Published in final edited form as:

Hum Immunol. 2016 March ; 77(3): 273–282. doi:10.1016/j.humimm.2015.10.018.

HLA Haplotype Validator for Quality Assessments of HLA Typing

Kazutoyo Osoegawa^{1,3}, Steven J. Mack³, Julia Udell³, David A. Noonan³, Steven Ozanne², Elizabeth Trachtenberg³, and Matthew Prestegard²

¹Department of Pathology, Stanford University, Stanford, CA, USA

²National Marrow Donor Program, Minneapolis, MN, USA

³Children's Hospital Oakland Research Institute, Oakland, CA, USA

Abstract

HLA alleles are observed in specific haplotypes, due to linkage disequilibrium (LD) between particular alleles. Haplotype frequencies for alleles in strong LD have been established for specific ethnic groups and racial categories.

Application of high-resolution HLA typing using Next Generation Sequencing (NGS) is becoming a common practice in research and clinical laboratory settings.

HLA typing errors using NGS occasionally occur due to allelic sequence imbalance or misalignment. Manual inspection of HLA genotypes is labor intensive and requires an in-depth knowledge of HLA alleles and haplotypes.

We developed the “HLA Haplotype Validator (HLAHapV)” software, which inspects an HLA genotype for both the presence of common and well-documented alleles and observed haplotypes. The software also reports warnings when rare alleles, or alleles that do not belong to recognized haplotypes, are found.

The software validates observable haplotypes in genotype data providing increased confidence regarding the accuracy of the HLA typing, thus reducing the effort involved in correcting potential HLA typing errors. The HLAHapV software is a powerful tool for quality control of HLA genotypes prior to the application of downstream analyses.

We demonstrated the use of the HLAHapV software for identifying unusual haplotypes leading to finding potential HLA typing errors.

Keywords

Haplotype; Linkage Disequilibrium; HLA-B~C haplotype block; HLA-DRB3/4/5~DRB1~DQB1 haplotype block

Corresponding Author: Kazutoyo Osoegawa, Address: Department of Pathology, Stanford University School of Medicine, 3155 Porter Drive, Palo Alto CA 94304, Tel: 650-724-0169, Fax: 650-724-0294, kazutoyo@stanford.edu.

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

1. Introduction

Human Leukocyte Antigen (HLA) genes are the most polymorphic genes in the human genome [1-3]. HLA genes contain numerous single nucleotide polymorphisms (SNPs) [4]. In addition to the accumulation of SNP variants, the high-levels of allelic polymorphism at these genes have evolved through intra- and intergenic recombination and short-tract gene conversions [5-7]. As of April 2015, 9,749 alleles have been described for HLA class I genes, and 3,274 for class II, totaling 13,023 alleles registered in IMGT/HLA Database version 3.20.0 [8].

HLA genotyping using next-generation sequencing (NGS) is becoming a popular strategy in research and clinical laboratories. NGS systems generate large numbers of “clonal” sequence reads derived from individual DNA molecules, in a massively parallel fashion. The clonal nature of NGS allows each sequence read being assigned to a single allele, resulting in HLA types with fewer ambiguities than those obtained from more widely used Sanger-sequencing based typing (SBT) methods [9, 10]. SBT has been the gold standard for so-called high-resolution HLA typing, in which the “core exons” that encode the antigen recognition site of HLA proteins (exons 2-3 for class I genes and exon 2 for class II genes) are typically sequenced [11]. SBT is augmented with sequence specific primer (SSP) or sequence specific oligonucleotide (SSO) probe technologies to resolve ambiguities [12]. NGS platforms generate many more sequence reads than SBT instruments, allowing non-core exons, introns and untranslated regions to be sequenced in addition to core exons. As a result, NGS platforms can return full-length (four-field) alleles and detect novel alleles [13]. NGS technologies also permit high-throughput HLA typing for large numbers of samples in a cost effective manner [14], permitting large-scale studies. The ability to obtain high-resolution HLA typing using NGS is quickly expanding our knowledge of genetic variation for HLA genes.

Genes on a given chromosome are said to be linked, if alleles at respective genes do not assort independently, those alleles are said to be in linkage disequilibrium (LD) [15]. The HLA-C and HLA-B genes are situated within a 90-kb region at chromosome 6p21.33 [2]. Allele combinations of these two genes are often preserved, and are likely to have been derived from a shared ancestral chromosome segment, due to LD. The LD between HLA-B and HLA-C is often called the HLA-B~C haplotype block [16]. Similar to the HLA-B~C block, HLA-DRB3/4/5, HLA-DRB1, HLA-DQA1 and HLA-DQB1 genes within HLA class II region are located in a 150 - 210-kb region at chromosome 6p21.32 [2]. As a consequence, alleles of these genes are also in strong LD, and constitute the HLA-DR~DQ block [17].

Haplotype frequencies have been estimated and reported in various publications [18-23]. Accurate haplotype frequency estimation is of importance for hematopoietic stem cell donor match prediction and for helping more patients identify suitably matched donors. Bioinformatics groups validated various computational tools for haplotype frequency estimation using data sets derived from hematopoietic stem cell donor registries in France, Germany, The Netherlands, UK and United States [24].

More recently, haplotype frequencies were estimated for 5 broad and 21 detailed race categories in 6.59 million individuals using an expectation–maximization (EM) algorithm [25]. It has been recognized that haplotypes follow a heavy tail distribution across all population/racial groups [26]. In addition haplotype frequencies were overestimated when sample sizes were small [27]. Therefore, some of the rare haplotypes in the reference table may not be real, or haplotype frequencies for some population/racial groups may be overestimated. Nevertheless, it is meaningful to review potential haplotypes from HLA genotypes. Based on the haplotype frequency information that we used as “reference” haplotypes [25], it is feasible to expect to observe specific “reference” haplotypes for HLA-B and -C, and -DR and -DQ alleles. In addition to these reference haplotypes, specific HLA alleles have been previously characterized as belong to “common” and “well-documented” (CWD) categories [28].

HLA typing using NGS is generally performed using commercially available HLA typing software. Although the software automatically generates the first pass of HLA typing, it is laborious to review the HLA typing from NGS platforms due to: 1) large numbers of sequence reads; 2) frequent contamination of sequence reads from other genes (e.g., pseudogenes); 3) inclusion of non-core exons and introns; and 4) the large number of samples processed. Any potential HLA typing errors have to be identified by manual inspection, and then corrected by manual edits and/or supplementary experiments.

Many factors contribute to causing HLA typing errors. For instance, HLA typing errors are often caused by shallow sequence coverage, allelic sequence imbalances or complete DNA sequence dropouts. These are generally triggered by biased allelic amplification. Unusual haplotypes can be predicted using reference haplotype frequencies from the previous study [25]. The unusual haplotypes may be real, or may be caused by HLA typing errors. Using this logic, potential HLA typing errors could be identified by the presence of rare HLA alleles that are not CWD or the presence of unusual haplotypes, and corrected by confirmatory secondary experiments. It is time-consuming to manually search for such unusual alleles and/or haplotypes. In addition, such a search requires extensive experience and knowledge of HLA alleles and haplotypes. Those who are HLA novices spend extensive hours reviewing their data to scan through the CWD list and haplotype frequency lists for potential errors. This level of inspection of DNA sequence alignments may result in a reviewer making manual changes to their results, which may lead to different reviewers generating different HLA typing results.

