Title

Challenges for the standardized reporting of NGS HLA genotyping: Surveying gaps between clinical and research laboratories

Journal

Human Immunology, 82(11)

ISSN

0198-8859

Authors

Osoegawa, Kazutoyo
Montero-Martín, Gonzalo
Mallempati, Kalyan C
et al.

Publication Date

2021-11-01

DOI

10.1016/j.humimm.2021.08.011

Peer reviewed

Published in final edited form as:

Hum Immunol. 2021 November ; 82(11): 820–828. doi:10.1016/j.humimm.2021.08.011.

Challenges for the Standardized Reporting of NGS HLA Genotyping: Surveying Gaps Between Clinical and Research Laboratories

Kazutoyo Osoegawa[1,*], **Gonzalo Montero-Martín**[1], **Kalyan C. Mallempati**[1], **Miranda Bauer**[3], **Robert P. Milius**[4], **Martin Maiers**[4], **Marcelo A. Fernández-Viña**[1,2], **Steven J. Mack**[5]

[1]Histocompatibility & Immunogenetics Laboratory, Stanford Blood Center, Palo Alto, CA 94303 USA

[2]Department of Pathology, Stanford University School of Medicine, Palo Alto, CA 94304 USA

[3]Vendor and Laboratory Services, National Marrow Donor Program, Minneapolis, MN, USA

[4]Bioinformatics Research, National Marrow Donor Program, Minneapolis, MN, USA

[5]Department of Pediatrics, University of California, San Francisco, Oakland, CA, 94609 USA

Abstract

Next generation sequencing (NGS) is being applied for HLA typing in research and clinical settings. NGS HLA typing has made it feasible to sequence exons, introns and untranslated regions simultaneously, with significantly reduced labor and reagent cost per sample, rapid turnaround time, and improved HLA genotype accuracy. NGS technologies bring challenges for cost-effective computation, data processing and exchange of NGS-based HLA data. To address these challenges, guidelines and specifications such as Genotype List (GL) String, Minimum Information for Reporting Immunogenomic NGS Genotyping (MIRING), and Histoimmunogenetics Markup Language (HML) were proposed to streamline and standardize reporting of HLA genotypes. As part of the 17th International HLA and Immunogenetics Workshop (IHIW), we implemented standards and systems for HLA genotype reporting that included GL String, MIRING and HML, and found that misunderstanding or misinterpretations of these standards led to inconsistencies in the reporting of NGS HLA genotyping results. This may be due in part to a historical lack of centralized data reporting standards in the histocompatibility and immunogenetics community. We have worked with software and database developers, clinicians and scientists to address these issues in a collaborative fashion as part of the Data Standard Hackathons (DaSH) for NGS. Here we report several categories of challenges to the consistent exchange of NGS HLA genotyping data we have observed. We hope to address these challenges in future DaSH for NGS efforts.

[*]Correspondence: Kazutoyo Osoegawa, kazutoyo@stanford.edu.

Keywords

allelic ambiguity; genotypic ambiguity; GL String; MIRING; HML

1. Introduction

Many data standards for interpreting and sharing DNA sequences have been defined and applied by the genomic and genetic research community. FASTQ has emerged as a common file format for sharing sequence read data by combining both the nucleotide sequence and an associated per base quality score [1]. The Sequence Alignment/Map (SAM) format is a generic alignment format for storing read alignments in the context of reference sequences [2]. The variant call format (VCF) is a generic format for storing DNA polymorphism data such as SNPs, insertions, deletions and structural variants, alongside rich annotations [3]. All three have been widely adapted for many bioinformatics tools since next generation sequencing (NGS) technology emerged and gained popularity.

The classical Human Leukocyte Antigen (HLA) genes are recognized as the most polymorphic loci in the human genome [4, 5]. They display extensive nucleotide variation and are very difficult to characterize using single nucleotide polymorphisms (SNPs) in a clinically meaningful manner. The World Health Organization (WHO) Nomenclature Committee for factors of the HLA System has established a system that assigns unique allele names based on the constellation of SNPs and more complex multinucleotide polymorphisms within each HLA gene [6]. Historically, core-exon sequences, encoding the antigen recognition domain, were targeted for HLA typing using Sanger sequencing-based typing (SBT) methods, and SBT HLA genotypes were primarily reported using truncated two-field allele names. The implementation of NGS for HLA typing has made it feasible to sequence all exons, along with introns and untranslated regions, and to potentially report untruncated four-field HLA allele names. Currently, three HLA class I *(HLA-A, -B* and *-C),* and eight HLA class II *(HLA-DRB3, -DRB4, -DRB5, -DRB1, -DQA1, -DQB1, -DPA1* and *-DPB1)* genes are routinely genotyped for transplantation therapy and immunogenetic research.

Within the histocompatibility and immunogenetics (H&I) community, specific data standards have been developed for sharing and interpreting HLA genotype data. The goal in defining these standards has been to facilitate uninterrupted data exchanges between NGS HLA typing software and either laboratory information management systems (LIMS) or analytical tools developed for interpreting HLA genotyping data. Genotype List (GL) String has been proposed as a standard format for reporting HLA genotypes [7]. GL String is a grammar that applies a set of hierarchical delimiters (+, ^, /, | and ~), described in Table 1, to precisely define the relationships between alleles, lists of possible alleles, genotypes, lists of possible genotypes, phased alleles and multilocus unphased genotypes for an individual, as a precise representation of a specific genotyping result. The Minimum Information for Reporting Immunogenomic NGS Genotyping (MIRING) guidelines define the minimal set of data and meta-data needed to understand an HLA genotyping result in the context of the NGS system that generated it [8]. Histoimmunogenetics Markup Language (HML) is

an electronic eXtensible Markup Language (XML) format designed for exchange of HLA genotyping data, with extensions developed for next-generation sequencing (NGS) that conform to the MIRING reporting guidelines [9].

Historically, the vendors of HLA genotyping software and databases, along with clinical HLA laboratories, have been inclined to create their own independent data reporting systems. Each clinical HLA laboratory defines a unique data exchange system by working with HLA genotyping system vendors and LIMS vendors. While this may be sufficient for day-to-day clinical operations, it becomes an obstacle when clinical HLA laboratories participate in research collaborations and must exchange data with collaborating laboratories. Problems arise when these H&I laboratories use publicly available software for data interpretation, as investigators must spend significant amounts of time learning how to use the software, and determining how to format their data for the software.

As part of the 17th International HLA and Immunogenetics Workshop (IHIW), analytical tools e.g., HLA Haplotype Validator (HLAHapV) [10], Bridging ImmunoGenomic Data-Analysis Workflow Gaps (BIGDAWG) [11], haplObserve and Phased or Unphased Linkage Disequilibrium (POULD) [12], were developed and updated to operate using GL String, MIRING, and HML formatted data. Despite the requirement to use these data standards for the 17th IHIW, we encountered many instances in which the smooth flow of data from the HLA typing laboratories to the analytic software was not possible. Here we describe informatics challenges experienced by HLA laboratories participating in 17th IHIW research projects in the use of publicly available tools for the exchange of HLA genotyping data.

