Human Immunology xxx (2015) xxx–xxx

Contents lists available at ScienceDirect

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

# A gene feature enumeration approach for describing HLA allele polymorphism

Steven J. Mack

*Children's Hospital Oakland Research Institute, 5700 Martin Luther King Jr. Way, Oakland, CA 94609, USA*

ARTICLE INFO

*Article history:*
Received 2 February 2015
Revised 24 September 2015
Accepted 24 September 2015
Available online xxxx

*Keywords:*
HLA
Nomenclature
Gene feature enumeration
Next generation sequencing
IHIW
17th workshop

ABSTRACT

HLA genotyping via next generation sequencing (NGS) poses challenges for the use of HLA allele names to analyze and discuss sequence polymorphism. NGS will identify many new synonymous and non-coding HLA sequence variants. Allele names identify the types of nucleotide polymorphism that define an allele (non-synonymous, synonymous and non-coding changes), but do not describe how polymorphism is distributed among the individual features (the flanking untranslated regions, exons and introns) of a gene. Further, HLA alleles cannot be named in the absence of antigen-recognition domain (ARD) encoding exons. Here, a system for describing HLA polymorphism in terms of HLA gene features (GFs) is proposed. This system enumerates the unique nucleotide sequences for each GF in an HLA gene, and records these in a GF enumeration notation that allows both more granular dissection of allele-level HLA polymorphism and the discussion and analysis of GFs in the absence of ARD-encoding exon sequences.


## 1. Introduction

The human leucocyte antigen (HLA) genes are well known as the most polymorphic loci in the human genome. The extensive sequence polymorphism known for the HLA alleles is curated by the ImMunoGeneTics (IMGT)/HLA Database [1], which annotates the individual features for each gene [nucleotide sequences of each exon, intron and flanking untranslated region (UTR)] and gene product (encoded protein sequences). Here, exons, introns and UTRs are collectively referred to as gene features (GFs) to distinguish them from "sequence features" described elsewhere [2].

The World Health Organization Nomenclature Committee for factors of the HLA system (HLA Nomenclature Committee) assigns a unique allele name to each unique HLA nucleotide sequence [3]. Each HLA allele name consists of four colon-delimited fields (e.g., *HLA-A*01:01:01:01*). The first field identifies the allele family (for all genes but *HLA-DPB1*); the second field enumerates the unique protein sequences for the alleles in a given allele family, in the order in which they were identified; the third field enumerates sequences with synonymous substitutions for a given protein sequence, in the order in which they were identified; and the fourth field enumerates sequences with nucleotide substitutions in UTRs and introns for a given synonymous sequence in an exon, in the order in which they were identified. *HLA-DPB1* lacks allele-families; the first field identifies unique protein sequences for all but the *DPB1*02* and **04* alleles, for which two distinct protein sequences each are known [3–5].

The IMGT/HLA Database is updated every three months, and the number of named HLA gene and pseudogene sequences increases with each update. For example, 9946 HLA alleles had been named as of December of 2013 [6]; this number increased to 12,242 in December of 2014 [7], and 13,412 HLA alleles have been named as of July of 2015. Increases in the number of new allele sequences included in the database have followed the adoption of new genotyping technologies by the Histocompatibility and Immunogenetics (H&I) community, often in conjunction with international HLA and immunogenetics workshops (IHIWs).

The IMGT/HLA Database annotation, based on European Molecular Biology Laboratory (EMBL) formats, is available as hla.dat and hla.xml files from ftp.ebi.ac.uk. These files identify and characterize the nucleotide sequences corresponding to specific GFs for each HLA allele. As illustrated in Table 1, each HLA gene can have a different number of GFs, but all HLA genes have a 3′ and 5′ UTR, at least four exons and at least three introns. However, for most HLA genes, full-length sequence is unavailable for the majority of

*Abbreviations:* ARD, antigen recognition domain; EMBL, European Molecular Biology Laboratory; GF, gene feature; GFE, gene feature enumeration; HLA, human leucocyte antigen; IHIW, international HLA and immunogenetics workshop; IMGT, ImMunoGeneTics; NGS, next generation sequencing; UTR, untranslated region.

