogue criteria [48]. The designation of common alleles refer to alleles occurring at frequencies ≥0.001, well-documented alleles are observed at least five times in unrelated individuals, and rare alleles observed less than five times in unrelated individuals. Novel alleles describe exon variants discovered in the present study that are not listed in the IPD-IMGT/HLA Database release 3.25.0.

Allele frequency classifications are summarized in Table 6. The proportion of common alleles (n = 243) ranged from 49.5% at *HLA-B* to 72.7% at *HLA-DQA1* and *HLA-DQB1*. Well-documented alleles (n = 6) ranged from 1.1% at *HLA-B* to 6.1% at *HLA-DQA1*; well-documented alleles were not observed at *HLA-A*, *HLA-DRB1*, *HLA-DRB3/4/5*, and *HLA-DQB1* loci.

Rare alleles were highest at *HLA-A* (49.2%) and lowest at *HLA-DQA1* (21.2%). Novel alleles were detected at *HLA-B*, *HLA-DRB3/4/5,* and *HLA-DQB1* loci in single individuals. In contrast a novel allele was identified forty times at the *HLA-DPA1* locus.

3.3. Novel alleles

During the study, we identified three different novel alleles containing exon variants, characteristics of which are summarized in Table 7. At the time of writing this manuscript the novel allele sequences were compared to sequences in the most current version, 3.34.0, of the IPD-IMGT/HLA Database, to check whether these allele sequences had been submitted by other groups and new HLA names had been assigned. A nonsynonymous variant in exon 1 of *HLA-B*35:01:01:02* was detected in one individual, this sequence was previously submitted by another group and has been named as *HLA-B*35:347*. Synonymous substitutions in exon 3 and exon 1 of *HLA-DRB3*02:02:01:02 and HLA-DQB1*05:01:01:03* were identified on two chromosomes in 2 individuals; these alleles still remains novel. These sequences were submitted to GenBank (https://www.ncbi.nlm.nih.gov/genbank) and the IPD-IMGT/HLA Database. The sequence variants have been assigned new names. They are *HLA-DRB3*02:02:20* and *HLA-DQB1*05:01:33*. Two synonymous and one nonsynonymous

Table 3
HLA-DRB1 and *HLA-DRB3/4/5* allele frequencies observed in 2248 European Americans.

HLA-DRB1	Count (2n)	Frequency	HLA-DRB3/4/5	Count (2n)	Frequency
DRB1*01:01:01	386	0.0872	DRB*00:00	667	0.1484
DRB1*01:01:04	1	0.0002	DRB3*01:01:02:01/DRB3*01:01:02:02‡	645	0.1435
DRB1*01:02:01	63	0.0142	DRB3*01:16	2	0.0004
DRB1*01:03	51	0.0115	DRB3*02:02:01:01	205	0.0456
DRB1*03:01:01:01SG	516	0.1166	DRB3*02:02:01:02	662	0.1472
DRB1*03:04:01	1	0.0002	DRB3*02:02:01:02x1	1	0.0002
DRB1*04:01:01:01SG	391	0.0883	DRB3*02:02:06	1	0.0002
DRB1*04:02:01	67	0.0151	DRB3*02:16	1	0.0002
DRB1*04:03:01	41	0.0093	DRB3*02:24	11	0.0025
DRB1*04:04:01	160	0.0362	DRB3*03:01:01	210	0.0467
DRB1*04:05:01	23	0.0052	DRB4*01:01:01:01	239	0.0532
DRB1*04:06:02	1	0.0002	DRB4*01:02	3	0.0007
DRB1*04:07:01	44	0.0099	DRB4*01:03:01:01/DRB4*01:03:01:03‡	928	0.2064
DRB1*04:08:01	22	0.0050	DRB4*01:03:01:02N	154	0.0343
DRB1*04:11:01	1	0.0002	DRB4*01:03:02	52	0.0116
DRB1*07:01:01:01SG	558	0.1261	DRB4*01:03:03	6	0.0013
DRB1*08:01:01	95	0.0215	DRB5*01:01:01	576	0.1281
DRB1*08:02:01	1	0.0002	DRB5*01:01:02	1	0.0002
DRB1*08:03:02	15	0.0034	DRB5*01:02	48	0.0107
DRB1*08:04:01	8	0.0018	DRB5*01:05	2	0.0004
DRB1*08:10	2	0.0005	DRB5*01:20	1	0.0002
DRB1*09:01:02	39	0.0088	DRB5*02:02	80	0.0178
DRB1*10:01:01:01	39	0.0088	DRB5*02:03	1	0.0002
DRB1*11:01:01:01	250	0.0565			
DRB1*11:01:02	2	0.0005			
DRB1*11:02:01	7	0.0016			
DRB1*11:03:01	29	0.0066			
DRB1*11:04:01	182	0.0411			
DRB1*11:13:02	1	0.0002			
DRB1*11:36	1	0.0002			
DRB1*11:43	1	0.0002			
DRB1*12:01:01:03	90	0.0203			
DRB1*13:01:01:01SG	257	0.0581			
DRB1*13:02:01	208	0.0470			
DRB1*13:03:01	50	0.0113			
DRB1*13:05:01	18	0.0041			
DRB1*13:10	1	0.0002			
DRB1*14:01:01	11	0.0025			
DRB1*14:03:01	1	0.0002			
DRB1*14:04:01	6	0.0014			
DRB1*14:07:01	2	0.0005			
DRB1*14:54:01	85	0.0192			
DRB1*15:01:01:01SG	568	0.1283			
DRB1*15:02:01:01SG	48	0.0109			
DRB1*15:02:02	1	0.0002			
DRB1*15:03:01:01SG	1	0.0002			
DRB1*15:14	1	0.0002			
DRB1*16:01:01	71	0.0160			
DRB1*16:02:01:02	9	0.0020			

Abbreviations: 2n, total chromosome count; x1, denotes novel exon variant; SG, short tandem repeat (STR) allele ambiguity group; ‡ Ambiguous pairs of alleles due to unsequenced regions. Further details are summarized in Supplementary Table 1.

substitutions located within exon 3 defines a novel *HLA-DPA1* allele detected at 40 chromosomes in 40 individuals. The closest allele, according to nucleotide sequence similarity, was *HLA-DPA1*02:02:01*. Not surprisingly, this fairly common novel allele was found to have been submitted by several other groups, identified in both European and non-European ancestry individuals, to the IPD-IMGT/HLA Database. The allele has been named as *HLA-DPA1*02:07* in which there are three types of 4-field variants; *HLA-DPA1*02:07:01:01*, *HLA-DPA1*02:07:01:02*, and *HLA-DPA1*02:07:01:03*. In this study allele *HLA-DPA1*02:07:01:01* was identified in all 40 individuals.

3.4. Ewens-Watterson homozygosity test

Allele frequency distributions at *HLA* loci that showed no statistically significant deviation from Hardy-Weinberg expectations were evaluated by the normalized deviate of the Ewens-Watterson homozygosity statistic (F_{nd}). The results are summarized in Table 8. At *HLA-DQA1*, and *HLA-DQB1* F_{nd} values were negative and differed significantly ($P = 0.015$ and 0.009 respectively) from the expectation of neutral evolution ($F_{nd} = 0$) in the direction of balancing selection.

3.5. Comparison of allele frequencies to the National Marrow donor Program registry

We compared the observed allele frequencies from the European American sample with the allele frequencies of the USA 'Be the Match' hematopoietic cell donor registry managed by the National Marrow Donor Program (NMDP). The NMDP donor registry is one of the largest in the world and has collated details from more than 16 million adult stem cell donors and cord blood units, of which ~12.3 million (77%) are white donors (https://bethematch.org). The majority of the HLA alleles listed in the donor registry are defined at 1–2 field resolution or classified by G groups, pertaining to groups of alleles for a given HLA gene with identical nucleotide sequences across the antigen-recognition site; exons 2 and 3 for class I loci and exon 2 for class II loci. For comparative purposes, we converted the observed European-American

Table 4
HLA-DQA1 and *HLA-DQB1* allele frequencies observed in 2248 European Americans.

HLA-DQA1	Count (2n)	Frequency	HLA-DQB1	Count (2n)	Frequency
DQA1*01:01:01:02SG	401	0.0921	DQB1*02:01:01	499	0.1151
DQA1*01:01:02	63	0.0145	DQB1*02:02:01:01	394	0.0909
DQA1*01:02:01:01SG	566	0.1301	DQB1*02:02:01:02	7	0.0016
DQA1*01:02:01:04SG	198	0.0455	DQB1*02:26	1	0.0002
DQA1*01:02:02	86	0.0198	DQB1*03:01:01:01	376	0.0868
DQA1*01:02:04	1	0.0002	DQB1*03:01:01:02	54	0.0125
DQA1*01:03:01:01	45	0.0103	DQB1*03:01:01:03	464	0.1071
DQA1*01:03:01:02SG	255	0.0586	DQB1*03:02:01	453	0.1045
DQA1*01:04:01:01SG	93	0.0214	DQB1*03:02:02	2	0.0005
DQA1*01:04:01:03	6	0.0014	DQB1*03:03:02:01	143	0.0330
DQA1*01:04:02	4	0.0009	DQB1*03:03:02:02/DQB1*03:03:02:03†	38	0.0088
DQA1*01:05:01	40	0.0092	DQB1*03:04:01	10	0.0023
DQA1*01:05:02	1	0.0002	DQB1*03:05:01	9	0.0021
DQA1*01:06	1	0.0002	DQB1*03:12	1	0.0002
DQA1*01:07Q	1	0.0002	DQB1*03:19:01	8	0.0019
DQA1*01:10	5	0.0012	DQB1*03:22	1	0.0002
DQA1*02:01:01:01SG	545	0.1252	DQB1*04:02:01	107	0.0247
DQA1*03:01:01	427	0.0981	DQB1*05:01:01:01	63	0.0145
DQA1*03:02	38	0.0087	DQB1*05:01:01:02	41	0.0095
DQA1*03:03:01:01	341	0.0784	DQB1*05:01:01:03	402	0.0928
DQA1*04:01:01	81	0.0186	DQB1*05:01:01:03x1	1	0.0002
DQA1*04:01:02:01	7	0.0016	DQB1*05:02:01	95	0.0219
DQA1*04:01:02:02	1	0.0002	DQB1*05:03:01:01/DQB1*05:03:01:02†	102	0.0235
DQA1*04:02	8	0.0018	DQB1*05:04	6	0.0014
DQA1*05:01:01:01	54	0.0124	DQB1*06:01:01	46	0.0106
DQA1*05:01:01:02	427	0.0981	DQB1*06:01:03	2	0.0005
DQA1*05:01:01:03	19	0.0044	DQB1*06:02:01	534	0.1232
DQA1*05:03	7	0.0016	DQB1*06:03:01	281	0.0648
DQA1*05:05:01:01SG	500	0.1149	DQB1*06:04:01	141	0.0325
DQA1*05:05:01:05SG	99	0.0228	DQB1*06:07:01	1	0.0002
DQA1*05:08	1	0.0002	DQB1*06:09:01	50	0.0115
DQA1*05:09	13	0.0030	DQB1*06:16	1	0.0002
DQA1*06:01:01	18	0.0041	DQB1*06:39	1	0.0002