2. Materials and Methods

17th IHIW NGS HLA genotyping data were generated using five software platforms: Assign TruSight HLA (Illumina), HLA Twin (Omixon), MIA FORA (Immucor), NGSengine (GenDx) and TypeStream Visual (Thermo Fisher Scientific). IPD-IMGT/HLA Database release version 3.25.0 was used for the 17th IHIW. We reprocessed some data reported using IPD-IMGT/HLA Database release version 3.36.0 and 3.42.0 allele names, compared HLA genotyping results under these release versions using hlaGenotypeEvaluator (https://github.com/IHIW/hlaGenotypeEvaluator) [13], and used IPD-IMGT/HLA Database release version 3.44.0 for manual inspection. NGS HLA genotyping data were exported in HML format for review and comparative analyses. The current version of HML meets the MIRING guidelines, which require use of GL String-formatted genotypes. We have also included observations based on reevaluation of clinical genotypings that were performed either using a new version of an NGS HLA genotyping software, while keeping the pertinent IPD-IMGT/HLA Database version constant, or using a new version of the IPD-IMGT/HLA Database reference alignments, while keeping the genotyping software version constant.

There are three major commercially available clinical LIMS for the H&I community: HistoTrac (SystemLink, Inc), mTilda (HLA Data Systems) and Cytopar Histocompatibility Suite (Cytopar LLC). There are also laboratories that employ in-house developers to create homegrown LIMS specific to their institution.

Through this process, we have identified eight areas of focus where community effort and improvement are needed to facilitate better, more effective communication of NGS genotyping results – 1) community consensus between software developers, 2) consistent use of GL String notation, 3) improved reporting of genotyping ambiguity, 4) streamlined human review of genotyping results, 5) improved generation of consensus sequence, 6) improved detection of novel alleles, 7) standardized validation of HML messages, and 8) consistent application of IPD-IMGT/HLA Database versioning.

3. Challenges to Effective Data Exchange

3.1 Lack of Consensus across Software Development Parties

Data standardization is a key factor for successful collaboration between clinical HLA laboratories and research scientists.

In clinical HLA laboratories, NGS HLA genotyping is routinely performed for BMT patients and donors for HLA matching, and for solid organ recipients at pre-transplant stage, and solid organ donors for retrospective monitoring of donor specific antibody (DSA). The HLA laboratories are required to report both recipient and donor HLA genotyping results to a BM registry or donor center. The NMDP is the primary recipient of clinical HLA genotyping results in the United States, and these results are electronically transmitted using HML, which can be formatted in multiple ways. For research applications, HLA genotyping data and associated consensus sequences must be provided to a software application for analysis. These data are usually exchanged as text files for research applications. Effective data standardization requires consensus between NGS genotyping vendors, LIMS developers, and the developers of research software tools. Figure 1 illustrates some of the areas in an NGS workflow where a lack of consensus among these parties results in obstacles to collaboration.

FASTQ sequence files are transmitted from NGS instrument to HLA genotyping software. Data transmission from the HLA genotyping software to LIMS requires significant upfront efforts to meet each clinical laboratory's requirements, with adjustments made for NGS vendor-specific data formats (K. Osoegawa, personal communication). This process is costly and time-consuming, because clinical laboratories, NGS HLA genotyping software vendors, and commercial LIMS vendors and homegrown LIMS developers work independently, in an uncollaborative fashion, and without following publicly available data standards. The various types of NGS HLA genotyping software generate XML formatted HML-like output files. LIMS vendors indicate that these HML-like file formats differ between NGS platforms, requiring the development of vendor-specific XML parsers to extract the required HLA genotypes from the HML-like files for both clinical and research applications (K. Osoegawa, personal communication). Each LIMS vendor has established their own electronic transmission system for transferring HLA genotyping data from their LIMS to NMDP. However, it is currently not possible to generate a local HML output file from the vendor-based LIMS, hindering the extraction of data from LIMS for research applications.

Homegrown LIMS and vendor-based software each have benefits and drawbacks. Laboratories that make homegrown LIMS using an in house developer are able to more

dynamically make changes to their software. They can respond to major updates requested by their customers or other institutions quickly, and are able to expeditiously fix errors generated by their LIMS. It may take significant time for a vendor to roll out an upgrade to all of their customers or make a correction to their software. The decision by vendors to make free or low cost updates to their software can depend on how widespread the adoption will be by their existing customer base. Customization of software that is requested by a single laboratory is not cost-effective for the vendor, so labs may have to pay a premium for a feature tailored to their specific needs compared to a feature that can be adopted by multiple laboratories.

Vendor-based LIMS have the benefit of being prepackaged systems. The software is offered as a standardized base package with optional add-on features. Labs that use vendor based software do not have the ongoing expense of staffing an in-house developer, and can instead call on their software representative to handle the installation and maintenance process of the LIMS. Vendor based software generally uses the same formatting rules for reporting data for all their customers compared to homegrown programs, which lends itself better to standardization of reporting. Labs who use homegrown software may each have a unique way of collecting and reporting data, which can make it difficult to analyze by outside programs or for research purposes.

3.2 Inconsistency in Applying Genotype List (GL) String format

Table 1 defines the GL String delimiters [7], and presents examples of their application. Table 2 details three examples of GL String-related errors. In error 1, both the full-length *HLA-DPB1*04:01:01:01* allele and the truncated two-field *HLA-DPB1*04:01* allele are included in the same GL String. When a single full-length allele name, like *HLA-DPB1*04:01:01:01*, is reported, it indicates that this is the only possible allele. In contrast, when a truncated two-field allele name like, *HLA-DPB1*04:01*, is reported, it includes all third- and fourth-field allele names that begin with *HLA-DPB1*04:01*; there are 122 such possible *HLA-DPB1*04:01* alleles in IPD-IMGT/HLA Database release version 3.44.0. Using GL String notation, these can be represented as a slash-delimited ambiguous allele string (Supplementary Table 1).

In error 2, three possible *HLA-DRB1*15:01* alleles have been delimited with pipes (|) instead of slashes (/). We speculate that error 2 results from a misunderstanding of when to use a pipe (|) and a slash (/) (Table 1). It is sometimes impossible to establish phase between detected polymorphisms using short sequence reads, especially in the presence of an extended SNP desert. Genotypic ambiguity is reported when two or more possible genotypes are observed. GL String formatted genotypic ambiguity is represented using pipe (|) and plus (+) delimiters to identify all possible genotypes that cannot be distinguished. In contrast, GL String formatted allelic ambiguity is represented using a slash (/); e.g., *HLA-DRB1*15:01:01:01/HLA-DRB1*15:01:01:02/HLA-DRB1*15:01:01:03* indicates that these three alleles are not distinguishable using the HLA genotyping method applied [7].

