Human Immunology xxx (2015) xxx–xxx

Contents lists available at ScienceDirect

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

HLA Haplotype Validator for quality assessments of HLA typing

Kazutoyo Osoegawa [a,c,*], Steven J. Mack [c], Julia Udell [c], David A. Noonan [c], Steven Ozanne [b], Elizabeth Trachtenberg [c], Matthew Prestegaard [b]

[a] Department of Pathology, Stanford University, Stanford, CA, USA
[b] National Marrow Donor Program, Minneapolis, MN, USA
[c] Children's Hospital Oakland Research Institute, Oakland, CA, USA

ARTICLE INFO

Article history:
Received 31 May 2015
Revised 29 August 2015
Accepted 29 October 2015
Available online xxxx

Keywords:
Haplotype
Linkage Disequilibrium
HLA-B∼C haplotype block
HLA-DRB3/4/5∼DRB1∼DQB1 haplotype block

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, and 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 demonstrate the use of the HLAHapV software for identifying unusual haplotypes, which can lead to finding potential HLA typing errors.

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 October 2015, 10,297 alleles have been described for HLA class I genes, and 3,543 for class II, totaling 13,840 alleles registered in IMGT/HLA Database version 3.22.0 [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 to be 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

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.

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

http://dx.doi.org/10.1016/j.humimm.2015.10.018
0198-8859/

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 belonging 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 secondary experiments.

Many factors can contribute to 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]. These 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 by the presence of unusual haplotypes, and could be 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 can spend extensive hours reviewing their data by scanning 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, the World Marrow Donor Association (WMDA) working group discussed and suggested detecting HLA typing errors in US, UK, France, Dutch and German registries at the 14th International HLA and Immunogenetics Workshop (IHIW) [29].

In order to identify such errors or unusual haplotypes in a systematic way, we have developed "the HLA Haplotype Validator (HLAHapV) software", which checks each allele against the CWD catalog, 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 in the reference haplotype table, as demonstrated below.

2. Materials and methods

2.1. Development of the 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 the 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 (Fig. 1) [8].

2.3. Common and well-documented alleles (CWD)

The software reviews the CWD 2.0.0 catalog 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

Please cite this article in press as: K. Osoegawa et al., HLA Haplotype Validator for quality assessments of HLA typing, Hum. Immunol. (2015), http://dx.doi.org/10.1016/j.humimm.2015.10.018