Abbreviations: 2n, total chromosome count; x1, denotes novel exon variant; SG, short tandem repeat (STR) allele ambiguity group; † Ambiguous pairs of alleles due to unsequenced regions. Further details are summarized in Supplementary Table 1.

allele frequencies equal or greater than 1% at *HLA-A*, *HLA-C*, *HLA-B*, *HLA-DRB1,* and *HLA-DQB1* loci to G-group level and 2-field alleles then compared with the corresponding allele frequencies listed in the NMDP registry [17] for the USA white population (Fig. 2). The Pearson's correlation coefficient was above 0.95 for *HLA-DQB1* indicating a close similarity of European-American *HLA-DQB1* allele frequencies and the NMDP. Correlation coefficient values were slightly lower for *HLA-A*, *HLA-C HLA-B*, and *HLA-DRB1* (r = > 0.86–0.90).

In general, the most frequent alleles at all 5 loci tended to be lower than those listed in the NMDP registry. For example, the observed AF for *HLA-A*02:01g* is 26.2% compared to 29.6% in the registry, *HLA-B*08:01g* observed AF is 10.5% the registry is 12.5%, and *HLA-DRB1*15:01* observed AF is 12.8%, whilst the registry AF is higher at 14.4%. Such discrepancies can be mostly explained by some biases inherent to the recruitment practices of the NMDP registry. The registry recruits donor's dependent on the need of patients requiring HSCT, and since most patients requiring treatment have particular HLA genotypes associated with the disease in question the donor pool would be enriched for those HLA genotypes. In other words, the registry may not represent an accurate picture of a randomly selected population. Consequently the ranking of alleles and haplotypes observed in the present European-American dataset differs slightly to those ranked by the registry. The ranks of the allele frequencies compared to the NMDP registry [17] are listed in Supplementary Table 2.

3.6. Linkage disequilibrium and 2-locus haplotypes

3.6.1. Global linkage disequilibrium

Various global linkage disequilibrium (LD) measures were computed to assess the strength of association between two loci. The LD values (D', W_n, $W_{loc1/loc2}$, $W_{loc2/loc1}$) of neighboring loci pairs and some non-neighboring loci pairs that were deemed to be in relative close proximity are summarized in Table 9. LD values correlate with the physical distance and the recombination fraction between the loci on the chromosome; typically larger LD values are associated with shorter distances. The strength of association was greatest for the *HLA-DRB1~HLA-DQA1* haplotype (D' = 0.974, W_n = 0.742), followed by *HLA-DQA1~HLA-DQB1* (D' = 0.970, W_n = 0.668) and *HLA-DRB1~HLA-DQB1* (D' = 0.962, W_n = 0.685). LD between *HLA-DRB1~HLA-DPB1* (D' = 0.334, W_n = 0.209), and *HLA-DQB1~HLA-DPB1*(D' = 0.314, W_n = 0.183) loci is relatively weak and is due to the well-known recombination breakpoint between *HLA-DPB1* and other class II genes [49]. A LD plot illustrating global D' and W_n values for loci pairs are depicted in Fig. 3.

Interestingly the *HLA-C~HLA-B* loci pair that is often reported to have the greatest strength of association in both European [22] and non-European populations [18,22] was lower than expected when compared to previous studies and was ranked at number 3 by D' measure (0.938) and 4 by W_n (0.654). It is important to highlight that previous studies had often used 2-field HLA allele frequencies to assess LD parameters. Since the LD calculation is based on allele frequencies it is highly likely that there would be a notable difference in LD values generated using either 2- and 4-field data. To further explore this hypothesis we converted the observed alleles found in this study to 2-field and computed LD measures. Overall higher LD values were observed for 4-field versus 2-field loci pairs with the exception of *HLA-DPA1~HLA-DPB1* where LD values were higher at 2-field (D' = 0.915) compared to 4-field (D' = 0.906). These observations can be explained by the greater diversity of the different types of 4-field alleles for the majority of loci. On the other hand, the alleles of the *HLA-DPA1~HLA-DPB1* haplotypes are less diverse and predominantly consist of '*HLA-DPA1*01:03:01*' and '*HLA-DPB1*04:01:01*' 4-field variant alleles. A reduction of these *HLA-*

Table 5
HLA-DPA1 and *HLA-DPB1* allele frequencies observed in 2248 European Americans.

HLA-DPA1	Count (2n)	Frequency	HLA-DPB1	Count (2n)	Frequency
DPA1*01:03:01:01	592	0.1317	DPB1*01:01:01	213	0.0474
DPA1*01:03:01:02	1268	0.2822	DPB1*01:01:02	4	0.0009
DPA1*01:03:01:03	458	0.1019	DPB1*02:01:02/DPB1*02:01:19	620	0.1380
DPA1*01:03:01:04	825	0.1836	DPB1*02:01:04	2	0.0005
DPA1*01:03:01:05	527	0.1173	DPB1*02:02	33	0.0074
DPA1*01:03:04	2	0.0005	DPB1*03:01:01	376	0.0837
DPA1*01:03:05	1	0.0002	DPB1*04:01:01:01	1915	0.4263
DPA1*01:04	32	0.0071	DPB1*04:01:01:02	1	0.0002
DPA1*01:05	3	0.0007	DPB1*04:01:03	1	0.0002
DPA1*02:01:01:01	230	0.0512	DPB1*04:02:01:01	91	0.0203
DPA1*02:01:01:02	224	0.0498	DPB1*04:02:01:02	431	0.0960
DPA1*02:01:02	178	0.0396	DPB1*05:01:01	80	0.0178
DPA1*02:01:04	5	0.0011	DPB1*06:01:01	68	0.0151
DPA1*02:01:07	1	0.0002	DPB1*09:01:01	32	0.0071
DPA1*02:01:08	1	0.0002	DPB1*10:01:01	65	0.0145
DPA1*02:02:01x1	40	0.0089	DPB1*104:01	76	0.0169
DPA1*02:02:02	70	0.0156	DPB1*105:01	12	0.0027
DPA1*02:06	29	0.0065	DPB1*11:01:01	99	0.0220
DPA1*03:01	7	0.0016	DPB1*124:01	5	0.0011
DPA1*04:01x1	1	0.0002	DPB1*13:01:01/DPB1*107:01	63	0.0140
			DPB1*130:01	1	0.0002
			DPB1*131:01	1	0.0002
			DPB1*138:01	4	0.0009
			DPB1*14:01:01	87	0.0194
			DPB1*15:01:01	36	0.0080
			DPB1*16:01:01	30	0.0067
			DPB1*17:01	50	0.0111
			DPB1*19:01	31	0.0069
			DPB1*20:01:01	14	0.0031
			DPB1*23:01:01	32	0.0071
			DPB1*25:01	1	0.0002
			DPB1*259:01	1	0.0002
			DPB1*26:01:02	1	0.0002
			DPB1*34:01	4	0.0009
			DPB1*35:01:01	1	0.0002
			DPB1*350:01	3	0.0007
			DPB1*45:01	3	0.0007
			DPB1*52:01	1	0.0002
			DPB1*59:01	1	0.0002
			DPB1*81:01	2	0.0005
			DPB1*90:01	1	0.0002

Abbreviations: 2n, total chromosome count; x1, denotes novel exon variant; ǂ Ambiguous pairs of alleles due to unsequenced regions. Further details are summarized in Supplementary Table 1.

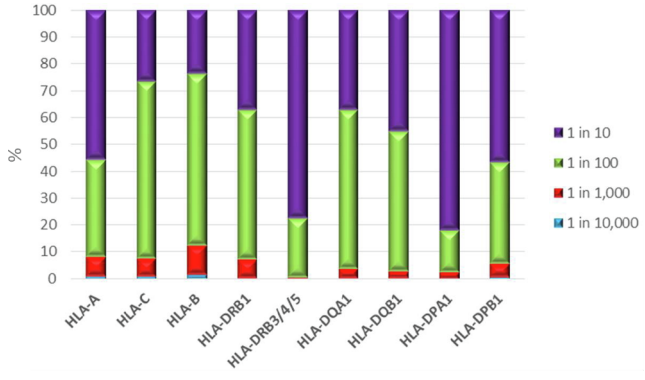

Fig. 1. Cumulative frequencies of class I and class II HLA alleles. The colored bars represent the cumulative frequencies of alleles observed at 1 in 10,000 (blue bar), 1 in 1,000 (red bar), 1 in 100 (green bar), and 1 in 10 (purple bar).