In error 3, the genotype *HLA-DPB1*04:01:01:01|HLA-DPB1*126:01+HLA-DPB1*04:02:01:02|HLA-DPB1*105:01* genotype is delimited with slashes (/) instead of pipes (|). The *HLA-DPB1*04:01:01:01* and *HLA-DPB1*126:01* alleles share identical exon

2 sequences, as do the *HLA-DPB1*04:02:01:01* and allele *HLA-DPB1*105:01* alleles. The *HLA-DPB1*04:01:01:01* and *HLA-DPB1*105:01* alleles share identical exon 3 sequences, as do the *HLA-DPB1*04:02:01:01* and *HLA-DPB1*126:01* alleles. Based on these exon 2 and 3 sequences, this genotype should have been reported as a genotypic ambiguity: *HLA-DPB1*04:01:01:01+HLA-DPB1*04:02:01:02|HLA-DPB1*126:01+HLA-DPB1*105:01.*

3.3 Programmatic Failure to Report Genotypic Ambiguities

As part of the 17th IHIW family haplotype project data, we identified a family in which the mother's *HLA-B* genotype was *HLA-B*51:01:01:01+HLA-B*53:01:01,* the father's was *HLA-B*35:08:01+HLA-B*14:02:01:01,* and the first child's was *HLA-B*14:02:01:01+HLA-B*53:01:01* (Table 3). The second child's *HLA-B* genotype was reported as *HLA-B*35:01:01:02+HLA-B*53:24,* which did not match the parental *HLA-B* alleles. As described in section 3.2, genotypic (aka, phase) ambiguities occur in the presence of an extended SNP desert, a phenomenon that is frequently encountered for *HLA-DPB1,* but can occur at other HLA loci [14]. DNA sequence alignment suggested that the NGS HLA genotyping software (1) failed to phase informative SNPs during the sequence assembly stage, (2) reported consensus sequence for only one of two possible genotypes, and (3) did not report a genotype ambiguity: *HLA-B*35:08:01+HLA-B*53:01:01|HLA-B*35:01:01:02+HLA-B*53:24* (Figure 2). There is currently (as of IPD-IMGTIHLA Database release version 3.45.0) no genomic reference sequence for *HLA-B*53:24,* and there are no informative SNPs in a 465 nucleotide-long region spanning the 3' region of exon 2, intron 2 and the 5' region of exon 3 of *HLA-B*35:08:01*, *HLA-B*53:01:01* and *HLA-B*35:01:01:02* (Figure 2). We were able to detect this HLA genotype reporting error because we had HLA genotypes for all family members. Without these family data, this error would likely have gone unidentified. This example highlights the importance that vendors ensure that genotyping software accurately report HLA genotype ambiguity, especially in instances when sequence phase is unknown. MIRING provides guidelines for accurately describing consensus sequences with known and unknown phase relationships. We speculate that unexpected, unphased sequences may have occured in this case if the fragment size of the DNA sequencing library was smaller than optimal (e.g., < 450 bp).

4. Required Human Review of Genotyping Results

4.1 Importance of Manual Review of HLA Alleles and Haplotypes

One of the shortcomings of PCR-based enrichment procedures is the potential for 'allele dropout' due to amplification failure. It is crucial to review each software-generated HLA genotype to detect potential allele dropout. It may be feasible to detect allele dropout by testing the same sample using a different method, e.g. sequence-specific oligonucleotide probe (SSOP). However, it is costly and time-consuming to perform confirmatory experiments for all subjects, especially in research or other high throughput settings when HLA genotypes are generated for hundreds or thousands of subjects. A reasonably cost-effective procedure to detect allele dropout is to review common HLA haplotypes that have been characterized and published for various ethnic groups or countries [15–18]. There are also computational tools to automatically predict haplotypes [10–12, 19]. As part of the 17th IHIW family haplotype project, we

encountered a subject with an *HLA-DRB1*07:01:01:01/HLA-DRB1*07:01:01:02~HLA-DQA1*02:01:01:01/HLA-DQA1*02:01:01:02~ HLA-DQB1*02:02:01:01* haplotype, but we could not detect an *HLA-DRB4* allele expected based on the common haplotype analysis using NGS. Two siblings and a parent in this family carried the same *DR~DQ* haplotype, but we did not detect the expected *HLA-DRB4* allele using NGS HLA typing for them either. We performed SSOP genotyping for these individuals, and were able to confirm the presence of the *HLA-DRB4*01:01:01:01* allele. We hypothesized that there could be an unknown sequence variant located near the 3'-end of an NGS PCR primer, that lead to the initial PCR amplification failure of *HLA-DRB4* sequences. This exemplifies a technical limitation of amplicon-based NGS HLA typing assays, as well as the importance of reviewing HLA haplotypes and following up inconsistencies using a different method [20].

In addition, it is important to be aware of the presence of unusual haplotypes. NGS HLA typing systems are capable of capturing such haplotypes. For example, we identified a subject with a *HLA-DRB4*01:03:01:05+HLA-DRB5*01:01:01:01^HLA-DRB1*01:01:01:01+HLA-DRB1*04:05:01:04^HLA-DQA1*01:01:01:01+HLA-DQA1*03:03:01:03^HLA-DQB1*04:01:01:01+HLA-DQB1*05:01:01:03* genotype. The imputed haplotypes were *HLA-DRB4*01:03:01:05~HLA-DRB1*04:05:01:04~HLA-DQA1*03:03:01:03~HLA-DQB1*04:01:01:01+*<u>*HLA-DRB5*01:01:01:01~HLA-DRB1*01:01:01:01*</u>*~HLA-DQA1*01:01:01:01~HLA-DQB1*05:01:01:03*. The second *DRB* haplotype (<u>*HLA-DRB5*01:01:01:01~HLA-DRB1*01:01:01:01*</u>) does not conform to the broad structural *DRB* haplotypes described by Andersson [21]. We confirmed the presence of *HLA-DRB5*01:01:01:01* allele by visual inspection of the sequence alignments.

4.2 Detecting Errors of Consensus Sequence Assembly

Current NGS HLA typing systems examine available HLA gene sequences, including introns. Erroneous DNA sequence assembly from FASTQ files often leads to an inaccurate HLA genotype. Here, we present three cases, illustrated in Figure 3, where errors in assembly resulted in consensus sequences that incorrectly incorporated SNPs.