The HLA community recognized the need of a tool identifying HLA typing errors even before NGS was applied to HLA. For example, it was suggested and discussed to detect HLA typing errors in Registries in US, UK, France, Netherlands and Germany at the International HLA and Immunogenetics Workshop (IHIW) by the World Marrow Donor Association (WMDA) working group [29].

In order to identify such errors or unusual haplotypes in a systematic way, we have developed software, “HLA Haplotype Validator (HLAHapV)”, which checks each allele against the CWD catalogue, then reports reference haplotypes for HLA-B and HLA-C, and for HLA-DRB3/4/5, HLA-DRB1, HLA-DQA1 and HLA-DQB1. The software generates

warning reports when orphan alleles, which do not belong to any reference haplotypes, are found, resulting in the formation of unusual haplotypes. In addition, the software calculates the likelihood of each haplotype pair, if multiple haplotypes are found from the allele combinations, and ranks each haplotype combination. These reports provide increased confidence regarding the accuracy of the HLA typing, when reference haplotypes are found. It also provides more time for careful analysis (and potentially re-typing) of unusual HLA alleles or haplotypes for potential errors in the HLA typing. It is important to note that the HLAHapV software is used as a validator of observable haplotypes in genotype data, and not a genotyping validator. However, genotyping errors could be revealed when haplotypes were not confirmed from the reference haplotype table as demonstrated below.

2. Materials and Methods

2.1. Development of HLA Haplotype Validator (HLAHapV)

To identify and isolate unusual alleles and haplotypes from HLA genotyping data in an automated manner, we have developed computer software, which we have named “HLA Haplotype Validator (HLAHapV)”. The software has been developed using Java 1.7, and is available via GitHub (<https://github.com/nmdp-bioinformatics/ImmunogeneticDataTools>). The software is accompanied by built-in JUnit tests (<http://junit.org/>), to serve as a basic regression suite, in order to mitigate against introduction of software defects.

2.2. IMGT (IMMunoGeneTics)/HLA Database

The software uses IMGT (IMMunoGeneTics)/HLA Database v3.20.0 as the default database, but the user can provide a specific database version that is used to generate HLA typing data via a runtime argument (Figure 1) [8].

2.3. Common and Well-Documented alleles (CWD)

The software reviews the CWD 2.0.0 catalogue to determine which alleles listed in the genotype data are CWD [28]. To overcome differences between the database versions that were used for generating HLA genotype and current version of HLA database, the software first looks up a specific HLA allele in the default or user-specified database version to convert the allele name to be a specific accession number using a lookup table (ftp://ftp.ebi.ac.uk/pub/databases/imgt/mhc/hla/Allelelist_history.txt). The software then searches for the specific allele in the CWD table using the accession number. If an allele is not found in the CWD table, the allele is reported in the “nonCwdWarnings.log” file.

2.4. Haplotype tables

We used the “Be The Match Registry Haplotype Frequencies” table (<https://bioinformatics.bethematchclinical.org/HLA-Resources/Haplotype-Frequencies/Be-The-Match-Registry-Haplotype-Frequencies/>) as the reference table for the software [25]. This table includes haplotype frequencies estimated for 6.59 million individuals in 5 broad and 21 detailed race categories using an expectation–maximization (EM) algorithm [25]. The haplotype reference table, which used the IMGT/HLA database version 3.4.0, contains alleles that are described with the first two fields of HLA allele nomenclature, representing protein level assignment [25]. Alleles with amino acids identical in the antigen recognition

were combined as a group represented by “g” at the end of the allele name [25]. To accommodate more recent IMGT/HLA database version, the software is capable of using “G group allele code” table (<https://bioinformatics.bethematchclinical.org/hla-resources/allele-codes/>) with IMGT/HLA database version 3.15.0 as a runtime argument. The HLAHapV software converts four field allele names obtained from the NGS genotyping to the two field allele names using the conversion table. The software searches all possible HLA-B~C haplotypes constructed from HLA-B and HLA-C alleles, and HLA-DR~DQ haplotypes from HLA-DRB3/4/5, HLA-DRB1 and HLA-DQB1 alleles in the genotype data against the reference haplotypes. The software is capable of producing information for any set of loci, provided that a frequency file is provided and configured. We tested the software including HLA-A genotypes, and generating reports of haplotypes for HLA-A~B~C and HLA-DR~DQ. When matches are found, the software extracts the frequency of the matching haplotypes for each population/racial group from the reference table and reports them in either the “linkages.log” or “linkageWarnings.log” file. When at least four haplotypes are found, two from each haplotype block, the haplotypes are reported in the “linkages.log” file. The software reports frequencies of all possible haplotypes for 5 broad and 21 detailed population/racial groups separately in “linkages.log” file. When zero or one haplotypes are found for either haplotype block, the subject is reported in the “linkageWarnings.log” file. Homozygotes are evaluated as having two haplotypes for each haplotype block.

2.5. Haplotype Pair Ranking Using Relative Haplotype Frequencies

The software initially captures every reference haplotype found in the reference haplotype table, it may, therefore, report more than one haplotype per allele in the “linkages.log” file. In addition to reporting reference haplotypes, the software provides a haplotype combination rank calculated from the frequencies of each haplotype for each population/racial group. The software calculates the likelihood of the different possible haplotypes relative to one another and their populations as a whole.

For example, HLA-B and HLA-C genotypes of individuals with alleles B_1 and B_2 at HLA-B and alleles C_1 and C_2 at HLA-C are described as $B_1 + B_2 \wedge C_1 + C_2$. There are two possible HLA-B~C haplotype combinations: $B_1 \sim C_1 + B_2 \sim C_2$ and $B_1 \sim C_2 + B_2 \sim C_1$. Taking the probability $p_\theta(B_i \sim C_j)$ to be the haplotype frequency for HLA alleles B_i and C_j respectively in population θ , we can then estimate the probability of each haplotype combination as follows:

$$\begin{aligned}\hat{p}_\theta(B_1 \sim C_1 + B_2 \sim C_2) &= p_\theta(B_1 \sim C_1) \cdot p_\theta(B_2 \sim C_2) \\ \hat{p}_\theta(B_1 \sim C_2 + B_2 \sim C_1) &= p_\theta(B_1 \sim C_2) \cdot p_\theta(B_2 \sim C_1)\end{aligned}$$

We can then define the likelihood $\mathcal{L}_\theta$ of each individual combination in this population as follows:

$$\begin{aligned}\mathcal{L}_\theta(B_1 \sim C_1 + B_2 \sim C_2) &= \frac{\hat{p}_\theta(B_1 \sim C_1 + B_2 \sim C_2)}{\hat{p}_\theta(B_1 \sim C_1 + B_2 \sim C_2) + \hat{p}_\theta(B_1 \sim C_2 + B_2 \sim C_1)} \\ \mathcal{L}_\theta(B_1 \sim C_2 + B_2 \sim C_1) &= \frac{\hat{p}_\theta(B_1 \sim C_2 + B_2 \sim C_1)}{\hat{p}_\theta(B_1 \sim C_1 + B_2 \sim C_2) + \hat{p}_\theta(B_1 \sim C_2 + B_2 \sim C_1)}\end{aligned}$$

In cases of allelic ambiguity, the denominator of this expression would sum the estimated probabilities of each possible haplotype pairings. This allows the likelihood of each possible haplotype combination to be described relative to each other, in the case that there is more than one reference haplotype found for each allele. The estimated haplotype pairs are reported in the “haplotypePairs.log” file. The software reports frequencies of all possible haplotype combinations for 5 broad and 21 detailed population/racial groups separately in “haplotypePairs.log” file. When zero or one reference haplotypes are found from either haplotype block, the subject is reported in the “haplotypePairWarnings.log” file. The “linkageWarnings.log” and “haplotypePairWarnings.log” files contain the same subjects. The “linkageWarnings.log” and “haplotypePairWarnings.log” files should be carefully reviewed for any potential HLA genotyping errors. Note that the identified haplotypes and haplotype combinations are based on a pure statistical prediction, not experimentally determined in the lab. Some of the rare haplotypes in the reference table may not be real.