DP alleles to 2-field would lead to a predominance of *HLA-DPA1*01:03~HLA-DPB1*04:01* haplotypes, essentially a 1:1 ratio, leading to higher LD values at the 2-field.

The complementary pair of cALD values for $W_{HLA\text{-}C/HLA\text{-}B}$ and $W_{HLA\text{-}B/HLA\text{-}C}$ of 0.846 and 0.757 respectively indicates that there is more variation of *HLA-B* alleles compared to *HLA-C* alleles. A similar pattern

Table 6
Classification of class I and class II HLA allele frequencies observed in 2248 European Americans.

Locus	*k*	Common Alleles, n (%)	Well-Documented Alleles, n (%)	Rare Alleles, n (%)	Novel Alleles, n (%)
HLA-A	63	32 (50.8)	0 (0.0)	31 (49.2)	0 (0.0)
HLA-C	67	36 (53.7)	1 (1.5)	30 (44.8)	0 (0.0)
HLA-B	93	46 (49.5)	1 (1.1)	45 (48.4)	1 (1.1)
HLA-DRB1	49	33 (67.3)	0 (0.0)	16 (32.7)	0 (0.0)
HLA-DRB3/4/5	23	14 (60.9)	0 (0.0)	8 (34.8)	1 (4.3)
HLA-DQA1	33	24 (72.7)	2 (6.1)	7 (21.2)	0 (0.0)
HLA-DQB1	33	24 (72.7)	0 (0.0)	8 (24.2)	1 (3.0)
HLA-DPA1	20	12 (60.0)	1 (5.0)	6 (30.0)	1 (5.0)
HLA-DPB1	41	22 (53.7)	1 (2.4)	18 (43.9)	0 (0.0)
Total	422	243	6	169	4

Common alleles denotes allele frequencies (AF) ≥ 0.001; Well-Documented alleles represent alleles that have been observed at least five times in unrelated individuals; Rare alleles denotes alleles that have been observed less than five times in unrelated individuals ; Novel alleles refers to novel exon variants identified.

of cALD measures are observed for haplotypes *HLA-A~HLA-B*, *HLA-A~HLA-C*, and *HLA-DRB3/4/5~HLA-DRB1*. cALD measures for haplotypes *HLA-B~HLA-DRB1* and *HLA-DRB1~HLA-DQA1* indicate greater

Table 7
Novel alleles identified in the European American population.

Most similar allele	Nucleotide Substitution[a]	Gene region	Codon/Amino Acid change[b]	New allele name	GenBank accession number
HLA-B*35:01:01:02	GGG > TGG	Exon 1	−10/Gly > Trp	HLA-B*35:347	MF069209
HLA-DRB3*02:02:01:02	ACG > ACA	Exon 3	145/Thr > Thr	HLA-DRB3*02:02:20	MK297329
HLA-DQB1*05:01:01:03	GTA > GTG	Exon 1	−18/Val > Val	HLA-DQB1*05:01:33	MK297328
HLA-DPA1*02:02:01	GTG > ATG	Exon 3	91/Val > Met	HLA-DPA1*02:07:01:01	KP774801
	CCA > CCG		127/Pro > Pro		
	GTA > GTG		154/Val > Val		

[a] Nucleotide of the previously reported allele is listed first, differences are underlined.
[b] Amino acid encoded by the previously reported allele is shown first.

Table 8
The Ewens-Watterson homozygosity test for selective neutrality at the *HLA-A, HLA-DRB1*, *HLA-DQA1,* and *HLA-DQB1* loci in European Americans.

Locus	Observed *F*	Expected *F*	Normalized deviate of *F* (F_{nd})	*P*-value
HLA-A	0.129	0.088	1.464	0.923
HLA-DRB1	0.076	0.114	−0.967	0.100
HLA-DQA1	0.088	0.171	−1.254	**0.015**
HLA-DQB1	0.084	0.171	−1.307	**0.009**

Abbreviations: *F,* homozygosity statistic derived from the sum of the squared allele frequencies; *P*, probability value of obtaining *F* under neutral evolution that is less than or equal to the observed *F statistic*. Loci that are significant at the 0.05 level are shown in boldface type.

diversity of *HLA-B* and *HLA-DRB1* alleles compared to *HLA-DRB1* and *HLA-DQA1* alleles respectively. cALD values for $W_{HLA\text{-}DRB1/HLA\text{-}DQA1}$ (0.864) and $W_{HLA\text{-}DQA1/HLA\text{-}DRB1}$ (0.929) suggests more variation at *HLA-DRB1* than *HLA-DQA1* loci. For *HLA-DRB1~HLA-DQB1*, *HLA-DRB1~HLA-DPB1*, *HLA-DQA1~ HLA-DQB1, HLA-DQB1~HLA-DPB1* and *HLA-DPA1~ HLA-DPB1* haplotypes, the complementary cALD values are very similar suggesting no significant variation of alleles at each locus.

3.6.2. Allele-level linkage disequilibrium

A total of 600 *HLA-A~HLA-B*, 514 *HLA-A~HLA-C*, 244 *HLA-C~HLA-B*, 645 *HLA-B~HLA-DRB1*, *95* HLA-DRB1~HLA-DRB3/4/5, 91 *HLA-DRB1~HLA-DQA1*, 115 *HLA-DRB1~HLA-DQB1*, 86 *HLA-DQA1~HLA-DQB1*, and 100 *HLA-DPA1~HLA-DPB1* allele-level haplotypes were estimated. The 2-locus haplotypes presented in Table 10 are restricted to haplotypes estimated at frequencies of 5% and above in significant LD. Supplementary Tables 3–11 shows a comprehensive list of 2-locus haplotypes that occurred at least four times for various loci pairs; haplotypes with counts of three or less were excluded because the EM algorithm is known to be unreliable for estimating rare haplotypes. The EM algorithm estimates haplotype frequencies from unphased genotype data under the assumption of HWE. Due to the extensive HWE deviation observed at multiple loci, the haplotype frequency data presented in this study should be used with caution.

Haplotypes bearing *HLA-A* alleles had the weakest associations, whilst maximal associations were observed at both *HLA-DRB1*07:01:01:01SG~HLA-DQA1*02:01:01:01SG* ($D'_{ij} = 1$), and *HLA-DQA1*05:01:01:02~HLA-DQB1*02:01:01* ($D'_{ij} = 1$). Diversity at the 4-field has generated distinctive haplotypes that would not be evident with 2-field alleles, clearly shown at the *HLA-DPA1~HLA-DPB1* haplotypes where different *HLA-DPA1*01:03:01* 4-field alleles associate with different *HLA-DPB1* alleles. *HLA-DPA1*01:03:01:02* and *HLA-DPA1*01:03:01:04* both associate tightly with *HLA-DPB1*04:01:01:01*

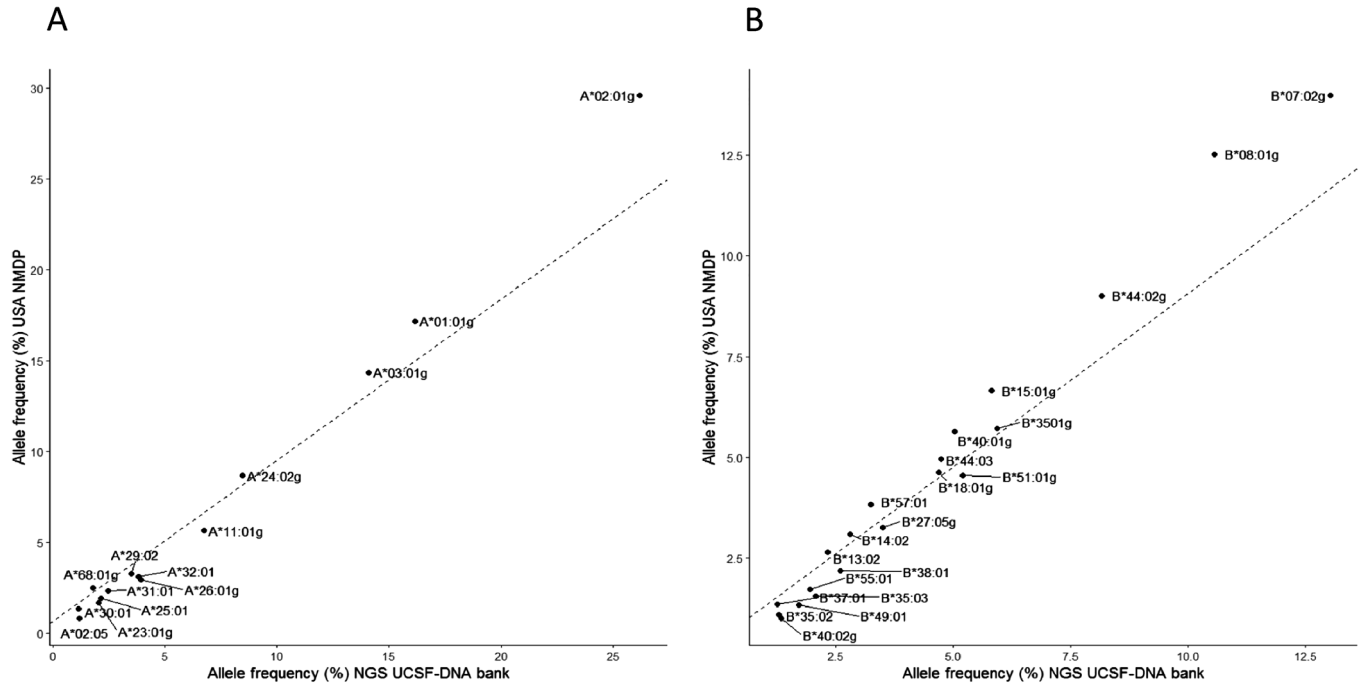

Fig. 2. Comparison of USCF European American HLA allele frequencies (%) with the NMDP registry USA white population. HLA alleles characterized by NGS in the European American sample were reduced to G groups to match the NMDP G group data. (A) *HLA-A* allele frequency, (B) *HLA-B* allele frequency, (C) *HLA-C* allele frequency, (D) *HLA-DQB1* allele frequency, (E) *HLA-DRB1* allele frequency.