*E-mail address:* SJMACK@CHORI.ORG

http://dx.doi.org/10.1016/j.humimm.2015.09.016
0198-8859/

**Table 1**
Maximum Lengths of gene features in 11 HLA genes in IMGT/HLA Database release 3.21.1.

| Locus | 5′ UTR | Exon 1 | Intron 1 | Exon 2 | Intron 2 | Exon 3 | Intron 3 | Exon 4 | Intron 4 | Exon 5 | Intron 5 | Exon 6 | Intron 6 | Exon 7 | Intron 7 | Exon 8 | 3′ UTR |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| *HLA-A* | 300 | 73 | 130 | 293 | 242 | 314 | 600 | 280 | 102 | 117 | 442 | 33 | 142 | 48 | 169 | 5 | 300 |
| *HLA-B* | 284 | 73 | 129 | 272 | 250 | 281 | 575 | 277 | 104 | 120 | 441 | 33 | 107 | 44 | | | 364 |
| *HLA-C* | 283 | 73 | 130 | 277 | 250 | 297 | 587 | 277 | 124 | 138 | 440 | 33 | 107 | 48 | 164 | 5 | 171 |
| *HLA-DPA1* | 523 | 100 | 3584 | 246 | 340 | 282 | 214 | 155 | | | | | | | | | 4331 |
| *HLA-DPB1* | 366 | 100 | 4536 | 264 | 4014 | 282 | 547 | 111 | 329 | 30 | | | | | | | 963 |
| *HLA-DQA1* | 746 | 82 | 3858 | 249 | 445 | 282 | 429 | 155 | | | | | | | | | 436 |
| *HLA-DQB1* | 530 | 109 | 1458 | 270 | 2889 | 291 | 517 | 111 | 485 | 24 | 611 | 14 | | | | | 206 |
| *HLA-DRB1* | 607 | 100 | 10306 | 272 | 3464 | 282 | 702 | 111 | 487 | 24 | 1142 | 14 | | | | | 325 |
| *HLA-DRB3* | 327 | 100 | 7681 | 270 | 2302 | 282 | 684 | 111 | 473 | 24 | 799 | 14 | | | | | 579 |
| *HLA-DRB4* | 313 | 100 | 9563 | 270 | 2741 | 282 | 704 | 111 | 474 | 24 | 302 | 14 | | | | | 570 |
| *HLA-DRB5* | 0 | 100 | 0 | 270 | 0 | 282 | 0 | 111 | 0 | 24 | 0 | 14 | | | | | 0 |

The maximum length of the nucleotide sequences for each gene feature (GF) [untranslated region (UTR), exon or intron] for each HLA gene in 12,332 HLA alleles in IMGT/HLA Database release 3.21.1 is shown.
Blank cells indicate that no GF exists for that gene. Values of 0 indicate that no sequences for that GF have been included in available IMGT/HLA Database annotations.

alleles. As illustrated in Fig. 1, nucleotide sequences for more than 60% of HLA-A, -B, -C, and –DRB1 alleles in IMGT/HLA Database version 3.21.1 are available only for exons 2 and 3 of the class I genes, and exon 2 of the class II genes, as these exons encode the antigen recognition domain (ARD). Fewer than 8% of the alleles at these loci have full-length sequences, describing nucleotide sequence for all of an allele's GFs. Many of these full-length sequences have been generated using next generation sequencing (NGS) technologies, and the number of HLA alleles included in the database seems poised to increase dramatically as NGS technologies become widely used for HLA genotyping by the H&I and genomics communities, and as part of the 17th IHIW.