2.6. Executing the software

The HLAHapV software can be obtained from GitHub (<https://github.com/nmdp-bioinformatics/ImmunogeneticDataTools>), and cloned to a local computer. Compilation, packaging and execution of the software are managed using Maven (<https://maven.apache.org>). The software takes one or more file names containing HLA genotyping data as a first command line argument. The file format includes sample name and HLA genotyping data separated by common delimiters (tab or comma). The HLA genotyping data is represented using Genotype List (GL) String format [30]. The software takes a second, optional argument to specify one of the supported haplotype tables. It is possible to specify one of seven IMGT/HLA Database versions (3.20.0, 3.19.0, 3.18.0, 3.15.0, 3.12.0, 3.11.0 or 3.10.0) as an optional third argument. The software uses 3.20.0 as a default database version if the third argument is not provided. As new IMGT/HLA Database versions are released, and are correlated to the haplotype frequencies, the software will support them. Updates will continue to be available from GitHub. An example of command line is shown in Figure 1. Six output files are generated (Supplemental Materials): 1) nonCwdWarnings.log; 2) haplotypePairs.log; 3) haplotypePairWarnings.log; 4) linkages.log; 5) linkageWarnings.log; and 6) immuno.log. The contents of the first five files are described above. The immuno.log is a basic error-logging file, identifying issues such as an invalid GL string format or unexpected runtime errors. The software takes any HLA typing as long as the input is formatted with GL string. However, we do not recommend testing HLA genotypes containing very long ambiguity string that are generated from non-core exon sequences due to core exon sequence dropout. Such long ambiguity string consumes unnecessary computational times to calculate all possible haplotype combinations in the GL string.

2.7. Validation of the software

HLA genotypes were generated from the sequences of the following exons: HLA-A exons 1-5, HLA-B exons 1-5, HLA-C exons 1-7, HLA-DPA1 exons 2-3, HLA-DPB1 exons 2-3, HLA-DQA1 exons 2-3, HLA-DQB1 exons 2-3 and HLA-DRB1/3/4/5 exons 2-3. DNA sequences were obtained using the GS FLX+ DNA sequencing system (Roche 454). HLA genotypes were generated using SeqNext-HLA software (JSI medical systems) with IMGT database 3.10.0. HLA genotypes were exported as XML files using talkMaster software

connected with the SeqNext-HLA software. Each HLA genotype is obtained from the XML files, and organized in GL String format using programs developed using Java and Perl. We analyzed HLA genotypes derived from 150 subjects, organized in 50 trios using HLAHapV software.

3. Results

3.1. Identifying potential HLA typing error using HLAHapV software

We used HLA genotyping data from 150 subjects (Subjects 1 – 150) from 50 families (50 trios) to demonstrate the utility of the HLAHapV software. We observed 44 HLA-B alleles including 1 novel allele, 25 HLA-C, 17-HLA-DQB1, 25 HLA-DRB1, 3 HLA-DRB3, 4 HLA-DRB4 and 3 HLA-DRB5 from these 150 subjects. Of 150 subjects, 145 subjects were reported without any warning message in the “haplotypePairs.log” file, and 5 subjects were reported in the “haplotypeWarnings.log” file. The same subsets of subjects are reported in “linkages.log” and “linkageWarnings.log” files, respectively.

We visually inspected each case, and found that 134 of 145 subjects had heterozygous haplotypes for both the HLA-B~C and HLA-DR~DQ haplotype blocks in the “haplotypePairs.log” file. Of the remaining 11 subjects, 6 subjects were homozygous for the HLA-DR~DQ haplotype block, and 5 subjects were homozygous for both the HLA-B~C and HLA-DR~DQ blocks. We confirmed that the software treated each homozygous haplotype as two haplotypes.

Of 145 subjects reported in haplotypePairs.log file, 138 subjects consist of 46 trios. These families had consistent allele inheritance, as well as fairly stable HLA-B~C and HLA-DR~DQ haplotype block inheritance from the parents to the children. Table 1 shows an inheritance of HLA-B~C and HLA-DR~DQ haplotype blocks from parents to a child in a single family.

Of the 5 subjects with warning messages, 2 subjects had unknown or “novel” alleles for HLA-B gene. These two subjects had a parent and a child relationship in the same family confirming the transmission of a novel HLA-B allele from the parent to the child. Interestingly, all three remaining subjects (S39, S95 and S134) had HLA-B*38:27. We reviewed these three cases in detail as follows.

3.1.1. Subject 39—For subject 39, we found a pair of HLA-DR~DQ haplotypes with a high estimated relative frequency in the “haplotypeWarnings.log” file. These haplotypes are HLA-DRB1*07:01~HLA-DRB4*01:01g~HLA-DQB1*02:01g + HLA-DRB1*04:02~HLA-DRB4*01:01g~HLA-DQB1*03:02g. The HLA-B and HLA-C genotypes of the subject 39 were reported as: HLA-B*14:02:01+HLA-B*38:27^HLA-C*08:02:01/HLA-C*08:34+HLA-C*12:03:01:01/HLA-C*12:03:01:02/HLA-C*12:34 (Table 2A). When the “linkageWarnings.log” file is reviewed, only one haplotype, HLA-B*14:02~HLA-C*08:02, was reported, but no HLA-B*38:27~C*12:03 haplotype was identified in the reference table by the HLAHapV software (Table 2A). We reviewed the original DNA sequence assignment in SeqNext-HLA and found that only 7 of 195 sequence reads (3.6%) for HLA-B exon5 were manually allocated as a second allele and used to call HLA-B*38:27 (Figure 2A). This

small fraction of DNA sequences likely represents sequencing “noise”, and we most likely identified these sequences as a second allele in error. The sequences of HLA-B*38:01:01 were excluded as a possibility, because the “noise” sequence does not match the exon5 sequence of HLA-B*38:01:01. Only exon 2 and exon 3 sequences are available for HLA-B*38:27 in the IMGT/HLA Database. We were not able to distinguish HLA-B*38:27 from HLA-B*38:01:01 using exons 2 and 3 sequences, because our exon2 primer overlaps with the 3'-end of exon2 and the nucleotide that distinguishes HLA-B*38:27 from HLA-B*38:01:01 could not be evaluated [31]. We concluded that the HLA-B types of the subject 39 should be reported as: HLA-B*14:02:01+ HLA-B*38:01:01/HLA-B*38:27.