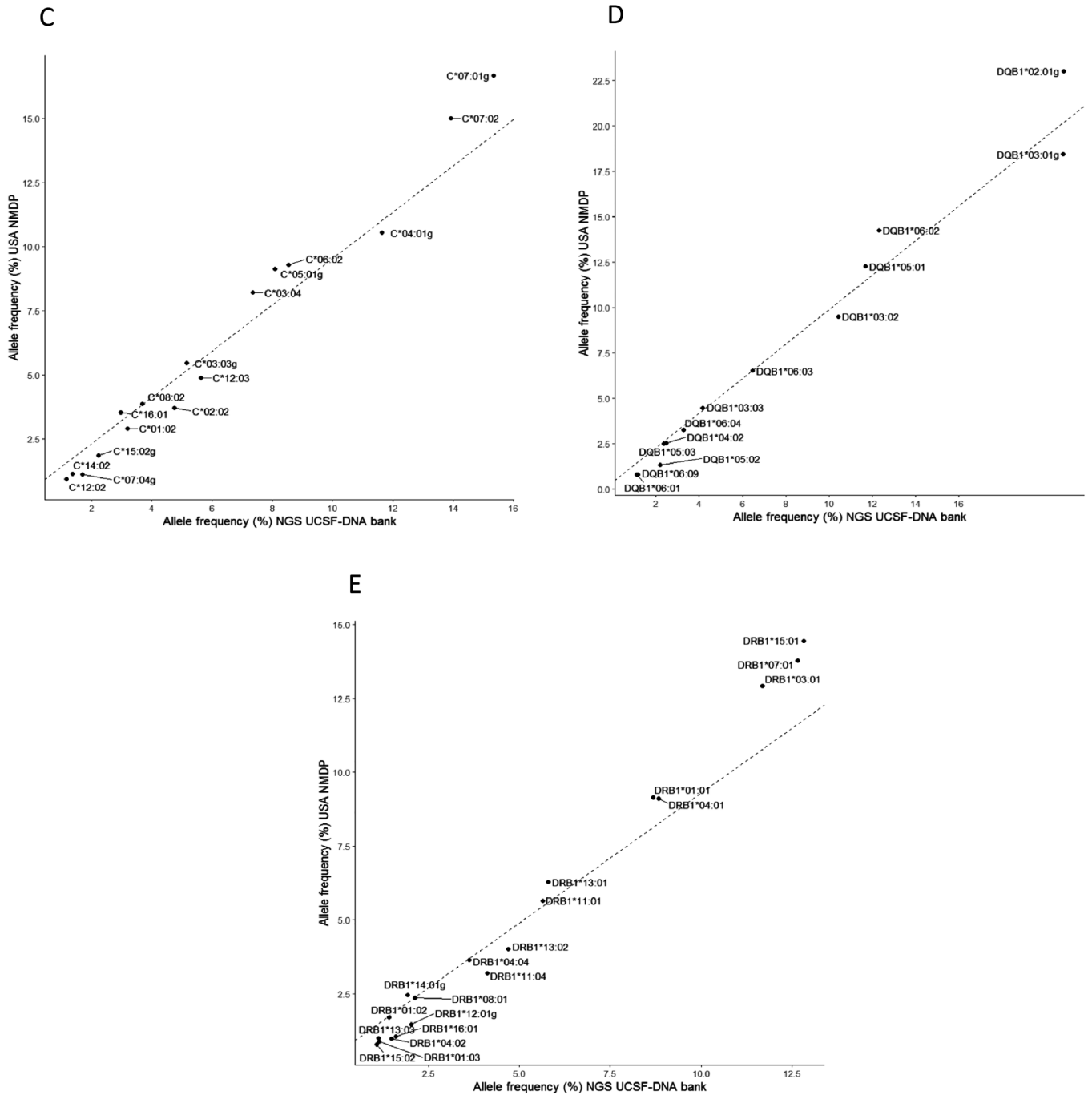

Fig. 2. (*continued*)

(D'_{ij} = 0.794, and 0.830 respectively), *HLA-DPA1*01:03:01:01* with *HLA-DPB1*02:01:02/HLA-DPB1*02:01:19* (D'_{ij} = 0.825), whereas *HLA-DPA1*01:03:01:05* associates strongly with *HLA-DPB1*04:02:01:02* (D'_{ij} = 0.979). Allele *HLA-DPA1*01:03:01:03* exhibits strong LD with *HLA-DPB1*03:01:01* (D'_{ij} = 0.979).

3.7. Extended haplotypes with and without HLA-DP loci

We examined the associations of alleles in haplotypes consisting of 11 loci (*HLA-A*, *HLA-C*, *HLA-B*, *HLA-DRB1*, *HLA-DRB3/4/5*, *HLA-DQA1*, *HLA-DQB1*, *HLA-DPA1*, *HLA-DPB1*), 9 loci excluding the *HLA-DP* loci, and 4 loci (*HLA-DRB1*, *HLA-DRB3/4/5*, *HLA-DQA1*, *HLA-DQB1*). We excluded the *HLA-DP* loci due to the weaker association of *HLA-DP* alleles with alleles at other HLA loci. Comprehensive lists of extended haplotypes estimated, with a minimum count of four, are shown in Supplementary Tables 12 (11 loci), 13 (9 loci), and 14 (6 loci). For the 11-locus haplotypes the frequencies of the top fifteen haplotypes ranged from 0.31% to 3.0% and *HLA-A*01:01:01~HLA-C*07:01:01:01~B*08:01:01:01~ HLA-DRB3*01:01:02:01 /DRB3*01:01:02:02~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* was the most frequent haplotype. This haplotype in the absence of the *HLA-DP* loci was also the most frequent (6.6%). The *HLA-A*01~HLA-C*07~HLA-DRB1*03~HLA-DQB1*02* haplotype corresponds to the most common 8.1 ancestral haplotype in Northern European populations, however previous studies have only reported 2-field or G-group level data for only some of the loci [50,51]. As expected, the common extended haplotypes were essentially composed of different combinations of the 2-locus haplotypes found in strong LD, such as *HLA-C*07:01:01:01~HLA-*

Table 9
Global linkage disequilibrium estimates between pairs of neighboring loci otherwise in parentheses.

Locus Pair	*D*	*D′*	W_n	$W_{loc1/loc2}$	$W_{loc2/loc1}$	*P-value*
(HLA-A~HLA-B)	0.005	0.519	0.310	0.447	0.368	0.000*
HLA-A~HLA-C	0.006	0.483	0.313	0.402	0.365	0.000*
HLA-C~HLA-B	0.007	0.938	0.645	0.846	0.757	0.000*
(HLA-B~HLA-DRB1)	0.005	0.592	0.369	0.420	0.506	0.000*
HLA-DRB3/4/5~HLA-DRB1	0.015	0.938	0.637	0.881	0.695	0.000*
HLA-DRB1~HLA-DQA1	0.012	0.974	0.742	0.864	0.929	0.000*
(HLA-DRB1~HLA-DQB1)	0.011	0.962	0.685	0.842	0.887	0.000*
(HLA-DRB1~HLA-DPB1)	0.006	0.334	0.209	0.253	0.258	0.000*
HLA-DQA1~HLA-DQB1	0.013	0.970	0.668	0.920	0.908	0.000*
(HLA-DQB1~HLA-DPB1)	0.006	0.314	0.183	0.260	0.239	0.000*
HLA-DPA1~HLA-DPB1	0.040	0.906	0.650	0.788	0.788	0.000*

Abbreviations: *D*, linkage disequilibrium (LD) parameter quantifying the deviation of the observed from expected haplotype frequencies; *D′*, normalized measure of *D* ranging from 0 to +1, weights the contribution to LD of specific allele pairs by the product of their allele frequencies (Hedrick 1987); W_n, (also known as Cramér's V statistic), a normalized measure of LD ranging from 0 to +1. It is the re-expression of the Pearson's chi-square statistic for deviations between observed and expected haplotype frequencies (Cramér 1946); $W_{loc1/loc2}$, asymmetric measure of LD for locus '1' conditioned on locus '2'; $W_{loc2/loc1}$, asymmetric measure of LD for locus '2' conditioned on locus '1' (Thomson & Single 2014); *P*, probability values for log- likelihood ratio tests. *P*-values < 0.05 is indicative of overall significant LD.

*B*08:01:01:01* and *HLA-DQA1*05:01:01:02~HLA-DQB1*02:01:01*. Due to the strong LD the frequencies of extended haplotypes were significantly lower than observed for 2-locus haplotypes.

4. Discussion

In this study, we report high-resolution alleles and haplotypes characterized at 3–4 fields in a large sample of European Americans. The NGS approach combined with long-range PCR allows for a comprehensive coverage and sequencing of HLA genes and permits detection and analyses of variants that would otherwise go undetected using lower resolution molecular-based methods. The data generated from this study provides a valuable resource for future population studies conducted in both similar and dissimilar ethnic groups and complements the ever-expanding pool of global HLA frequency data. In addition, our data holds value in the clinical laboratory setting where it can be used as a reference point for expected allele frequency distributions in European Americans, and guide searches for patient donor HLA matching.