In case 1, we identified a *HLA-DPB1* genotype, *HLA-DPB1*05:01:01:01+HLA-DPB1*135:01*, using IPD-IMGT/HLA Database release version 3.36.0. After we re-processed the same FASTQ files using IPD-IMGT/HLA Database release version 3.42.0, using the same NGS genotyping software version, the *HLA-DPB1* genotype was reported as *HLA-DPB1*03:01:01:01+HLA-DPB1*135:01.* In the first genotype, *HLA-DPB1*05:01:01:01+HLA-DPB1*135:01*, DNA sequences corresponding to exon 2, intron 2 and exon 3 for *HLA-DPB1*03:01:01:01* or *HLA-DPB1*104:01:01:01:01* had been completely ignored. In the second genotype, *HLA-DPB1*03:01:01:01+HLA-DPB1*135:01*, the rs11551421 SNP "A" variant in exon 4 had not been included in consensus sequence for two possible alleles by the NGS genotyping software, but had been included in only a single consensus sequence, resulting in incorrect HLA genotypes (Figure 3A). After careful review of the sequence alignments, we concluded that both genotyping results were incorrect, and that the genotype should have been reported as *HLA-DPB1*03:01:01:01+HLA-DPB1*05:01:01:01|HLA-DPB1*104:01:01:01:01+HLA-DPB1*135:01.*

In case 2, we identified two incorrect genotypes, *HLA-DPB1*03:01:01:10+HLA-DPB1*104:01:01:01* and *HLA-DPB1*03:01:01:10+HLA-DPB1*124:01:01:01,* in which an *HLA-DPB1*03:01:01:01* allele was incorrectly reported as *HLA-DPB1*03:01:01:10.* The genotypes in these cases should have been reported as *HLA-DPB1*03:01:01:01+HLA-DPB1*104:01:01:01* and *HLA-DPB1*03:01:01:01+HLA-DPB1*124:01:01:01.* The erroneous *HLA-DPB1*03:01:01:10* allele was reported because the rs112104961 SNP "G" variant in intron 2 for the *HLA-DPB1*104:01:01:01* and *HLA-DPB1*124:01:01:01* alleles was erroneously incorporated into the consensus sequence of the *HLA-DPB1*03:01:01:01* allele (Figure 3B). The nearly equal number of sequence reads containing rs112104961 SNP "G" and "T" variants were clearly observed in the sequence alignment view, but the rs112104961 SNP "T" variant was not used for the consensus sequence assembly.

In case 3, we identified a subject with the *HLA-DQB1*03:01:01:07+HLA-DQB1*03:01:01:12* genotype. This case was originally identified in the 17th IHIW Family Haplotype Project [18]. We reprocessed FASTQ files from these family members using a more recent version of the HLA genotyping software and IPD-IMGT/HLA Database release version 3.35.0. We extensively reviewed how HLA allele combinations affect consensus DNA sequence assembly. The rs41263783 SNP in intron 2 distinguishes these alleles (Figure 3C). The NGS HLA genotyping software reported the correct *HLA-DQB1* genotype, but reported only a single consensus sequence representing the *HLA-DQB1*03:01:01:07* allele. These examples reveal the complexity of assembling highly polymorphic HLA genes.

We speculate that these errors may occur because genotyping software developers have primarily focused on returning a genotype result; genotyping algorithms may not be optimized for possible alternative sequence combinations, and may be less focused on returning accurate consensus sequences. HLA genotyping error can be manually corrected using a software function, but the corresponding consensus sequences are not updated. Identifying novel alleles via manual interpretation is very labor-intensive, as discussed below (Section 4.3). Automation of this process may be more cost-effective and efficient for clinical HLA laboratories, but this automation will only be possible if accurate consensus sequences are available. Without demand from clinical HLA typing laboratories and/or regulatory agencies (e.g., APHIA, ASHI and EFI) for accurate consensus sequences that reflect the genotyping result, there may not be an incentive for vendors to address this issue.

4.3 Evaluating the Biological Significance of Novel Allele Sequences

In routine clinical NGS HLA genotyping, we often encounter HLA nucleotide sequences that are not included in the release version of the IPD-IMGT/HLA Database being used by the NGS HLA genotyping software (novel sequence variants). It is clinically important to determine if a novel sequence variant conveys any biological consequences. In some cases, we can identify the corresponding HLA allele name for a novel sequence variant by reviewing the most recent IPD-IMGT/HLA Database release. We identified a subject for which NGS HLA genotyping software called the *HLA-DQB1*05:01:01:03* allele, but also reported a single nucleotide mismatch (T) at SNP rs9273650 in *HLA-DQB1* exon 4 using IPD-IMGT/HLA Database release version 3.36.0. The correct allele, *HLA-DQB1*05:01:35,*

appeared in IPD-IMGT/HLA Database release version 3.37.0. This SNP variant results in a synonymous change, and to our knowledge, clinical significance has not been reported.

Unlike the previous case, we often encounter novel alleles that have not been reported even in the most updated version of the IPD-IMGT/HLA Database. Reporting novel alleles via manual interpretation is very labor-intensive. To facilitate identifying and reporting novel alleles in an automated fashion, we developed hlaPoly, an R software package [22]. In addition, we recently revised a collection of standard reference alleles that can be used to report novel alleles [23]. It is important to note that even if a nonsynonymous change is identified, it is often difficult to determine if that nonsynonymous change has any significant impact in clinical outcomes. For example, a nonsynonymous change (rs11551421 SNP) in exon4 distinguishes *HLA-DPB1*03:01:01:01* and *HLA-DPB1*104:01:01:01* (Figure 3A). This change was reported to have a limited functional role in allorecognition of *HLA-DPB1*03:01*/*HLA-DPB1*104:01* in unrelated stem cell donor selection [24], but little is known about changes in the downstream immunological response [25, 26]. It is also important to note that while current NGS methods together with their related HLA genotyping software are able to detect coding (exon) variants with a relatively high accuracy, detection and characterization of non-coding variants as well as new alleles are still major challenges using the currently available tools [27].

5. Proper Use of Histoimmunogenetics Markup Language

We have also observed multiple non-HLA character strings (e.g., “NO CALL”, “N/A”, “Insufficient data”, etc.) reported in the GL String field in HML documents. When genotyping for a locus has failed, no value should be reported in the “<glstring>“ tag in HML; these allele-calling failures should be reported outside of the GL String field. For genotype dropout information, we recommend adding property tags under <allele-assignment> (Figure 4A), under <typing-method> or even under one of the sequencing methods like <sbt-ngs> (Figure 4B). Property tags are name/value pairs that are coordinated between the sender and receiver and should represent a well-defined value-set.