Subject 39 is a child of subjects 37 and 38. To confirm our findings, we further reviewed the parental HLA typing from subjects 37 and 38. The HLA-B and HLA-C genotypes of subject 37 were reported as: HLA-B*38:01:01/HLA-B*38:27+HLA-B*49:01:01^HLA-C*07:01:31/HLA-C*07:103/HLA-C*07:230+HLA-C*12:03:01:01/HLA-C*12:03:01:02/HLA-C*12:34 (Table 2A). The HLAHapV reported the highest-ranking haplotype combination for the HLA-B~C haplotype block for subject 37 is HLA-B*38:01~HLA-C*12:03 + HLA-B*49:01~HLA-C*07:01 (Table 2A). The HLA-B and HLA-C genotypes of subject 38 were reported as: HLA-B*14:02:01+HLA-B*35:02:01/HLA-B*35:211^HLA-C*04:01:01:01/HLA-C*04:01:01:02/HLA-C*04:01:01:03/HLA-C*04:01:01:04/HLA-C*04:01:01:05/HLA-C*04:20/HLA-C*04:117+HLA-C*08:02:01/HLA-C*08:34 (Table 2A). Similarly, the HLAHapV reported the highest-ranking haplotype combination for HLA-B~C haplotype block for subject 38 is HLA-B*35:02~HLA-C*04:01g + HLA-B*14:02~HLA-C*08:02 (Table 2A). These parental HLA-B~C haplotype estimations assured the HLA typing error of HLA-B in subject 39, thus HLA-B types of the subject 39 should be reported as: HLA-B*14:02:01+ HLA-B*38:01:01/HLA-B*38:27.

3.1.2. Subject 134—Subject 134 is a parent of subject 135. We found a pair of HLA-DR~DQ haplotypes with a high estimated relative frequency in the “haplotypeWarnings.log” file for subject 134. These haplotypes are HLA-DRB1*07:01~HLA-DRB4*01:01g~HLA-DQB1*02:01g + HLA-DRB1*04:02~HLA-DRB4*01:01g~HLA-DQB1*03:02g. The HLA-B and HLA-C types of the subject 134 were reported as: HLA-B*14:02:01+HLA-B*38:27^HLA-C*08:02:01/HLA-C*08:34+HLA-C*12:03:01:01/HLA-C*12:03:01:02/HLA-C*12:34 (Table 2B). We inspected the original DNA sequence assignment data in SeqNext-HLA for subject 134, and found that only 6 of 397 (1.5%) sequence reads for HLA-B exon5 were manually allocated as a second allele and used to call HLA-B*38:27 (not shown). This is the same error that we found in subject 39. To verify our findings, we reviewed the HLA-B and HLA-C genotypes of subject 135: HLA-B*18:01:01:01/HLA-B*18:01:01:02/HLA-B*18:51+HLA-B*38:01:01/HLA-B*38:27^HLA-C*12:03:01:01/HLA-C*12:03:01:02/HLA-C*12:34+HLA-C*12:03:01:01/HLA-C*12:03:01:02/HLA-C*12:34. The HLA-B genotypes of subject 135 confirm that the HLA-B genotypes of subject 134 should be reported as: HLA-B*14:02:01+ HLA-B*38:01:01/HLA-B*38:27 (Table 2B).

These subject 39 and 134 examples represent the utility and power of the HLAHapV software for identifying the potential HLA genotyping errors. In other words, the HLAHapV software correctly identified subjects who may contain potential HLA typing errors.

3.1.3. Subject 95—We reviewed the original DNA sequence assignment of subject 95 in SeqNext-HLA and found that about 50% of sequence reads for HLA-B exon5 were automatically allocated as a second allele and used to call HLA-B*38:27 (Figure 2B). Unlike above two cases, the exon5 sequence in this case is unlikely to be “noise”. We found a pair of HLA-DR~DQ haplotypes with a high estimated relative frequency in the “haplotypeWarnings.log” file for subject 95. These haplotypes are HLA-DRB1*13:01~HLA-DRB3*01:01~HLA-DQB1*06:03g + HLA-DRB1*11:04~HLA-DRB3*02:02g~HLA-DQB1*03:01g. The HLA-B and HLA-C genotypes of subject 95 were reported as: HLA-B*13:02:01/HLA-B*13:02:11+HLA-B*38:27^HLA-C*06:02:01:01/HLA-C*06:02:01:02+HLA-C*12:03:01:01/HLA-C*12:03:01:02/HLA-C*12:34. (Table 2C). Subject 95 is a parent of subject 96. The HLA-B and HLA-C types of subject 96 were reported as: HLA-B*13:02:01/HLA-B*13:02:11+HLA-B*14:02:01^HLA-C*06:02:01:01/HLA-C*06:02:01:02+HLA-C*08:02:01/HLA-C*08:34 (Table 2C). HLAHapV reported the highest-ranking haplotype combination for the HLA-B~C haplotype block for subject 96 as HLA-B*13:02g~HLA-C*06:02g + HLA-B*14:02~HLA-C*08:02. It appears that Subject 96 inherited the HLA-B*13:02g^HLA-C*06:02g haplotype from subject 95. Therefore, we were not able to use the trio data to determine if the HLA-B*38:27 allele assignment is correct for subject 95.

3.2. Haplotype pair estimation

Of 150 subjects analyzed, 145 subjects were reported in the “haplotypePairs.log” and “linkages.log” files, and 5 subjects were reported in the “haplotypeWarnings.log” and “linkageWarnings.log” files. From the initial analysis we found that 46 families had consistent allele inheritance from the parents to the children. As described above, two of 5 subjects in the “haplotypeWarnings.log” file had a novel HLA-B allele, and had a parent and a child relationship in the same family. We also corrected HLA-B typing errors from subjects 39 and 134 as described in sections 3.1.1 and 3.1.2, these families displayed consistent HLA allele inheritance from the parents to child. We were not able to find reference haplotypes from subject 95, but we confirmed correct HLA allele inheritance from the parents (subject 94 and subject 95) to a child (subject 96) (Table 2C). We concluded that all 50 families displayed consistent HLA allele inheritance from the parents to the children based on the visual inspection of each HLA allele for each gene from each subject.

We further reviewed the initial 46 families that were originally reported in the “haplotypePairs.log” file. We found that a HLA-B~C and HLA-DR~DQ haplotype from the highest-estimated relative frequency haplotype pair of the parents is transmitted to the child and forms the highest-estimated relative frequency haplotype pairs of the child in 44 of 46 families. The results supports recognized evidence of LD between the HLA-B and -C and HLA-DR and -DQ loci, which is the foundation of the HLAHapV software, and indicate the high confidence of the relative frequency haplotype pair estimates in the reference haplotype data. In addition to the expected haplotype inheritances, unusual haplotype inheritances are observed in two families. These families are further described in section 4.2.

4. Discussion

We have developed software that is primarily intended for quality control (QC) and quality assessment (QA) of HLA genotypes obtained from DNA sequences generated from the NGS systems. The software, HLA Haplotype Validator (HLAHapV), attempts the following four steps.