The allele families identified in this European American population are consistent with those detected in other European American populations. However, compared to our previous study conducted by Cao et al. in 2001 [22], which examined the distribution of class I alleles, we observed an increased number of alleles per locus at all loci typed. For example, we detected 63, 67, and 93 alleles at *HLA-A*, *HLA-C*, and *HLA-B* loci in comparison the Cao study identified 28, 22, and 47 alleles respectively. The difference in the number of unique alleles identified is obviously explained by the different resolution of the typing methods used in both studies; alleles in the Cao study were characterized by PCR-SSOP typing which only examined nucleotide sequences of exons 2 and 3, therefore polymorphisms outside of this region were not analyzed also indistinguishable alleles were assigned as the lowest number in the ambiguous group. Other factors to consider were that less HLA alleles were identified and assigned in the IPD-IMGT/HLA Database back in 2001 compared to 2016 when version 3.25.0 was released. Also, the Cao study included a small number of individuals (n = 265). However, it is somewhat surprising that the heterozygosity values do not always exhibit the same pattern as the allelic differences. At *HLA-C* and *HLA-B* the observed heterozygosity indexes estimated from this study were higher than the Cao study but the heterozygosity index at the *HLA-A* locus was slightly lower (0.869) than found in the Cao study (0.898).

In our study at the HLA class I loci, the heterozygosity index was lowest for *HLA-A* which can be explained by the dominance of a few alleles that occur at high frequency such as *HLA-A*02:01:01:01* (26.1%), *HLA-A*01:01:01:01* (16.1%), and *HLA-A*03:01:01:01* (13.4%). It can be postulated that the high-frequency of *HLA-A*02:01* alleles are under selective pressure to provide protection from human cancers. A study by Wiedenfeld and colleagues showed a negative association of the presence of *HLA-A*02:01* and the mutated oncoprotein, *p53* [5]. On the other hand, the distribution of *HLA-B* alleles exhibited near maximal heterozygosity, a finding that has been consistently reported in several population studies [13,22,23,52]. It is speculated that the *HLA-B* alleles have undergone overdominant pathogen-mediated balancing selection [53]. In addition, a high level of gene conversions at the *HLA-B* locus, particularly at exon 3, has been reported which could

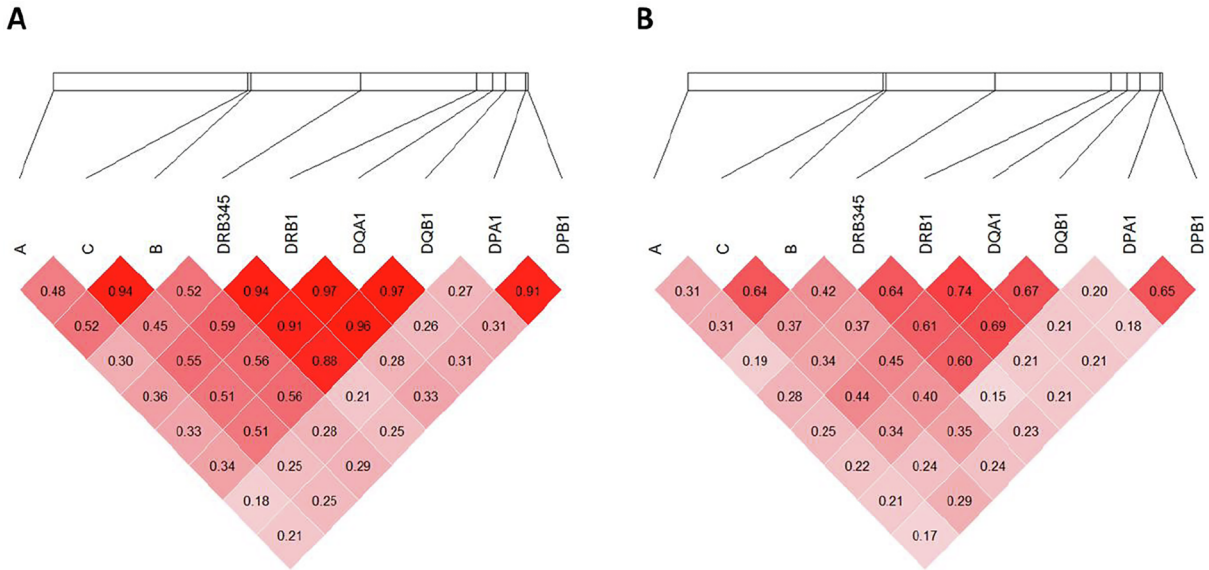

Fig. 3. Plots of global linkage disequilibrium (LD) for pairwise HLA loci based on (A) *D′* values, (B) W_n values.

Table 10
Allele-level linkage disequilibrium and haplotype frequencies ≥ 5% for loci pairs.

Loci Pair	Haplotypes	HF	D'_{ij}
HLA-A~HLA-B	A*01:01:01:01~B*08:01:01:01	0.082	0.739
	A*03:01:01:01~B*07:02:01	0.057	0.351
HLA-A~HLA-C	A*01:01:01:01~C*07:01:01:01	0.085	0.526
	A*03:01:01:01~C*07:02:01:03	0.053	0.336
HLA-C~HLA-B	C*07:02:01:03~B*07:02:01	0.122	0.977
	C*07:01:01:01~B*08:01:01:01	0.103	0.992
	C*05:01:01:02~B*44:02:01:01	0.065	0.946
HLA-B~HLA-DRB1	B*08:01:01:01~DRB1*03:01:01:01SG	0.084	0.778
	B*07:02:01~DRB1*15:01:01:01SG	0.076	0.527
HLA-DRB3/4/5~HLA-DRB1	DRB5*01:01:01~DRB1*15:01:01:01SG	0.127	0.994
	DRB3*01:01:02:01/DRB3*01:01:02:02~DRB1*03:01:01:01SG	0.097	0.804
	DRB4*01:03:01:01/DRB4*01:03:01:01:03~DRB1*04:01:01:01SG	0.087	0.984
	DRB*00:00~DRB1*01:01:01	0.087	1.000
	DRB3*02:02:01:02~DRB1*11:01:01:01	0.053	0.934
	DRB4*01:01:01:01~DRB1*07:01:01:01SG	0.053	0.974
HLA-DRB1~HLA-DQA1	DRB1*15:01:01:01SG~DQA1*01:02:01:01SG	0.127	0.983
	DRB1*07:01:01:01SG~DQA1*02:01:01:01SG	0.126	1.000
	DRB1*03:01:01:01SG~DQA1*05:01:01:02	0.098	0.997
	DRB1*01:01:01~DQA1*01:01:01:02SG	0.083	0.982
	DRB1*13:01:01:01SG~DQA1*01:03:01:02SG	0.056	0.966
	DRB1*04:01:01:01SG~DQA1*03:03:01:01	0.055	0.682
	DRB1*11:01:01:01~DQA1*05:05:01:01SG	0.053	0.917
HLA-DRB1~HLA-DQB1	DRB1*15:01:01:01SG~DQB1*06:02:01	0.121	0.980
	DRB1*03:01:01:01SG~DQB1*02:01:01	0.115	0.998
	DRB1*07:01:01:01SG~DQB1*02:02:01:01	0.091	0.997
	DRB1*01:01:01~DQB1*05:01:01:03	0.083	0.976
	DRB1*13:01:01:01SG~DQB1*06:03:01	0.058	0.987
	DRB1*11:01:01:01~DQB1*03:01:01:03	0.053	0.944
HLA-DQA1~HLA-DQB1	DQA1*01:02:01:01SG~DQB1*06:02:01	0.123	0.993
	DQA1*05:05:01:01SG~DQB1*03:01:01:03	0.100	0.917
	DQA1*05:01:01:02~DQB1*02:01:01	0.098	1.000
	DQA1*03:01:01~DQB1*03:02:01	0.094	0.961
	DQA1*02:01:01:01SG~DQB1*02:02:01:01	0.091	0.994
	DQA1*01:01:01:02SG~DQB1*05:01:01:03	0.091	0.994
	DQA1*03:03:01:01~DQB1*03:01:01:01	0.062	0.780
	DQA1*01:03:01:02SG~DQB1*06:03:01	0.058	0.983
HLA-DPA1~HLA-DPB1	DPA1*01:03:01:02~DPB1*04:01:01:01	0.249	0.794
	DPA1*01:03:01:04~DPB1*04:01:01:01	0.165	0.830
	DPA1*01:03:01:01~DPB1*02:01:02/DPB1*02:01:19	0.112	0.825
	DPA1*01:03:01:05~DPB1*04:02:01:02	0.094	0.979
	DPA1*01:03:01:03~DPB1*03:01:01	0.082	0.979

Abbreviations: HF, haplotype frequency; D'_{ij}, standardized measure of D_{ij} (LD between two alleles at two different loci) ranging from 0 to +1 (Lewontin 1964).

also explain the high levels of alleles at this locus [54]. *HLA-B* alleles have been reported to be associated with a number of infectious disease, such as *HLA-B*27:05* and *HLA-B*57:01* confer long-term non-progression from Human Immunodeficiency virus (HIV) infections to acquired immune deficiency disease (AIDS). *HLA-B*57:01* is also protective against hepatitis C virus (HCV) infection, and *HLA-B*08* is associated with faster disease progression in HCV and HIV [55]. The major role of *HLA-C* molecules has always been assumed to regulate the cytotoxicity of natural killer (NK) cells by acting as a ligand for killer-cell immunoglobulin-like receptors (KIRs) expressed on NK cells [56]. However it has been observed that *HLA-C* restricted CD8 + T-cell responses (CTL) epitopes have been identified in chronic infections such as Epstein-Barr virus (EBV) and HIV infections [55]. Compared to *HLA-A* and *HLA-B*, *HLA-C* is expressed at relatively low levels (10% of classical MHC molecules) [57]. Also the *HLA-C*-peptide complex is less stable for peptide presentation, therefore more sensitive to HLA loss that can be detected by KIR molecules. The non-classical *HLA-F*, *HLA-G*, and *HLA-E* molecules are less polymorphic than classical class I genes; the advantage of low diversity at these loci is that more alleles would create holes in the T-cell repertoire because of self-tolerance and therefore additional diversity would be negatively selected.