6. IPD-IMGT/HLA Database Version Consistency and Informatics Challenges

Each quarterly IPD-IMGT/HLA Database release includes new sequences and allele names, and can include minor changes to extant sequences and allele names as well. It is important that genotype calls made under a given IPD-IMGT/HLA Database release should only be made using sequences and allele names present in that version. The 17th IHIW data was collected using IPD-IMGT/HLA Database release version 3.25.0. Of the 14,815 alleles of the 11 classical HLA loci in release version 3.25.0, DNA sequences for 1584 HLA alleles were extended from partial coding sequence or cDNA sequence to genomic DNA sequences, and 357 genomic sequences were updated (mostly extended). 13562 alleles of the 11 classical HLA loci were added to the database between releases 3.25.0 and 3.42.0. Although available genomic DNA sequences in IPD-IMGT/HLA Database have increased, the presence of partial DNA sequences may still introduce informatics challenges for accurate HLA genotype assignments [28]. In addition, the constant increase of HLA

alleles with every IPD-IMGT/HLA Database release requires more and more processing time for some of the currently available HLA genotyping software, thus possibly affecting to the turnaround time for clinical NGS HLA reporting.

When a new NGS HLA genotyping software version is released from a vendor, the software has to be validated prior to its use for clinical tests. Though a vendor may introduce two variables (e.g., new software along with a new IPD-IMGT/HLA Database release version) at the same time for improved results, it is common practice for laboratories to validate only one variable at a time. As part of NGS HLA genotyping software validation, we reprocessed FASTQ files generated for the 17th IHIW QC project using IPD-IMGT/HLA Database release version 3.25.0 (the fixed factor), with a new version of the HLA genotyping software (the variable being validated), and compared the results with those from the 17th IHIW using using hlaGenotypeEvaluator [13]. We observed a discordant genotype, *HLA-DQB1*03:01:01:01/HLA-DQB1*03:276N+HLA-DQB1*03:01:01:01/HLA-DQB1*03:276N,* that is not a possible genotype using IPD-IMGT/HLA Database release version 3.25.0, because the *HLA-DQB1*03:276N* allele appeared in IPD-IMGT/HLA Database release version 3.32.0 [29]. We can only explain this discordant result by reasoning that the *HLA-DQB1*03:276N* allele had been hard-coded in the software to be reported as ambiguous with the *HLA-DQB1*03:01:01:01* allele, even though the allele *HLA-DQB1*03:276N* did not exist in the IPD-IMGT/HLA Database release version 3.25.0. HLA genotyping software developers need to ensure that the genotype calls made under a given database version are only made using sequences and allele names present in that version.

7. Conclusions

The Data Standard Hackathons for NGS have been central in discussing challenges and issues for data standards with representatives of HLA laboratory directors, academic and non-academic scientists and software engineers. The group has been efficiently identifying many issues described in this manuscript, and developing tools to capture and address these issues. However, new technologies are arising rapidly, challenging the H&I community to cope with the speed of innovation, and determine how to best incorporate these innovations into cutting-edge research design and day-to-day clinical tests, all under strict regulations.

Data standards will become increasingly important as the H&I community adopts more contemporary informatics approaches (e.g., moving from manual data entry and formatting to automated data transmission), and as the broader genomic and healthcare communities look to H&I for new research and clinical solutions. The integration of the guidelines and specifications developed for the H&I field into technical standards that have already been embraced by the larger healthcare community (e.g., Global Alliance for Genomics and Health [30] and Health Level Seven International Fast Healthcare Interoperability Resources [31]) will be key for the integration of sequence-based HLA genotyping reports into clinical systems. Collaboration across the H&I community – involving clinicians, HLA laboratory directors, research scientists and software engineers for both genotyping and LIMS systems – will be critically important for the development of technical standards that will make this broader vision possible. Ultimately, an international organization that defines data reporting

standards for the overall H&I community, including both clinical and research laboratories, is needed. The first steps to establishing such an entity could be taken by regional regulatory organizations (e.g., APHIA, ASHI and EFI), by facilitating collaborative discussions around data standards, with the goal of establishing an international standard.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgements

This work was supported by National Institutes of Health (NIH) National Institute of Allergy and Infectious Disease (NIAID) grant R01AI128775 (BM, MM, SM). The content is solely the responsibility of the authors and does not necessarily reflect the official views of the NIAID, NIH, or United States government. We thank the Stanford Blood Center for the support and promotion of the 17th IHIW endeavor, and the Data Standard Hackathons (https://github.com/nmdp-bioinformatics/dash/wiki) for NGS for the community development of tools, services and standards for Histocompatibility and Immunogenetics efforts.

Abbreviations:

ASHI	American Society for Histocompatibility and Immunogenetics
APHIA	Asia-Pacific Histocompatibility and Immunogenetics Association
BIGDAWG	Bridging ImmunoGenomic Data-Analysis Workflow Gaps
BMT	Bone Marrow Transplantation
CSV	comma-separated value
DaSH	Data Standards Hackathons
DSA	Donor Specific Antibody
EFI	European Federation for Immunogenetics
GL	Genotype List
H&I	histocompatibility and immunogenetics
HLA	Human Leukocyte Antigen
HLAHapV	HLA Haplotype Validator
HML	Histoimmunogenetics Markup Language
LIMS	laboratory information management system
MIRING	Minimum Information for Reporting Immunogenomic NGS Genotyping
NGS	next generation sequencing
NMDP	National Marrow Donor Program

POULD	Phased or Unphased Linkage Disequilibrium
SAM	Sequence Alignment/Map
SBT	Sanger sequencing-based Typing
SSOP	sequence specific oligonucleotide probe
TSV	tab-separated value
VCF	Variant Call Format
WHO	World Health Organization
XML	eXtensible Markup Language

References:

[1]. Cock PJ, Fields CJ, Goto N, Heuer ML, Rice PM: The Sanger FASTQ file format for sequences with quality scores, and the Solexa/Illumina FASTQ variants. Nucleic Acids Res 2010;38:1767. [PubMed: 20015970]

[2]. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer Net al. : The Sequence Alignment/Map format and SAMtools. Bioinformatics 2009; 25:2078. [PubMed: 19505943]

[3]. Danecek P, Auton A, Abecasis G, Albers CA, Banks E, DePristo MAet al. : The variant call format and VCFtools. Bioinformatics 2011; 27:2156. [PubMed: 21653522]

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

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

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

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

[8]. Mack SJ, Milius RP, Gifford BD, Sauter J, Hofmann J, Osoegawa Ket al. : Minimum information for reporting next generation sequence genotyping (MIRING): Guidelines for reporting HLA and KIR genotyping via next generation sequencing. Hum Immunol 2015;76:954. [PubMed: 26407912]

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

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

[11]. Pappas DJ, Marin W, Hollenbach JA, Mack SJ: Bridging ImmunoGenomic Data Analysis Workflow Gaps (BIGDAWG): An integrated case-control analysis pipeline. Hum Immunol 2016;77:283. [PubMed: 26708359]

[12]. Osoegawa K, Mack SJ, Prestegaard M, Fernandez-Vina MA: Tools for building, analyzing and evaluating HLA haplotypes from families. Hum Immunol 2019; 80:633. [PubMed: 30735756]