First, the software passes the HLA genotypes through the GL Service to check the accuracy of GL String format [30]. Second, the individual HLA alleles are filtered against the CWD 2.0.0 catalogue [28]. For example, an erroneous HLA allele name that may be generated by a typographical error or other reasons [32] is identified and reported at this stage. Third, reference haplotypes for the HLA-B~C and HLA-DR~DQ haplotype blocks, derived from a haplotype table that contains frequency data for 6.59 million individuals assigned to 5 broad and 21 detailed race categories are searched [25]. Lastly, the highest-estimated relative frequency haplotype pairs are reported first in order of the HLA-B~C haplotype, followed by HLA-DR~DQ haplotype.

4.1 Identification of Potential Genotyping Errors

We identified two potential HLA-B genotyping errors from subjects 39 and 134 using HLAHapV software (Table 2AB). It is important to correct such errors before HLA genotypes are used for downstream analyses. The HLAHapV software automatically generated a small list of subjects with HLA types that needed to be reviewed, thereby saving labor.

In addition, subject 95, for which HLA-B*38:27 was reported, is listed in both the “linkageWarnings.log” and “haplotypePairs.log” files. The HLA-B and HLA-C genotypes of subject 95 were reported as: HLA-B*13:02:01/HLA-B*13:02:11+HLA-B*38:27^HLA-C*06:02:01:01/HLA-C*06:02:01:02+HLA-C*12:03:01:01/HLA-C*12:03:01:02/HLA-C*12:34 (Table 2C). The HLA-B*13:02g~HLA-C*06:02g haplotype for subject 95 is found as one possible haplotype in the “linkageWarnings.log” file. Thirty-four haplotypes containing HLA-B*38:27 and 444 haplotypes containing HLA-C*12:03g are reported in the reference HLA-B~C haplotype block table, but the HLA-B*38:27~HLA-C*12:03g haplotype is not found among the reference haplotypes [25]. Although the software flagged subject 95 due to the identification of rare or unknown haplotypes, the HLA typing data of subject 95 appears to be correct. It is valuable to point out that there is a possibility that the HLA-B*38:27 of subject 95 may not be correct, because only exon 2 and exon 3 sequences are available for HLA-B*38:27 in the IMGT/HLA Database. The SeqNext-HLA software subjectively assigned HLA-B*38:27 without any sequence information of exon 5 of HLA-B*38:27. It is possible that this HLA-B allele of subject 95 may be identified as a novel HLA-B*38 allele when exon 5 sequence for HLA-B*38:27 becomes available.

4.2. Identification of Unexpected Haplotype Transmission

We observed the transmission of unexpected haplotypes in two families. For the first family, subject 70 and 71 are the parents of subject 72. The HLAHapV software reported HLA-B*52:01g~HLA-C*12:02 + HLA-B*50:01~HLA-C*06:02g as the highest-estimated relative

frequency haplotype pair for subject 70's HLA-B~C haplotype block. Similarly, the software reported HLA-B*44:03~HLA-C*16:01 + HLA-B*57:01g~HLA-C*06:02g as the highest-estimated relative frequency haplotype pair for subject 71's HLA-B~C haplotype blocks. Interestingly, the software reported HLA-B*52:01g~HLA-C*06:02g + HLA-B*57:01g~HLA-C*06:02g is the only one possible haplotype pair for subject 72. Examination of the primary sequence data eliminates the possibility of HLA-C*12:02 having been incorrectly assigned as HLA-C*06:02g in subject 72. Although HLA-B*50:01~HLA-C*12:02 + HLA-B*52:01g~HLA-C*06:02g is listed as a possible haplotype pair for subject 70, the frequency of this pair is 7.32×10^{-12} while the frequency of the highest-estimated relative frequency haplotype pair, HLA-B*52:01g~HLA-C*12:02 + HLA-B*50:01~HLA-C*06:02g, is 1.19×10^{-4} (Table 3A).

For the second family, subject 76 and 77 are the parents of subject 78. The HLAHapV software reported HLA-B*52:01g~HLA-C*12:02 + HLA-B*50:01~HLA-C*06:02g as the highest-estimated relative frequency (1.19×10^{-4}) haplotype pair for subject 76's HLA-B~C haplotype blocks (Table 3B). Another haplotype pair, HLA-B*50:01~HLA-C*12:02 + HLA-B*52:01g~HLA-C*06:02g, is listed as a possible haplotype pair with much lower relative frequency (7.32×10^{-12}) for subject 76 as subject 70 (see above). Similarly, the software reported HLA-B*53:01~HLA-C*04:01g + HLA-B*14:02~HLA-C*08:02 as the highest-ranking haplotype pair for subject 77's HLA-B~C haplotype blocks. The software reported HLA-B*14:02~HLA-C*08:02 + HLA-B*52:01g~HLA-C*06:02g is the highest-ranking haplotype pair for subject 78 (Table 3B).

In these two families, both children inherited HLA-B*52:01g~HLA-C*06:02g. The frequency of this HLA-B~C haplotype in the reference Caucasian race category is 1.22×10^{-5} and its ranking in this race category is 436th [25]. It is not possible to know whether the parent inherited this rare haplotype or if the rare haplotype was formed during meiotic recombination when HLA alleles are transmitted from the parent to the child. These examples suggest that haplotype prediction, and haplotype pair estimation will not be perfect; in other words, there are always exceptions that do not fit the established models. Users of the HLAHapV software will always need to be aware of the presence of such exceptions.

4.3. Inclusion of HLA-A in the haplotype analysis

The software is capable of producing information for any set of loci, provided that a frequency file is provided and configured. We tested the software including HLA-A genotypes, extracting HLA-A~C~B haplotype frequencies from the reference table and estimating HLA-A~C~B haplotype combinations. We did not find additional subjects by adding HLA-A genotypes in the analyses (data not shown). It is of importance to note that we observed more possible haplotype combinations when HLA-A was included in the haplotype combination analyses. For example, Subject 3 had HLA-A, HLA-B, HLA-C genotypes as follows: HLA-A*01:01g+HLA-A*32:01~HLA-B*41:01+HLA-B*44:03~HLA-C*04:01g+HLA-C*17:01g. The relative frequencies of HLA-B~C haplotype combination for HLA-B*41:01~HLA-C*17:01g + HLA-B*44:03~HLA-C*04:01g were nearly 100%, while HLA-B*41:01~HLA-C*04:01g + HLA-B*44:03~HLA-C*17:01g were

0% for all 5 broad race group. When HLA-A was included in the analysis, the relative frequencies of haplotype combinations for HLA-A*01:01g~HLA-B*41:01~HLA-C*17:01g + HLA-A*32:01~HLA-B*44:03~HLA-C*04:01g were 63.94% for AFA, 97.51% for API, 53.07% for CAU, 51.10% for HIS and 50.89% for NAM, respectively. The relative frequencies of haplotype combinations for HLA-A*01:01g~HLA-B*44:03~HLA-C*04:01g + HLA-A*32:01~HLA-B*41:01~HLA-C*17:01g were 36.06% for AFA, 2.49% for API, 46.93% for CAU, 48.90% for HIS and 49.11% for NAM, respectively. These results confirm the evidence that HLA-C and HLA-B are in very strong LD, while HLA-A and HLA-C~B haplotype block are in weaker LD. Adding other loci that are not in strong LD, such as HLA-A, might not provide as distinctive information as the haplotype blocks in strong LD, thus likely less effective for identifying genotype errors.