Of the class II loci, *HLA-DPB1* exhibited the lowest level of observed heterozygosity (0.768), which again can be explained by the dominance of a high-frequency allele, *HLA-DPB1*04:01:01:01* (42.6%). In contrast *HLA-DRB1* had the largest heterozygosity index (0.926) which is comparable to other European descent populations; Spain 0.927, Italy 0.930 but lower than found in non-European descent populations such as African-Americans 0.937, and USA Hispanics 0.946 (Creary et. al. unpublished data).

It is noteworthy that the *HLA-DPA1* locus had the lowest number of alleles of all the loci sequenced in this study, with the exception of alleles counted individually for *HLA-DRB3*, *HLA-DRB4*, and *HLA-DRB5* loci. Strikingly 81.7% of the *HLA-DPA1* alleles detected were 4-field variants of *HLA-DPA1*01:03:01*, this suggests that the *HLA-DPA1*01:03* allele is under selective pressure and only intronic variants of this allele are generated in the population. We postulate that the *HLA-DPA1*01:03* allele is 'fixed' in the population due to tight LD with *HLA-DPB1* alleles which exhibits far more allelic variation, and in terms of antigen presentation the latter locus holds more biological significance. We can view the biological function of *HLA-DPA1* molecules as a subunit that must pair with many protein variants encoded by *HLA-DPB1* in which diversification of *HLA-DPA1* could be restrained because structural variation would prevent pairing of *HLA-DPA1* with some *HLA-DPB1* alleles. In this sense, we see a similar pattern for beta-2 micro globulin which is a stable non-polymorphic component of MHC class I molecules that are necessary for the cell surface expression of

many class I molecules and the stability of the peptide binding grove. Like-wise the *HLA-DRA* gene shows virtually no protein sequence diversification possibly due to constraints to pairing with many proteins encoded by *HLA-DRB1*, *HLA-DRB3*, *HLA-DRB4* and *HLA-DRB5* genes.

We found that the distribution of 5 of the 11 loci deviated significantly from Hardy-Weinberg proportions, which collectively was mostly due to an excess of specific homozygous genotypes, and specific heterozygous genotypes that included many rare alleles. The deviation from HWE could be partly explained by recent genetic admixture since the majority of the rare alleles we found in our European American cohort are observed at elevated frequencies in non-European populations. These alleles can be considered as informative markers of ancestry. For instance comparing the allele frequency distribution found in this study (EuAm) with those reported in the allele frequencies database (www.allelefrequencies.net): *HLA-B*39:24:01* EuAm AF = 0.0002 but found at 0.082 in USA Spanish ancestry population; *HLA-B*15:29*, EuAm AF = 0.0002 in Israel Iran Jews (0.013), India (0.011); *HLA-C*07:06* EuAm AF = 0.0005, India (0.0538), South African India (0.1200), and China (0.0230). HWE deviations in a population may impact haplotype analyses and interpretation. The EM algorithm uses unphased genotype data along with the assumption of HWE to estimate haplotype frequencies (HFs). Particular HFs may be under or overestimated if the loci included in the analyses deviate significantly from HWE. To improve the accuracy of the estimation of two-locus haplotypes generated from HLA loci that deviated from HWE, Single *et. al.* proposed a 'collapsing over loci' procedure, where HFs are first estimated for extended haplotypes and then collapsed over loci to generate two-locus haplotypes; specific two-locus HFs are estimated indirectly by summing the HFs from the multi-locus estimation [58]. We explored this method using our data and found in general that the direct and indirect two-locus haplotype estimates were comparable. For example, we showed that the direct HF (0.1219) for *HLA-C*07:02:01:01~HLA-B*07:02:01* was very similar to the haplotype estimate of the indirect collapsing method (0.1217). Furthermore, the D'_{ij} values were also similar; 0.9770 (direct two-locus estimate) and 0.9787 (collapsing method). The same holds true for a less frequent haplotype *HLA-C*07:04:01:01~HLA-B*44:27:01*: direct HF = 0.0023, $D'_{ij} = 1$; indirect HF = 0.0022, $D'_{ij} = 1$). These findings suggest that overall the direct frequency estimates of two-locus haplotypes are accurate and could be used to make inferences about the European American population, however, the decision will be entirely at the discretion of the user.

Many studies have shown that the distributions of HLA allele frequencies are consistent with the action of balancing selection [15,16,18]. In this study we found significant evidence of balancing selection (negative F_{nd} value) at *HLA-DQA1* and *HLA-DQB1* loci. Alter and colleagues showed that positive F_{nd} values were restricted to class I genes analysed individually and particularly for haplotypes bearing *HLA-A* in numerous populations they analysed in their study [59]. Our finding of a positive F_{nd} value at the *HLA-A* locus resembles the Alter study. The excess homozygosity observed at the *HLA-A* locus in our study (observed $F = 0.129$, expected $F = 0.088$) suggests that homozygosity confers a selective advantage over heterozygosity and subpopulation structure.

NGS of extended long-range products has permitted characterization of difficult to sequence genomic regions or regions that were not routinely interrogated by historical HLA typing methods. For instance, we accurately and efficiently detected 4 non-expressed HLA variants in the entire cohort. The detection of null alleles is important in HSCT as misidentification may result in an HLA mismatch between patient and donor, which is highly likely to increase the risk of engraftment failure and severe graft-versus-host disease [60]. The location of null variants is ubiquitous and may lie within exons, both inside and outside of the antigen-recognition site (ARS), as well as non-coding regions. In this study, the 3 class I null allele variants are all located outside of the ARS; *HLA-C*04:09N* (deletion of A nucleotide in codon 341 exon 7), *HLA-B*51:11N* (insertion of C nucleotide in codon 185 exon 4), *HLA-B*57:79N* (deletion of G nucleotide in codon 268 exon 4). Also the *HLA-DRB4*01:03:01:02N* variant is located at the 3′ end of intron 1 / exon 2 boundary (G > A substitution position 9656). These findings suggest that all genomic regions should be sequenced in the clinical laboratory to ensure detection of all possible non-expressed alleles.

*HLA-DRB4*01:03:01:02N* was the most frequent null allele observed in the sample (AF = 3.4%). Interestingly we note that *HLA-DRB4*01:03:01:02N* is carried on a different extended haplotype than its expressed counterpart *HLA-DRB4*01:03:01:01/HLA-DRB4*01:03:01:03*. The *HLA-DRB4*01:03:01:02N* allele is in strong LD ($D'_{ij} = 0.95$) with *HLA-DRB1*07:01:01:01SG* and the most frequent extended haplotype is; *HLA-A*01:01:01~HLA-C*06:02:01:01~ HLA-B*57:01:01~HLA-DRB1*07:01:01:01SG~HLA-DRB4*01:03:01:02N~ HLA-DQA1*02:01:01:01SG~HLA-DQB1*03:03:02:01* (HF = 0.008). Whereas *HLA-DRB4*01:03:01:01/HLA-DRB4*01:03:01:03* is ubiquitous and the association with *HLA-DRB1*07:01:01:01SG* is weak ($D'_{ij} = 0.14$).

Frequent (≥0.25%) *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*. The *C*16:01:01:01~HLA-B*44:03:01:01* block is also 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. The distinct C~B and DQA1~DQB1 haplotype blocks suggest; (i) the C~B and DQA1~DQB1 haplotype alleles may act as markers for the *HLA-DRB4*01:03:01* and the *HLA-DRB4*01:01:01:01* alleles, (ii) it is highly likely that the 4-field *HLA-DRB1*07:01:01* allele differs on *HLA-DRB4*01:03:01:01*, *HLA-DRB4*01:03:01:02N*, and *HLA-DRB4*01:01:01:01* bearing haplotypes. Due to low complexity genomic regions we could not further refine the *HLA-DRB4*01:01/01:03~HLA-DRB1*07:01* associations. These findings, as well as other haplotype blocks comprised of SG alleles, strongly suggest that HLA diversity in the European Americans is likely to be greater than estimated.

We observed highly diverse two-locus haplotypes which was mainly due to silent nucleotide substitution differences in introns and untranslated regions captured by 4-field alleles. It appears that the physical distance and the recombination fractions between pairs of loci on the MHC greatly impact the LD measurement. We noted a number of allele pairs at maximal LD in some allele level *HLA-C~HLA-B*, *HLA-DQA1~HLA-DQB1*, *HLA-DRB1~HLA-DQB1*, and *HLA-DPA1~ HLA-DPB1* haplotypes. In general for 2-loci blocks the LD is tighter when alleles are defined at 4 fields compared to 2-field resolution. However for common 4 field alleles having associations with several alleles of the same 2/3 field group the D'_{ij} value decreases as seen in the case of *HLA-DPB1*04:01:01:01*, *HLA-DRB1*13:01:01:01SG*, *HLA-DQA1*01:03:01* and *HLA-DRB1*01:01* with two *HLA-DQA1*01:01:01:02/03* alleles.