[13]. Osoegawa K, Vayntrub TA, Wenda S, De Santis D, Barsakis K, Ivanova Met al. : Quality control project of NGS HLA genotyping for the 17th International HLA and Immunogenetics Workshop. Hum Immunol 2019;80:228. [PubMed: 30738112]

[14]. Pappas DJ, Lizee A, Paunic V, Beutner KR, Motyer A, Vukcevic Det al. : Significant variation between SNP-based HLA imputations in diverse populations: the last mile is the hardest. Pharmacogenomics J 2018;18:367. [PubMed: 28440342]

[15]. 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]

[16]. Creary LE, Gangavarapu S, Mallempati KC, Montero-Martin G, Caillier SJ, Santaniello Aet al. : Next-generation sequencing reveals new information about HLA allele and haplotype diversity in a large European American population. Hum Immunol 2019;80:807. [PubMed: 31345698]

[17]. Gonzalez-Galarza FF, McCabe A, Santos E, Jones J, Takeshita L, Ortega-Rivera NDet al. : Allele frequency net database (AFND) 2020 update: gold-standard data classification, open access genotype data and new query tools. Nucleic Acids Res 2020;48:D783. [PubMed: 31722398]

[18]. Osoegawa K, Mallempati KC, Gangavarapu S, Oki A, Gendzekhadze K, Marino SRet al. : HLA alleles and haplotypes observed in 263 US families. Hum Immunol 2019;80:644. [PubMed: 31256909]

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

[20]. Kong D, Lee N, Dela Cruz ID, Dames C, Maruthamuthu S, Golden Tet al. : Concurrent typing of over 4000 samples by long-range PCR amplicon-based NGS and rSSO revealed the need to verify NGS typing for HLA allelic dropouts. Hum Immunol 2021.

[21]. Andersson G: Evolution of the human HLA-DR region. Front Biosci 1998;3:d739. [PubMed: 9675159]

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

[23]. Matern BM, Mack SJ, Osoegawa K, Maiers M, Niemann M, Robinson Jet al. : Standard reference sequences for submission of HLA genotyping for the 18th International HLA and Immunogenetics Workshop. HLA 2021;97:512. [PubMed: 33719220]

[24]. Crivello P, Lauterbach N, Zito L, Sizzano F, Toffalori C, Marcon Jet al. : Effects of transmembrane region variability on cell surface expression and allorecognition of HLA-DP3. Hum Immunol 2013;74:970. [PubMed: 23619468]

[25]. Dixon AM, Drake L, Hughes KT, Sargent E, Hunt D, Harton JAet al. : Differential transmembrane domain GXXXG motif pairing impacts major histocompatibility complex (MHC) class II structure. J Biol Chem 2014;289:11695. [PubMed: 24619409]

[26]. Harton J, Jin L, Hahn A, Drake J: Immunological Functions of the Membrane Proximal Region of MHC Class II Molecules. F1000Res 2016;5.

[27]. Klasberg S, Surendranath V, Lange V, Schofl G: Bioinformatics Strategies, Challenges, and Opportunities for Next Generation Sequencing-Based HLA Genotyping. Transfus Med Hemother 2019;46:312. [PubMed: 31832057]

[28]. Sverchkova A, Anzar I, Stratford R, Clancy T: Improved HLA typing of Class I and Class II alleles from next-generation sequencing data. HLA 2019;94:504. [PubMed: 31496113]

[29]. Steiner NK, Hou L, Hurley CK: Characterizing alleles with large deletions using region specific extraction. Hum Immunol 2018;79:491. [PubMed: 29580811]

[30]. Rahimzadeh V, Dyke SO, Knoppers BM: An International Framework for Data Sharing: Moving Forward with the Global Alliance for Genomics and Health. Biopreserv Biobank 2016;14:256. [PubMed: 27082668]

[31]. Strasberg HR, Rhodes B, Del Fiol G, Jenders RA, Haug PJ, Kawamoto K: Contemporary clinical decision support standards using Health Level Seven International Fast Healthcare Interoperability Resources. J Am Med Inform Assoc 2021.

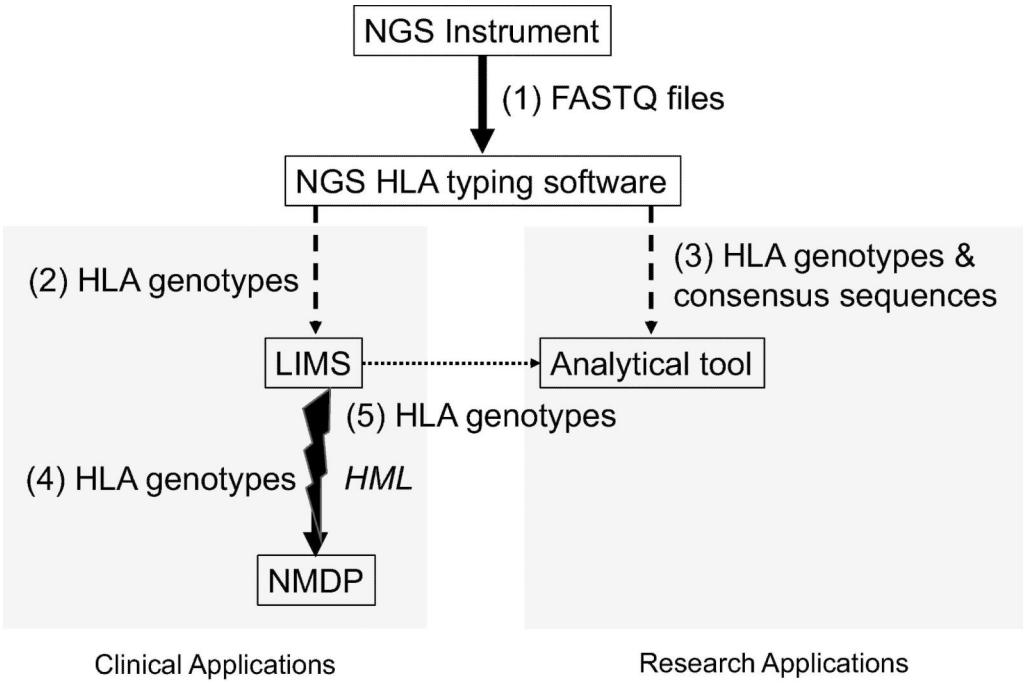

Figure 1:

NGS Data Workflows for Clinical and Research Applications

A generalized NGS genotyping data flow is depicted. The DNA sequences are generated at the top level. Solid bold lines indicate well-established standard workflows, with the bold lightning bolt indicating purely electronic data transmission. Dashed lines indicate laboratory specific workflows. The dotted line indicates that it is currently difficult or impossible to extract data from the LIMS for clinical research in analytical tools. (1) FASTQ files containing DNA sequences are imported into NGS HLA genotyping software. (2) NGS genotyping software generates reports in different formats, e.g. CSV, TSV, vendor specific XML or HML. HLA laboratories and LIMS vendors individually define which reporting file format is used to import HLA genotyping data into LIMS. There is no standard in this step; it is costly to develop a customized system for each HLA laboratory. (3) HLA genotypes can be extracted from NGS HLA genotyping software for H&I research, but this currently requires efforts to adjust the file format compatibility with the analytical tools. H&I vendors and research software developers have been working to standardize this workflow via DaSH. (4) LIMS vendors successfully established a pipeline for standardized electronic data transmission from the clinical database. Unrelated donor and recipient HLA genotypes are electronically transmitted using HML to NMDP. (5) However, HML files cannot currently be generated or transmitted from LIMS to local computers, and this is a major obstacle for collaborations between clinical and research laboratories. Increased participation by LIMS developers in future DaSH events may help to address this shortcoming.