4.4. Rare Haplotypes

When an observed pair of haplotype combinations is not found in the reference haplotype table, the subject is reported in Warnings files. Contrarily, when the software finds a pair of haplotype combinations in the reference table, the subject is reported in “haplotypePairs.log” file. The “linkages.log” reports haplotype frequencies of each population/racial group. The “haplotypePairs.log” file reports relative frequencies of each haplotype combination of each population/racial group. As noted in the Introduction, some of the rare haplotypes in the reference table may not be real [26, 27]. It is important to note that “warning” does not mean incorrect and that “no warning” does not mean correct. The software simply reports potentially erroneous haplotypes. It is up to the users to decide how to use and interpret the reports.

4.5. Conclusions

In summary, we have demonstrated that the HLAHapV software is a powerful validation tool for identifying potential HLA genotyping errors. It is difficult and labor intensive to identify erroneous HLA alleles and HLA typing errors in large HLA genotyping datasets. The software provides an opportunity to generate haplotype-validated HLA typing data in a systematic fashion, resulting in reduced labor costs for HLA genotyping.

The performance of the software will be tested against larger data sets when more genotype data with NGS become available (for example, in the 17th International HLA Workshop), and the algorithms of the software will be adjusted as needed in the future. The true value of the software will be more apparent when the community uses the software with a variety of data sets.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgements

The authors thank Dr. Martha Ladner (Children's Hospital Research Institute) for contributing to HLA typing. The work described here was performed with the support of National Institutes of Health (NIH) grants U01AI067068 (KO, SJM, JU, DAN and ET) awarded by the National Institute of Allergy and Infectious Diseases (NIAID) and R01GM19030 (SJM) awarded by the National Institute of General Medical Sciences (NIGMS). The content is

solely the responsibility of the authors and does not necessarily represent the official views of the NIH, NIGMS, NIAID or United States Government.

Abbreviations

HLA	Human Leukocyte Antigen
GL	Genotype List
CWD	Common and Well Documented
LD	Linkage Disequilibrium
IMGT	ImMunoGeneTics
NGS	Next Generation Sequencing
SBT	Sanger-sequencing Based Typing
SSO	Sequence Specific Oligonucleotide
SSP	Sequence Specific Priming

References

1. Mungall AJ, Palmer SA, Sims SK, Edwards CA, Ashurst JL, Wilming L, et al. The DNA sequence and analysis of human chromosome 6. *Nature*. 2003; 425:805. [PubMed: 14574404]
2. Stewart CA, Horton R, Allcock RJ, Ashurst JL, Atrazhev AM, Coghill P, et al. Complete MHC haplotype sequencing for common disease gene mapping. *Genome Res*. 2004; 14:1176. [PubMed: 15140828]
3. Shiina T, Hosomichi K, Inoko H, Kulski JK. The HLA genomic loci map: expression, interaction, diversity and disease. *J Hum Genet*. 2009; 54:15. [PubMed: 19158813]
4. Horton R, Wilming L, Rand V, Lovering RC, Bruford EA, Khodiyar VK, et al. Gene map of the extended human MHC. *Nat Rev Genet*. 2004; 5:889. [PubMed: 15573121]
5. Adamek M, Klages C, Bauer M, Kudlek E, Drechsler A, Leuser B, et al. Seven novel HLA alleles reflect different mechanisms involved in the evolution of HLA diversity: description of the new alleles and review of the literature. *Hum Immunol*. 2015; 76:30. [PubMed: 25500251]
6. Martinez-Laso J, Herraiz MA, Vidart JA, Penalzoza J, Barbolla ML, Jurado ML, et al. Polymorphism of the HLA-B*15 group of alleles is generated following 5 lineages of evolution. *Hum Immunol*. 2011; 72:412. [PubMed: 21376098]
7. von Salome J, Gyllenstein U, Bergstrom TF. Full-length sequence analysis of the HLA DRB1 locus suggests a recent origin of alleles. *Immunogenetics*. 2007; 59:261. [PubMed: 17345114]
8. Robinson J, Halliwell JA, Hayhurst JD, Flicek P, Parham P, Marsh SG. The IPD and IMGT/HLA database: allele variant databases. *Nucleic Acids Res*. 2015; 43:D423. [PubMed: 25414341]
9. Erlich H. HLA DNA typing: past, present, and future. *Tissue Antigens*. 2012; 80:1. [PubMed: 22651253]
10. Erlich HA. HLA typing using next generation sequencing: An overview. *Hum Immunol*. 2015
11. Lazaro A, Tu B, Yang R, Xiao Y, Kariyawasam K, Ng J, et al. Human leukocyte antigen (HLA) typing by DNA sequencing. *Methods Mol Biol*. 2013; 1034:161. [PubMed: 23775737]
12. Dunckley H. HLA typing by SSO and SSP methods. *Methods Mol Biol*. 2012; 882:9. [PubMed: 22665226]
13. Holcomb CL, Hoglund B, Anderson MW, Blake LA, Bohme I, Egholm M, et al. A multi-site study using high-resolution HLA genotyping by next generation sequencing. *Tissue Antigens*. 2011; 77:206. [PubMed: 21299525]