We also observed particular allele families were exclusively associated for instance *HLA-DQA1*01* allele group associates with either *HLA-DQB1*05* and *HLA-DQB1*06* families. Also *HLA-DQA1**02, 03, 04, 05, 06 alleles associate with *HLA-DQB1**02, 03 and 04 allele families. These specific *HLA-DQA1~HLA-DQB1* associations may result from structural complementation constraints. In contrast excluding founder effects, other haplotype associations, such as, *HLA-B**08~*HLA-C**07, could result from functional complementation that may result in wider peptide binding repertoires or in effective immune responses to pathogens. At *HLA-A~HLA-C* we note that two 4-field variants *HLA-C*06:02:01:01* and *HLA-C*06:02:01:03* are associated with *HLA-A*30:01:01* and *HLA-A*29:02:01:02* respectively. These findings suggest that strong LD across the HLA region appears to have evolved for selection of particular allele combinations. Specific HLA haplotypes may be selected because of the differential ability of the associated alleles to present an antigen or autoantigens, or its involvement in shaping the T-cell repertoire in the thymus. Examination of non-European populations that exhibit different patterns of LD would be

necessary in order to refine segregation patterns of alleles in European populations.

We have extended previous observations and shown that HLA haplotypes are highly conserved over several hundreds of kilo bases. To our knowledge, we are the first to report extended haplotypes comprised of 11 loci characterized at 3 – 4 fields in a European American population. Overall the distribution of the extended haplotypes is impacted by both the diversity of alleles at loci and LD between neighboring and non-neighboring loci.

In conclusion, the allele frequencies at class I and class II loci in our European American sample shows similarity to HLA distributions found in European ancestry populations. However we observed the presence of alleles that are consistent with the presence of genetic admixture in our European American sample. The extensive multi-locus HWE deviation indicates that the population data should be used with caution. We envision that our data will be useful for guiding searches of unrelated HSCTs and may be used to improve algorithms for predicting patient-donor matching.

Acknowledgements

We are grateful to the individuals who participated in this study and to Ms. R. Guerrero for sample processing. This work was supported by grant U19NS095774 (JRO, MFV) from the U.S. National Institutes of Health (NIH). The UCSF DNA biorepository is supported by RG-1611-26299 from the National Multiple Sclerosis Society. We thank the Michael J. Fox and Coriell biorepositories for access to de-identified DNA samples.

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, contributed to the interpretation of the results, and contributed to the final version of the manuscript. JRO and JAH were involved in the conception of the study and contributed to the final version of the manuscript. SG, KCM, and GMM performed the NGS genotyping assays. SC and AS were responsible for sample and data collection at the UCSF-DNA bank. All authors read and approved the final manuscript.

Declarations of Competing Interest

The authors declare no competing financial or other interests.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.humimm.2019.07.275.