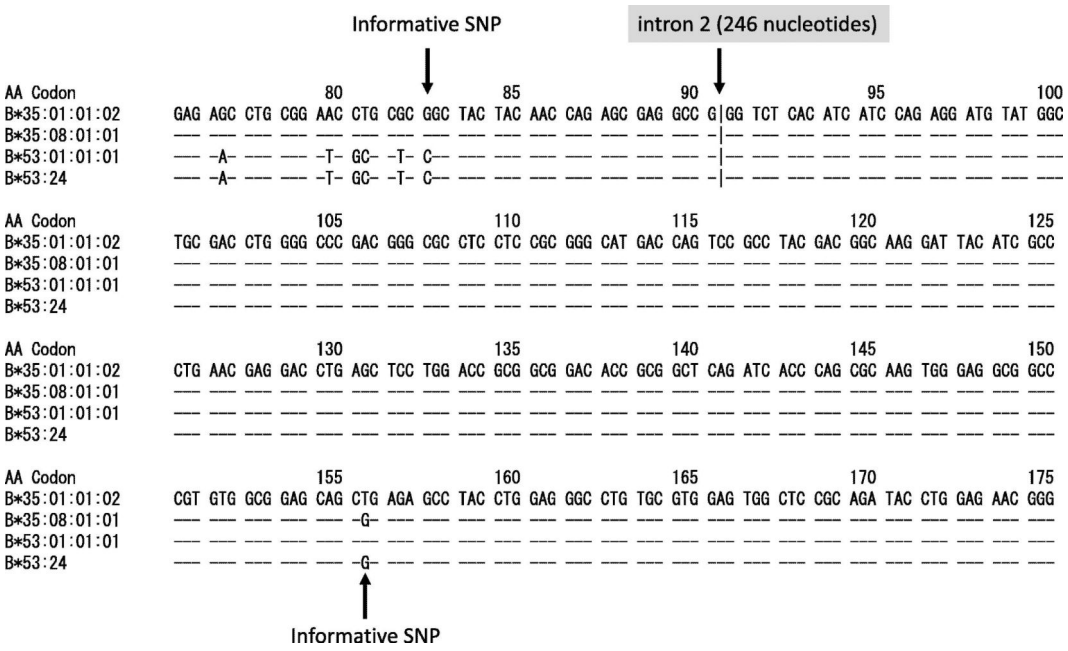

Figure 2:

Figure shows the exon 2 and exon 3 nucleotide sequence alignment of the *HLA-B*35:08:01:01*, *HLA-B*53:01:01:01, HLA-B*35:01:01:02,* and *HLA-B*53:24* alleles. The exon 2 and 3 boundary is between codon 91 positions 1 and 2, and intron 2 position and size are indicated with gray highlight. Two informative SNPs are also shown. A genotype ambiguity, *HLA-B*35:08:01+HLA-B*53:01:01|HLA-B*35:01:01:02+HLA-B*53:24*, could be reported when the NGS HLA genotyping software fails to phase SNPs between these exons.

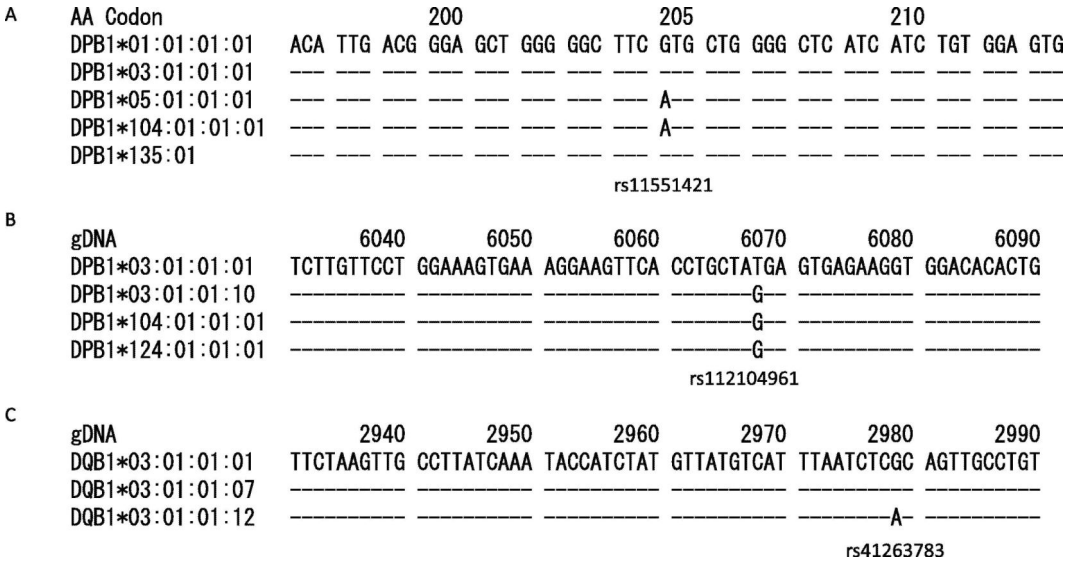

Figure 3:

DNA sequence alignments

Figure 3A shows the nucleotide sequence alignment of partial exon 4 sequences of the *HLA-DPB1*03:01:01:01, HLA-DPB1*05:01:01:01, HLA-DPB1*104:01:01:01:01* and *HLA-DPB1*135:01* alleles. Figure 3B shows the nucleotide sequence alignment of partial intron 2 sequences of *HLA-DPB1*03:01:01:01, HLA-DPB1*03:01:01:10*, *HLA-DPB1*104:01:01:01* and *HLA-DPB1*124:01:01:01* alleles. Figure 3C shows DNA sequence alignment of partial intron 2 sequences of alleles *HLA-DQB1*03:01:01:01*, *HLA-DQB1*03:01:01:07* and *HLA-DQB1*03:01:01:12.* SNPs rs11551421, rs112104961 and rs41263783 are shown in these figures. Failure of separating these SNPs as two distinct consensus sequences resulted in assigning incorrect HLA genotype assignments.