14. Moonsamy PV, Williams T, Bonella P, Holcomb CL, Hoglund BN, Hillman G, et al. High throughput HLA genotyping using 454 sequencing and the Fluidigm Access Array System for simplified amplicon library preparation. *Tissue Antigens*. 2013; 81:141. [PubMed: 23398507]
15. Slatkin M. Linkage disequilibrium--understanding the evolutionary past and mapping the medical future. *Nat Rev Genet*. 2008; 9:477. [PubMed: 18427557]
16. Ceppellini, R.; Curtioni, ES.; Mattiuz, PL.; Miggiano, V.; Scudeller, G.; Serra, A. Genetics of leukocyte antigens: a family study of segregation and linkage. Munksgaard; Copenhagen: 1967.
17. Fernandez Vina MA, Hollenbach JA, Lyke KE, Szein MB, Maiers M, Klitz W, et al. Tracking human migrations by the analysis of the distribution of HLA alleles, lineages and haplotypes in closed and open populations. *Philos Trans R Soc Lond B Biol Sci*. 2012; 367:820. [PubMed: 22312049]
18. Maiers M, Gragert L, Klitz W. High-resolution HLA alleles and haplotypes in the United States population. *Hum Immunol*. 2007; 68:779. [PubMed: 17869653]
19. Klitz W, Gragert L, Maiers M, Fernandez-Vina M, Ben-Nach Y, Benedek G, et al. Genetic differentiation of Jewish populations. *Tissue Antigens*. 2010; 76:442. [PubMed: 20860586]
20. Schmidt AH, Baier D, Solloch UV, Stahr A, Cereb N, Wassmuth R, et al. Estimation of high-resolution HLA-A, -B, -C, -DRB1 allele and haplotype frequencies based on 8862 German stem cell donors and implications for strategic donor registry planning. *Hum Immunol*. 2009; 70:895. [PubMed: 19683023]
21. Yang KL, Chen SP, Shyr MH, Lin PY. High-resolution human leukocyte antigen (HLA) haplotypes and linkage disequilibrium of HLA-B and -C and HLA-DRB1 and -DQB1 alleles in a Taiwanese population. *Hum Immunol*. 2009; 70:269. [PubMed: 19480851]
22. Schmidt AH, Solloch UV, Pingel J, Baier D, Bohme I, Dubicka K, et al. High-resolution human leukocyte antigen allele and haplotype frequencies of the Polish population based on 20,653 stem cell donors. *Hum Immunol*. 2011; 72:558. [PubMed: 21513754]
23. Qin Q, Su F, Xiao Y, Xing Z, Meng P, Chengya W, et al. Distribution of human leukocyte antigen-A, -B and -DR alleles and haplotypes at high resolution in the population from Jiangsu province of China. *Int J Immunogenet*. 2011; 38:475. [PubMed: 21816002]
24. Eberhard HP, Madbouly AS, Gourraud PA, Balere ML, Feldmann U, Gragert L, et al. Comparative validation of computer programs for haplotype frequency estimation from donor registry data. *Tissue Antigens*. 2013; 82:93. [PubMed: 23849067]
25. Gragert L, Madbouly A, Freeman J, Maiers M. Six-locus high resolution HLA haplotype frequencies derived from mixed-resolution DNA typing for the entire US donor registry. *Hum Immunol*. 2013; 74:1313. [PubMed: 23806270]
26. Slater N, Louzoun Y, Gragert L, Maiers M, Chatterjee A, Albrecht M. Power laws for heavy-tailed distributions: modeling allele and haplotype diversity for the national marrow donor program. *PLoS Comput Biol*. 2015; 11:e1004204. [PubMed: 25901749]
27. Pappas DJ, Tomich A, Garnier F, Marry E, Gourraud PA. Comparison of high-resolution human leukocyte antigen haplotype frequencies in different ethnic groups: Consequences of sampling fluctuation and haplotype frequency distribution tail truncation. *Hum Immunol*. 2015; 76:374. [PubMed: 25637668]
28. Mack SJ, Cano P, Hollenbach JA, He J, Hurley CK, Middleton D, et al. Common and well-documented HLA alleles: 2012 update to the CWD catalogue. *Tissue Antigens*. 2013; 81:194. [PubMed: 23510415]
29. Tiercy JM, Fischer G, Setterholm M. Quality control for registry HLA data. *Tissue Antigens*. 2007; 69(Suppl 1):13. [PubMed: 17445155]
30. Milius RP, Mack SJ, Hollenbach JA, Pollack J, Heuer ML, Gragert L, et al. Genotype List String: a grammar for describing HLA and KIR genotyping results in a text string. *Tissue Antigens*. 2013; 82:106. [PubMed: 23849068]
31. Bentley G, Higuchi R, Hoglund B, Goodridge D, Sayer D, Trachtenberg EA, et al. High-resolution, high-throughput HLA genotyping by next-generation sequencing. *Tissue Antigens*. 2009; 74:393. [PubMed: 19845894]
32. Hollenbach JA, Mack SJ, Gourraud PA, Single RM, Maiers M, Middleton D, et al. A community standard for immunogenomic data reporting and analysis: proposal for a STrengthening the

REporting of Immunogenomic Studies statement. Tissue Antigens. 2011; 78:333. [PubMed: 21988720]

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

Running a Test Data Set:

```
Command: mvn exec:java -  
Dexec.mainClass="org.dash.valid.LinkageDisequilibriumAnalyzer" -  
Dexec.args="<filename> <filename>"
```

```
Example: mvn exec:java -  
Dexec.mainClass="org.dash.valid.LinkageDisequilibriumAnalyzer" -  
Dexec.args="resources/test/hlaTypeExamples.txt" -Dorg.dash.frequencies="nmdp" -  
Dorg.dash.hladb="3.10.0"
```

Figure 1. Executing HLAHapV software

The software takes one or more file names containing HLA genotyping data as a first command line argument. The software takes a second, optional argument to specify one of the supported haplotype tables. It is possible to specify IMGT/HLA Database versions, 3.18.0, 3.15.0, 3.12.0, 3.11.0 and 3.10.0 as an optional third argument. The example shows that HLA genotyping data is stored in the file structure, “resources/test/hlaTypeExamples.txt”. We used “nmdp” as a second argument, specifying the reference haplotype dataset, and “3.10.0” for the database version that the HLA genotyping data was generated under as a third argument.

Figure 2. SeqNext-HLA software DNA sequence alignment view for HLA-B exon5

A) Seven of 195 sequence reads (the two lines indicated by the arrow) for HLA-B exon5 were manually allocated as second allele that was used to call HLA-B*38:27 on the basis of the exon 5 T113C sequence variant. This small fraction of DNA sequences is likely sequence “noise” that has most likely been identified as a second allele in error.

B) Two distinct alleles were equally amplified and sequenced for HLA-B exon5 in this case. Unlike Figure 2A, the exon5 sequence in this case is unlikely to represent sequence “noise” due to the comparable number of reads (99 vs. 83). HLA-B*38:27 was subjectively called

by SeqNext-HLA software on the basis of the exon 5 T113C sequence variant, because there is no exon5 sequence is available for HLA-B*38:27.

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

Table 1

Subject	Haplotype Block	Haplotype 1	Haplotype 2
1	HLA-B~C	HLA-B*44:03~HLA-C*04:01g	HLA-B*38:01~HLA-C*12:03g
	HLA-DR~DQ	HLA-DRB1*07:01~HLA-DRB4*01:01g~HLA-DQB1*02:01g	HLA-DRB1*11:01g~HLA-DRB3*02:02g~HLA-DQB1*03:01g
2	HLA-B~C	HLA-B*41:01~HLA-C*17:01g	HLA-B*57:03~HLA-C*06:02g
	HLA-DR~DQ	HLA-DRB1*04:03~HLA-DRB4*01:01g~HLA-DQB1*03:04	HLA-DRB1*07:01~HLA-DRB4*01:01g~HLA-DQB1*03:03g
3	HLA-B~C	HLA-B*41:01~HLA-C*17:01g	HLA-B*44:03~HLA-C*04:01g
	HLA-DR~DQ	HLA-DRB1*07:01~HLA-DRB4*01:01g~HLA-DQB1*02:01g	HLA-DRB1*07:01~HLA-DRB4*01:01g~HLA-DQB1*03:03g

Table 1 shows the highest likelihood HLA-B-C and HLA-DR~DQ haplotype pairs estimated by the HLAHapV software.
Haplotypes transmitted from parents to offspring are shown in bold.