References

[1] J. Trowsdale, J.C. Knight, Major histocompatibility complex genomics and human disease, Annu. Rev. Genomics Hum. Genet. Annu. Rev. G (2015) 301–323.
[2] L.E. Creary, K.C. Mallempati, S. Gangavarapu, S.J. Caillier, J.R. Oksenberg, M.A. Fernández-Viña, 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.
[3] N. Isobe, A. Keshavan, P.-A. Gourraud, A.H. Zhu, E. Datta, R. Schlaeger, S.J. Caillier, A. Santaniello, A. Lizée, D.S. Himmelstein, S.E. Baranzini, J. Hollenbach, B.A.C. Cree, S.L. Hauser, J.R. Oksenberg, R.G. Henry, Association of HLA genetic risk burden with disease phenotypes in multiple sclerosis, JAMA Neurol. 73 (2016) 795–802.
[4] W.T. Wissemann, E.M. Hill-Burns, C.P. Zabetian, S.A. Factor, N. Patsopoulos, B. Hoglund, C. Holcomb, R.J. Donahue, G. Thomson, H. Erlich, H. Payami, Association of Parkinson disease with structural and regulatory variants in the HLA region, Am. J. Hum. Genet. 93 (2013) 984–993.
[5] E.A. Wiedenfeld, M. Fernandez-Viña, J.A. Berzofsky, D.P. Carbone, Evidence for selection against human lung cancers bearing p53 missense mutations which occur within the HLA A*0201 peptide consensus motif, Cancer Res. 54 (1994) 1175–1177.
[6] D. Chen, J. Hammer, D. Lindquist, A. Idahl, U. Gyllensten, A variant upstream of HLA-DRB1 and multiple variants in MICA influence susceptibility to cervical cancer in a Swedish population, Cancer Med. 3 (2014) 190–198.
[7] T.I.H.C. International HIV Controllers Study, F. Pereyra, X. Jia, P.J. McLaren, A. Telenti, P.I.W. de Bakker, B.D. Walker, S. Ripke, C.J. Brumme, S.L. Pulit, M. Carrington, C.M. Kadie, J.M. Carlson, D. Heckerman, R.R. Graham, R.M. Plenge, S.G. Deeks, L. Gianniny, G. Crawford, J. Sullivan, E. Gonzalez, L. Davies, A. Camargo, J.M. Moore, N. Beattie, S. Gupta, A. Crenshaw, N.P. Burtt, C. Guiducci, N. Gupta, X. Gao, Y. Qi, Y. Yuki, A. Piechocka-Trocha, E. Cutrell, R. Rosenberg, K.L. Moss, P. Lemay, J. O'Leary, T. Schaefer, P. Verma, I. Toth, B. Block, B. Baker, et al., The major genetic determinants of HIV-1 control affect HLA class I peptide presentation, Science. 330 (2010) 1551–1557.
[8] M. McCormack, A. Alfirevic, S. Bourgeois, J.J. Farrell, D. Kasperavičiūtė, M. Carrington, G.J. Sills, T. Marson, X. Jia, P.I.W. de Bakker, K. Chinthapalli, M. Molokhia, M.R. Johnson, G.D. O'Connor, E. Chaila, S. Alhusaini, K.V. Shianna, R.A. Radtke, E.L. Heinzen, N. Walley, M. Pandolfo, W. Pichler, B.K. Park, C. Depondt, S.M. Sisodiya, D.B. Goldstein, P. Deloukas, N. Delanty, G.L. Cavalleri, M. Pirmohamed, HLA-A*3101 and carbamazepine-induced hypersensitivity reactions in Europeans, N. Engl. J. Med. 364 (2011) 1134–1143.
[9] R.J. Schutte, Y. Sun, D. Li, F. Zhang, D.A. Ostrov, Human leukocyte antigen associations in drug hypersensitivity reactions, Clin. Lab. Med. 38 (2018) 669–677.
[10] J. Robinson, J.A. Halliwell, J.D. Hayhurst, P. Flicek, P. Parham, S.G.E. Marsh, The IPD and IMGT/HLA database: allele variant databases, Nucleic Acids Res. 43 (2015) D423–D431.
[11] D. Meyer, V.R.C. Aguiar, B.D. Bitarello, D.Y.C. Brandt, K. Nunes, A genomic perspective on HLA evolution, Immunogenetics 70 (2018) 5–27.
[12] D. Meyer, R.M. Single, S.J. Mack, H.A. Erlich, G. Thomson, Signatures of demographic history and natural selection in the human major histocompatibility complex loci, Genetics 173 (2006) 2121–2142.
[13] T. Goeury, L.E. Creary, M.A. Fernandez-Vina, J.-M. Tiercy, J.M. Nunes, A. Sanchez-Mazas, Mandenka from senegal, HLA 91 (2018) 148–150.
[14] E. Moore, A. Grifoni, D. Weiskopf, V. Schulten, C.S.L. Arlehamn, M. Angelo, J. Pham, S. Leary, J. Sidney, D. Broide, A. Frazier, E. Phillips, S. Mallal, S.J. Mack, A. Sette, Sequence-based HLA-A, B, C, DP, DQ, and DR typing of 496 adults from San Diego, California, USA, Hum. Immunol. 79 (2018) 821–822.
[15] S.J. Mack, B. Tu, R. Yang, C. Masaberg, J. Ng, C.K. Hurley, Human leukocyte antigen-A, -B, -C, -DRB1 allele and haplotype frequencies in Americans originating from southern Europe: contrasting patterns of population differentiation between Italian and Spanish Americans, Hum. Immunol. 72 (2011) 144–149.
[16] S.J. Mack, B. Tu, A. Lazaro, R. Yang, A.K. Lancaster, K. Cao, J. Ng, C.K. Hurley, 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.
[17] M. Maiers, L. Gragert, W. Klitz, High-resolution HLA alleles and haplotypes in the United States population, Hum. Immunol. 68 (2007) 779–788.
[18] B. Tu, S.J. Mack, A. Lazaro, A. Lancaster, G. Thomson, K. Cao, M. Chen, G. Ling, R. Hartzman, J. Ng, C.K. Hurley, HLA-A, -B, -C, -DRB1 allele and haplotype frequencies in an African American population, Tissue Antigens. 69 (2007) 73–85.
[19] J.J. Chen, J.A. Hollenbach, E.A. Trachtenberg, J.J. Just, M. Carrington, K.S. Rønningen, A. Begovich, M.-C. King, S. McWeeney, S.J. Mack, H.A. Erlich, G. Thomson, Hardy-Weinberg testing for HLA class II (DRB1, DQA1, DQB1, AND DPB1) loci in 26 human ethnic groups, Tissue Antigens. 54 (1999) 533–542.
[20] Y. Pei, H. Huang, H. Li, J. Chen, G. Wu, Allelic and haplotype diversity of HLA-A, HLA-B and HLA-DRB1 gene at high resolution in the Nanning Han population, Int. J. Immunogenet. 45 (2018) 201–209.
[21] Y.R. Thorstenson, L.E. Creary, H. Huang, V. Rozot, T.T. Nguyen, F. Babrzadeh, S. Kancharla, M. Fukushima, R. Kuehn, C. Wang, M. Li, S. Krishnakumar, M. Mindrinos, M.A. Fernandez Viña, T.J. Scriba, M.M. Davis, 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.
[22] K. Cao, J. Hollenbach, X. Shi, W. Shi, M. Chopek, M.A. Ferna, 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, Hum. Immunol. 8859 (2001).
[23] C. Capittini, A. De Silvestri, M. Guarene, A. Pasi, C. Tinelli, C. Perotti, HLA-A, -B, -DRB1 allele and haplotype frequencies of 674 cord blood donors from North Italy, Hum. Immunol. 78 (2017) 412–413.
[24] S. Davey, J. Ord, C. Navarrete, C. Brown, HLA-A, -B and -C allele and haplotype frequencies defined by next generation sequencing in a population of 519 English blood donors, Hum. Immunol. 78 (2017) 397–398.
[25] P.A. Gourraud, D.J. Pappas, A. Baouz, M.L. Balère, F. Garnier, E. Marry, High-resolution HLA-A, HLA-B, and HLA-DRB1 haplotype frequencies from the French Bone Marrow Donor Registry, Hum. Immunol. 76 (2015) 381–384.
[26] E.W. Petersdorf, M. Malkki, C. O'hUigin, M. Carrington, T. Gooley, M.D. Haagenson, M.M. Horowitz, S.R. Spellman, T. Wang, P. Stevenson, High HLA-DP expression and graft-versus-host disease, N. Engl. J. Med. 373 (2015) 599–609.
[27] R. Thomas, R. Apps, Y. Qi, X. Gao, V. Male, C. O'hUigin, G. O'Connor, D. Ge, J. Fellay, J.N. Martin, J. Margolick, J.J. Goedert, S. Buchbinder, G.D. Kirk, M.P. Martin, A. Telenti, S.G. Deeks, B.D. Walker, D. Goldstein, D.W. McVicar, A. Moffett, M. Carrington, HLA-C cell surface expression and control of HIV/AIDS correlate with a variant upstream of HLA-C, Nat. Genet. 41 (2009) 1290.
[28] K. Cao, M. Chopek, M.A. Fernández-Viña, 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.
[29] K. Kotsch, J. Wehling, R. Blasczyk, 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.
[30] K. Kotsch, J. Wehllng, S. Köhler, R. Blasczyk, 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.
[31] M. Bunce, C.M. O'Neill, M.C.N.M. Barnardo, P. Krausa, M.J. Browning, P.J. Morris, K.I. Welsh, 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.
[32] N. Cereb, P. Maye, S. Lee, Y. Kong, S.Y. Yang, 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.
[33] K. Adhikari, J.C. Chacón-Duque, J. Mendoza-Revilla, M. Fuentes-Guajardo, A. Ruiz-Linares, The genetic diversity of the Americas, Annu. Rev. Genomics Hum. Genet. 18 (2017) 277–296.
[34] R.R. Humes, Karen R.; Jones, Nicholas A.; Ramirez, https://www.census.gov/prod/cen2010/briefs/c2010br-02.pdf. (Accessed August 20, 2018) United States Census Bureau, Overview of Race and Hispanic Origin: 2010, United States Census Bur. (2011) 1–23.
[35] I. Halder, M. Shriver, M. Thomas, J.R. Fernandez, T. Frudakis, A panel of ancestry informative markers for estimating individual biogeographical ancestry and admixture from four continents: utility and applications, Hum. Mutat. 29 (2008) 648–658.
[36] I. Halder, B.-Z. Yang, H.R. Kranzler, M.B. Stein, M.D. Shriver, J. Gelernter, Measurement of admixture proportions and description of admixture structure in different U.S. populations, Hum. Mutat. 30 (2009) 1299–1309.
[37] K. Bryc, E.Y. Durand, J.M. Macpherson, D. Reich, J.L. Mountain, The genetic ancestry of African Americans, Latinos, and European Americans across the United States, Am. J. Hum. Genet. 96 (2015) 37–53.
[38] C. Wang, S. Krishnakumar, J. Wilhelmy, F. Babrzadeh, L. Stepanyan, L.F. Su, D. Levinson, M.A. Fernandez-Vina, R.W. Davis, M.M. Davis, M. Mindrinos, High-throughput, high-fidelity HLA genotyping with deep sequencing, Proc. Natl. Acad. Sci. 109 (2012) 8676–8681.
[39] G. Andersson, Evolution of the human HLA-DR region, Front. Biosci. 3 (1998) d739–d745.
[40] A.K. Lancaster, R.M. Single, O.D. Solberg, M.P. Nelson, G. Thomson, PyPop update–a software pipeline for large-scale multilocus population genomics, Tissue Antigens. 69 (Suppl 1) (2007) 192–197.
[41] S.W. Guo, E.A. Thompson, Performing the exact test of Hardy-Weinberg proportion for multiple alleles, Biometrics 48 (1992) 361–372.
[42] W.J. Ewens, The sampling theory of selectively neutral alleles, Theor. Popul. Biol. 3 (1972) 87–112.
[43] G.A. Watterson, The homozygosity test after a change in population size, Genetics 112 (1986) 899–907.
[44] P.W. Hedrick, Gametic disequilibrium measures: proceed with caution, Genetics 117 (1987) 331–341.
[45] H. Cramer, Mathematical Methods of Statistics, I, Princeton University, Princeton, 1946.
[46] G. Thomson, R.M. Single, Conditional asymmetric linkage disequilibrium (ALD): extending the biallelic r2 measure, Genetics 198 (2014) 321–331.
[47] D.J. Pappas, W. Marin, J.A. Hollenbach, S.J. Mack, Bridging ImmunoGenomic Data Analysis Workflow Gaps (BIGDAWG): An integrated case-control analysis pipeline, Hum. Immunol. 77 (2016) 283–287.
[48] S.J. Mack, P. Cano, J.A. Hollenbach, J. He, C.K. Hurley, D. Middleton, M.E. Moraes, S.E. Pereira, J.H. Kempenich, E.F. Reed, M. Setterholm, A.G. Smith, M.G. Tilanus, M. Torres, M.D. Varney, C.E.M. Voorter, G.F. Fischer, K. Fleischhauer, D. Goodridge, W. Klitz, A.-M. Little, M. Maiers, S.G.E. Marsh, C.R. Müller, H. Noreen, E.H. Rozemuller, A. Sanchez-Mazas, D. Senitzer, E. Trachtenberg, M. Fernandez-Vina, Common and well-documented HLA alleles: 2012 update to the CWD catalogue, Tissue Antigens. 81 (2013) 194–203.
[49] M. Cullen, S.P. Perfetto, W. Klitz, G. Nelson, M. Carrington, High-resolution patterns of meiotic recombination across the human major histocompatibility complex, Am. J. Hum. Genet. 71 (2002) 759–776.
[50] C.M. Gambino, A. Aiello, G. Accardi, C. Caruso, G. Candore, Autoimmune diseases and 8.1 ancestral haplotype: an update, HLA. 92 (2018) 137–143.
[51] A. Lande, I. Andersen, T. Egeland, B.A. Lie, M.K. Viken, HLA -A, -C, -B, -DRB1, -DQB1 and -DPB1 allele and haplotype frequencies in 4514 healthy Norwegians, Hum. Immunol. 79 (2018) 527–529.
[52] K. Cao, A.M. Moormann, K.E. Lyke, C. Masaberg, O.P. Sumba, O.K. Doumbo, D. Koech, A. Lancaster, M. Nelson, D. Meyer, R. Single, R.J. Hartzman, C.V. Plowe, J. Kazura, D.L. Mann, M.B. Sztein, G. Thomson, M.A. Fernandez-Vina, Differentiation between African populations is evidenced by the diversity of alleles and haplotypes of HLA class I loci, Tissue Antigens. 63 (2004) 293–325.
[53] P. Parham, C.E. Lomen, D.A. Lawlor, J.P. Ways, N. Holmes, H.L. Coppin, R.D. Salter, A.M. Wan, P.D. Ennis, Nature of polymorphism in HLA-A, -B, and -C molecules, Proc. Natl. Acad. Sci. U.S.A. 85 (1988) 4005–4009.
[54] S.N. McAdam, J.E. Boyson, X. Liu, T.L. Garber, A.L. Hughes, R.E. Bontrop, D.I. Watkins, A uniquely high level of recombination at the HLA-B locus, Proc. Natl. Acad. Sci. U.S.A. 91 (1994) 5893–5897.
[55] A. Kosmrlj, E.L. Read, Y. Qi, T.M. Allen, M. Altfeld, S.G. Deeks, F. Pereyra, M. Carrington, B.D. Walker, A.K. Chakraborty, Effects of thymic selection of the T-cell repertoire on HLA class I-associated control of HIV infection, Nature 465 (2010) 350–354.
[56] S.K. Anderson, Molecular evolution of elements controlling HLA-C expression: Adaptation to a role as a killer-cell immunoglobulin-like receptor ligand regulating natural killer cell function, HLA 92 (2018) 271–278.
[57] J. Zemmour, P. Parham, Distinctive polymorphism at the HLA-C locus: implications for the expression of HLA-C, J. Exp. Med. 176 (1992) 937–950.
[58] R.M. Single, D. Meyer, J.A. Hollenbach, M.P. Nelson, J.A. Noble, H.A. Erlich, G. Thomson, Haplotype Frequency Estimation in Patient Populations: the effect of departures from Hardy-Weinberg proportions and collapsing over a locus in the HLA region, Genetic Epidemiol. 22 (2002) 186–195.
[59] I. Alter, L. Gragert, S. Fingerson, M. Maiers, Y. Louzoun, HLA class I haplotype diversity is consistent with selection for frequent existing haplotypes, PLOS Comput. Biol. 13 (2017) e1005693.
[60] H.-A. Elsner, R. Blasczyk, Immunogenetics of HLA null alleles: implications for blood stem cell transplantation, Tissue Antigens. 64 (2004) 687–695.