### *1.1. Application of NGS technology highlights current nomenclature limitations*

The four colon-delimited field nomenclature for HLA alleles developed in step with genotyping technologies, as greater insights into the nature and scope of HLA polymorphism became available [4,8–12]. While it provides insight into the *types* of polymorphism that distinguish alleles, this nomenclature does not identify the patterns and location of polymorphism across GFs at a given locus; the extent of the nucleotide sequence represented by an HLA allele name cannot be inferred from that name. The former issue has been partially addressed by extending allele names to identify those alleles that share identical ARD-encoding exon sequences (G groups of alleles, e.g., *HLA-A*01:01:01G*), as well as those alleles that encode identical ARD protein sequences (P groups of alleles, e.g., A*01:01P) [3], as these GFs constitute the largest fraction of the database. However, outside of the G group extension, alleles that share nucleotide sequences for other GFs cannot easily be identified. For example, class I alleles that share identical sequences for one of the ARD-encoding exons, but not the other, cannot be identified using G groups.

The sequences of ARD-encoding exons are required for all nucleotide sequence submissions to the HLA Nomenclature Committee via the IMGT/HLA Database, and novel nucleotide sequences for non-coding GFs must be submitted as part of full-length sequences. As a result, an HLA allele name cannot be assigned to a novel nucleotide sequence for an individual GF of interest (e.g., the 3′ UTR of *HLA-C* [13,14]) in the absence of nucleotide sequences for ARD-encoding GFs.

Klitz and Hedrick [15] have estimated that millions of alleles persist in the human population for each HLA gene. As NGS technologies extend sequence knowledge into non-ARD encoding GFs, the number of alleles distinguished by synonymous and non-coding variants can be expected to increase dramatically; for example, as illustrated in Table 1, introns 1 and 2 of class II genes can be several thousand nucleotides long, and are likely to have accumulated many nucleotide variants. These variants will be noted in the third and fourth fields of allele names, and it does not seem out of the case to imagine allele names like HLA-DRB1*01:01:100:1004 in the near future. As the number of full-length HLA gene sequences generated increases, it seems likely that a large fraction of them will be unique.

Given the inability to determine which GFs are represented in an HLA allele name, the inability to assign allele names to individual non-ARD-encoding GFs, and the impending likelihood of a large number of unique full-length gene sequences, the utility of the HLA nomenclature is limited for managing, exchanging, discussing and analyzing nucleotide sequences for HLA GFs without the context of ARD-encoding GFs.

Here, a gene feature enumeration (GFE) notation is proposed as a supplement to the current HLA nomenclature for the purposes of cataloging nucleotide sequence polymorphisms for non-ARD-encoding GFs, discussing and analyzing HLA alleles in the context of polymorphism distributed between GFs, and capturing novel nucleotide sequences for non-ARD-encoding GFs generated via NGS technologies. This GFE approach is being developed as part of the 17th IHIWS Informatics Component.

## 2. Gene feature enumeration

HLA allele name nomenclature enumerates non-synonymous, synonymous and non-coding nucleotide variants in the second through fourth fields of an allele name. To supplement this approach, the unique sequences in each GF of a given HLA gene can be sequentially numbered, and applied to construct a second name for that allele consisting of one field for each GF, containing the unique number for that GF nucleotide sequence and delimited by dashes, prefaced with the allele name followed by a 'w' (for Workshop) to identify the provisional nature of this notation [16,17]. This GFE notation is illustrated in Table 2.

For example, the *HLA-A* gene includes 17 GFs (Table 1); any *HLA-A* allele can be represented as a unique haplotype of 17 GFs. As illustrated in Table 2, and in Supplementary Table S1, *HLA-A*01:01:01:01* can be described in GFE notation as HLA-Aw1-1-1-1-1-1-1-1-1-1-1-1-1-1-1-1-1, identifying the constituent sequences of its GFs. In this case, the sequences for each *HLA-A*01:01:01:01 GF* are the first sequences numbered for those *HLA-A* GFs. The approach applied to assign these GF numbers is described in Section 2.1.

Not all nucleotide sequences for a GF of an HLA gene are the same length. In some cases, these length differences are due to incomplete sequence of the GF in question, and in other cases they are due to insertion–deletion mutations. Using GFE notation, GF

Please cite this article in press as: Mack SJ. A gene feature enumeration approach for describing HLA allele polymorphism. Hum Immunol (2015), http://dx.doi.org/10.1016/j.humimm.2015.09.016