Table 2A

Subject	HLA-B Genotype	HLA-C Genotype	HLA-B~C Haplotype 1	HLA-B~C Haplotype 2
Parent)	HLA-B*38:01:01/HLA-B*38:27+HLA-B*49:01:01	HLA-C*07:01:31/HLA-C*07:103/HLA-C*07:230+HLA-C*12:03:01:01/HLA-C*12:03:01:02/HLA-C*12:34	HLA-B*38:01~HLA-C*12:03g	HLA-B*49:01~HLA-C*07:01g
Parent)	HLA-B*14:02:01+HLA-B*35:02:01/HLA-B*35:211	HLA-C*04:01:01:01/HLA-C*04:01:01:02/HLA-C*04:01:01:03/HLA-C*04:01:01:04/HLA-C*04:01:01:05/HLA-C*04:20/HLA-C*04:117+HLA-C*08:02:01/HLA-C*08:34	HLA-B*35:02~HLA-C*04:01g	HLA-B*14:02~HLA-C*08:02
(Child)	HLA-B*14:02:01+HLA-B*38:27	HLA-C*08:02:01/HLA-C*08:34+ HLA-C*12:03:01:01/HLA-C*12:03:01:02/HLA-C*12:34	HLA-B*14:02~HLA-C*08:02	WARNING - No BC haplotype pairs detected

Table 2B

Subject	HLA-B Genotype	HLA-C Genotype	HLA-B~C Haplotype 1	HLA-B~C Haplotype 1
133 (Parent)	<u>HLA-B*15:18:01/HLA-B*15:226N+HLA-B*18:01:01/HLA-B*18:01:01/HLA-B*18:51</u>	<u>HLA-C*07:04:01/HLA-C*07:04:07+HLA-C*12:03:01:01/HLA-C*12:03:01:02/HLA-C*12:34</u>	<u>HLA-B*18:01g~HLA-C*12:03g</u>	HLA-B*15:18g~HLA-C*07:04g
134 (Parent)	<u>HLA-B*14:02:01+HLA-B*38:27</u>	<u>HLA-C*08:02:01/HLA-C*08:34+HLA-C*12:03:01:01/HLA-C*12:03:01:02/HLA-C*12:34</u>	<u>HLA-B*14:02~HLA-C*08:02</u>	WARNING - No BC haplotype pairs detected
135 (Child)	<u>HLA-B*18:01:01/HLA-B*18:01:01:02/HLA-B*18:51+HLA-B*38:01:01/HLA-B*38:27</u>	<u>HLA-C*12:03:01:01/HLA-C*12:03:01:02/HLA-C*12:34+HLA-C*12:03:01:01/HLA-C*12:03:01:02/HLA-C*12:34</u>	<u>HLA-B*18:01g~HLA-C*12:03g</u>	HLA-B*38:01~HLA-C*12:03g

Table 2C

subject	HLA-B Genotype	HLA-C Genotype	HLA-B~C Haplotype 1	HLA-B-C Haplotype 1
(Parent)	<u>HLA-B*14:02:01 + HLA-B*35:08:01/HLA-B*35:187</u>	<u>HLA-C*04:01:01:01/HLA-C*04:01:01:02/HLA-C*04:01:01:03/HLA-C*04:01:01:04/HLA-C*04:01:01:05/HLA-C*04:20/HLA-C*04:34</u>	<u>HLA-B*35:08~HLA-C*04:01g</u>	HLA-B*14:02~HLA-C*08:02
(Parent)	<u>HLA-B*13:02:01/HLA-B*13:02:11 + HLA-B*38:27</u>	<u>HLA-C*06:02:01:01/HLA-C*06:02:01:02 + HLA-C*12:03:01:01/HLA-C*12:03:01:02/HLA-C*12:34</u>	<u>HLA-B*13:02g~HLA-C*06:02g</u>	WARNING - No BC haplotype pairs detected.
(Child)	<u>B*13:02:01/HLA-B*13:02:11 + HLA-B*14:02:01</u>	<u>HLA-C*06:02:01:01/HLA-C*06:02:01:02 + HLA-C*08:02:01/HLA-C*08:34</u>	<u>HLA-B*13:02g~HLA-C*06:02g</u>	HLA-B*14:02~HLA-C*08:02

-B and HLA-C genotypes and predicted HLA-B-B~C haplotypes for subjects in three families (family members 37, 38 and 39; family members 133, 134 and 135; and family members 94, 95 and 96) are shown in Table 2. In each family, parental data is presented above the for the child. When known, parental HLA-B~C haplotypes and their constituent alleles are identified via underlining, **bold-face** or *italic bold-face*. The HLAHapV software reported a warning message for Subjects 39, 134 and 95, because no haplotype HLA-B*38:27 was identified by the HLAHapV software.

	Haplotype 1	Haplotype 2	Relative Frequency (CAU)
	HLA-B*52:01g~HLA-C*12:02	HLA-B*50:01~HLA-C*06:02g	1.19×10^{-4}
	HLA-B*50:01~HLA-C*12:02	<i>HLA-B*52:01g~HLA-C*06:02g</i>	7.32×10^{-12}
	HLA-DRB1*03:01~HLA-DRB3*02:02g~HLA-DQB1*02:01g	HLA-DRB1*11:04~HLA-DRB3*02:02g~HLA-DQB1*03:01g	6.92×10^{-4}
	HLA-B*44:03~HLA-C*16:01	<i>HLA-B*57:01g~HLA-C*06:02g</i>	9.02×10^{-4}
	HLA-B*44:03~HLA-C*06:02g	HLA-B*57:01g~HLA-C*16:01	3.60×10^{-10}
DRB4*01:01:01/HLA-DRB4*03:01N	HLA-DRB1*07:01~HLA-DRB4*01:01g~HLA-DQB1*02:01g	<i>HLA-DRB1*13:05~HLA-DRB3*02:02g~HLA-DQB1*03:01g</i>	2.81×10^{-4}
	<i>HLA-B*52:01g~HLA-C*06:02g</i>	<i>HLA-B*57:01g~HLA-C*06:02g</i>	4.11×10^{-7}
HLA-DRB3*02:02:01:01/HLA-DRB3*02:02:01:02	HLA-DRB1*13:05~HLA-DRB3*02:02g~HLA-DQB1*03:01g	HLA-DRB1*11:04~HLA-DRB3*02:02g~HLA-DQB1*03:01g	1.03×10^{-4}

	Haplotype 1	Haplotype 2	Relative Frequency (CAU)
A-DRB4*01:03:01:01/HLA-DRB4*01:03:01:02N/HLA-DRB4*01:03:01:03	HLA-B*52:01g~HLA-C*12:02	HLA-B*50:01~HLA-C*06:02g	1.19×10^{-4}
	HLA-B*50:01~HLA-C*12:02	<i>HLA-B*52:01g~HLA-C*06:02g</i>	7.32×10^{-12}
	HLA-DRB1*07:01~HLA-DRB4*01:01g~HLA-DQB1*02:01g	HLA-DRB1*13:02~HLA-DRB3*03:01~HLA-DQB1*06:09	8.34×10^{-4}
	HLA-B*53:01~HLA-C*04:01g	HLA-B*14:02~HLA-C*08:02	9.86×10^{-5}
/HLA-C*04:20/HLA-C*04:117+HLA-C*08:02:01/HLA-C*08:34	HLA-B*53:01~HLA-C*08:02	HLA-B*14:02~HLA-C*04:01g	4.81×10^{-11}
	HLA-DRB1*01:02~HLA-DRBX*NNNN~HLA-DQB1*05:01	<u>HLA-DRB1*01:02~HLA-DRBX*NNNN~HLA-DQB1*05:01</u>	2.28×10^{-4}
	HLA-B*14:02~HLA-C*08:02	<i>HLA-B*52:01g~HLA-C*06:02g</i>	3.31×10^{-7}
	<u>HLA-DRB1*01:02~HLA-DRBX*NNNN~HLA-DQB1*05:01</u>	HLA-DRB1*13:02~HLA-DRB3*03:01~HLA-DQB1*06:09	1.31×10^{-4}