A

```
<allele-assignment date="2021-02-01">
    <property name="dropout" value="HLA-A"/>
    <glstringHLA-B*40:01:02:01+HLA-B*07:02:01:01</glstring>
    <glstring>HLA-C*03:04:01:01+HLA-C*07:02:01:03</glstring>
</allele-assignment>
```

B

```
<typing-method>
  <sbt-ngs locus="HLA-A" />
    <property name="Insufficient data" value="dropout"/>
</typing-method>
```

Figure 4:

HML Property Tags

This figure includes two examples that illustrate how allele dropout can be reported using the HML <property> tag. Property tags contain name/value pairs. In Figure 4A, a property tag was added under <allele-assignment>. In Figure 4B, a property tag was added under the <sbt-ngs> typing method.

Table 1.

Genotype List String Delimiters and their Usage

A: Delimiters				
Delimiter	Name	Usage	Example	Note
+	Plus	Gene copy	HLA-A*24:02:01:01+HLA-A*02:06:01:01	Two distinct HLA-A alleles are identified as present using “+”.
^	Caret	Gene separator	HLA-B*35:01:01:02+HLA-B*51:01:01:01^HLA-C*03:03:01:01+HLA-C*15:02:01:01	HLA-B and HLA-C genotypes are separated by “^”.
/	Forward-Slash	Allele ambiguity	HLA-DQB1*05:03:01:01/HLA-DQB1*05:03:01:02+HLA-DQB1*03:01:01:01^HLA-DRB1*14:04:01+HLA-DRB1*04:08:01	Two indistinguishable HLA-DQB1 alleles are represented using “/”.
|	Pipe	Genotype ambiguity	HLA-DPB1*04:02:01:02+HLA-DPB1*04:01:01:01|HLA-DPB1*105:01+HLA-DPB1*126:01	Two possible HLA-DPB1 genotypes are represented using “|”.
~	Tilde	Gene phase	HLA-A*02:06:01:01~HLA-C*03:03:01:01~HLA-B*35:01:01:02	HLA-A, HLA-C and HLA-B alleles are experimentally or analytically confirmed on the same chromosome using “~”.
B: Extended Genotypes and Haplotypes represented by GL String				
	Delimiter	GL String		
Combined genotype	+,^, /, |	HLA-A*24:02:01:01+HLA-A*02:06:01:01^HLA-B*35:01:01:02+HLA-B*51:01:01:01^HLA-C*03:03:01:01+HLA-C*15:02:01:01^HLA-DPB1*04:02:01:02+HLA-DPB1*04:01:01:01|HLA-DPB1*105:01+HLA-DPB1*126:01^HLA-DQB1*05:03:01:01/HLA-DQB1*05:03:01:02+HLA-DQB1*03:01:01:01^HLA-DRB1*14:04:01+HLA-DRB1*04:08:01		

Combined Observed Haplotypes	~, +	HLA-A*02:06:01:01~HLA-C*03:03:01:01 ~HLA-B*35:01:01:02~HLA-DRB4*01:03:01:01/HLA-DRB4*01:03:01:03~HLA-DRB1*04:08:01~HLA-DQA1*03:03:01:01~HLA-DQB1*03:01:01:01~HLA-DPA1*01:03:01:02~HLA-DPB1*04:01:01:01/HLA-DPB1*04:01:01:02+HLA-A*24:02:01:01~HLA-C*15:02:01:01~HLA-B*51:01:01:01~HLA-DRB3*02:02:01:01~HLA-DRB1*14:04:01~HLA-DQA1*01:04:02~HLA-DQB1*05:03:01:01/HLA-DQB1*05:03:01:02~HLA-DPA1*01:03:01:05~HLA-DPB1*04:02:01:02

A: Table shows GL String delimiters [7]. Care should be taken to ensure that each delimiter is used in the proper context. For example, the pipe symbol should **never** be used to delimit ambiguous alleles at a locus; each pipe symbol should always be accompanied by at least two plus symbols. Ambiguous alleles at a locus should **always** be delimited using the forward-slash symbol. When gene phase is observed/confirmed by HLA allele segregation analyses within a family or MHC region sequencing, the tilde sign is used to represent gene phase or haplotype, but should not be used to represent predicted phase based on known haplotypes.

B: Together, the examples from Table A are combined in a single Genotype List String (Top). Two observed haplotypes are represented using tildes (~) and connected with plus (+) signs from family segregation analyses generated as part of the 17th IHIW family haplotype project [18]. The genotypes from the other family members are omitted.

Table 2:

Improperly formatted GL String

Error	Improperly formatted GL String	Possible Intended GL String	Comment
1	*HLA-DPB1*02:01:02+HLA-DPB1*04:01:01:01|HLA-DPB1*02:01:02+***HLA-DPB1*04:01***	*HLA-DPB1*02:01:02+HLA-DPB1*04:01:01:01*	Truncated two-field allele
2	*HLA-DRB1*15:01:01:01|HLA-DRB1*15:01:01:02|HLA-DRB1*15:01:01:03*	*HLA-DRB1*15:01:01:01/ HLA-DRB1*15:01:01:02/HLA-DRB1*15:01:01:03*	Incorrect usage of the pipe (|) delimiter
3	*HLA-DPB1*04:01:01:01| HLA-DPB1*126:01+HLA-DPB1*04:02:01:02| HLA-DPB1 *105:01*	*HLA-DPB1*04:01:01:01+HLA-DPB1*04:02:01:02|HLA-DPB1*126:01+HLA-DPB1*105:01*	Incorrect allelic ambiguities due to incorrect usage of the slash (/) delimiter

Table shows improperly formatted GL Strings, and the most likely intended genotypes. The erroneous elements are shown in bold in the leftmost column.

Table 3:

Reporting error of genotypic ambiguity

Relationship	***HLA-B* genotype**
Father	*HLA-B*35:08:01+HLA-B*14:02:01:01*
Mother	*HLA-B*51:01:01:01+HLA-B*53:01:01*
Child A	*HLA-B*14:02:01:01 +HLA-B*53:01:01*
Child B	*HLA-B*35:08:01+HLA-B*53:01:01*|***HLA-B*35:01:01:02+HLA-B*53:24***

Table shows *HLA-B* genotypes from a quartet family. Paternal alíeles are underlined, and maternal alleles are not. Only the boldface *HLA-B*35:01:01:02+HLA-B*53:24* genotype was originally reported for Child B. Based on the genotypes of Father, Mother and Child A, Child B may not carry the *HLA-B*35:01:01:02+HLA-B*53:24* genotype. After reviewing the DNA sequence alignment of *HLA-B*35:01:01:02, -B*35:08:01, -B*53:01:01* and *-B*53:24* (Figure 2), we concluded that this was a genotype reporting error, and that a genotypic ambiguity, *HLA-B*35:08:01+HLA-B*53:01:01*|*HLA-B*35:01:01:02+HLA-B*53:24*, was not reported.