Clinical Chemistry 62:12
000–000 (2016)

Generation of Full-Length Class I Human Leukocyte Antigen Gene Consensus Sequences for Novel Allele Characterization

Peter M. Clark,[1] Jamie L. Duke,[1] Deborah Ferriola,[1] Valia Bravo-Egana,[1] Tunde Vago,[2] Aniqa Hassan,[1] Anna Papazoglou,[1] and Dimitri Monos[1,3*]

BACKGROUND: Routine, high-resolution human leukocyte antigen (HLA) genotyping by next generation sequencing within clinical immunogenetics laboratories can now provide the full-length gene sequence characterization of fully phased HLA alleles. This powerful technique provides insights into HLA variation beyond the traditionally characterized antigen recognition domain, providing sequence annotation across the entire gene including untranslated and intronic regions and may be used to characterize novel alleles from massively parallel sequencing runs.

METHODS: We evaluated the utility of the Omixon Holotype HLA assay to generate credible, fully phased full-length gene consensus sequences for 50 individuals at major histocompatibility complex, class I, A (*HLA-A*), *HLA-B*, and *HLA-C* loci (300 genotyped alleles in total) to identify and characterize novel class I HLA alleles using our downstream analytical pipeline.

RESULTS: Our analysis revealed that 7.7% (23/300) of genotyped class I HLA alleles contain novel polymorphisms. Interestingly, all of the novel alleles identified by our analysis were found to harbor sequence variations within intronic regions of the respective locus. In total our analysis identified 17 unique novel class I HLA alleles from 23 of the 300 genotyped alleles and generated full-length gene sequence annotations for 9 previously incompletely annotated HLA class I allele sequences derived from 14 of the 300 genotyped alleles.

CONCLUSIONS: The demonstrated utility of the Omixon Holotype HLA assay in combination with our downstream analytical framework to generate fully phased, full-length gene consensus sequences for the identification and characterization of novel HLA alleles, facilitates the study of HLA polymorphism beyond the antigen recognition domain in human health and disease.

Human leukocyte antigen (HLA)[4] molecules are potent regulators of the immune response, conferring protection against foreign pathogens through the presentation of HLA class I and class II antigens to $CD8^+$ T cells and CD4+ T cells, respectively, as well as killer immunoglobulin-like receptors *(1)*. As a result of their essential immunological function, HLA genes are key determinants in organ transplantation compatibility *(2, 3)* and have been associated with a variety of autoimmune and infectious diseases *(4–6)*.

The ubiquitously expressed transmembrane class I HLA molecule is a heterodimer consisting of a highly polymorphic heavy (α) chain that is linked to the β_2-microglobulin light chain through noncovalent interaction with the α3 domain of the heavy chain. The class I HLA genes (located on the short arm of chromosome 6) each encode the α chain of their respective heterodimer (approximately 43 kDa), whereas β_2-microglobulin is transcribed as a separate gene from a locus on chromosome 15. Major histocompatibility complex, class I, A (*HLA-A*)[5] and *HLA-C* both contain 8 exons, and *HLA-B* contains 7 exons. The α1 and α2 domains of the heavy chain are encoded by exons 2 and 3 of the HLA gene respectively and together form the peptide-binding cleft, which defines the binding specificity of the antigen recognition domain (ARD). For each gene, exon 1 encodes the leader peptide (also termed signal peptide), exon 4 encodes the α3 domain, exon 5 encodes the transmem-

[1] Department of Pathology and Laboratory Medicine, The Children's Hospital of Philadelphia, Philadelphia, PA; [2] Omixon Biocomputing, Budapest, Hungary; [3] Department of Pathology and Lab Medicine, Perelman School of Medicine, University of Pennsylvania, Philadelphia PA.

[*] Address correspondence to this author at: The Children's Hospital of Philadelphia, Philadelphia PA, 19104. Fax 215-590-6361; e-mail monosd@email.chop.edu.

Received May 20, 2016; accepted August 19, 2016.
Previously published online at DOI: 10.1373/clinchem.2016.260661

[4] Nonstandard abbreviations: HLA, human leukocyte antigen; ARD, antigen recognition domain; RFLP, restriction fragment length polymorphism; PCR-SSOP, PCR-sequence specific oligonucleotide probe; SSP-PCR, single specific primer PCR; SBT, Sanger sequence based typing; NGS, next generation sequencing; IMGT, ImMunoGeneTics information system; qPCR, quantitative PCR; MSA, multiple sequence alignment; UTR, untranslated region.

[5] Human genes: *HLA-A,* major histocompatibility complex, class I, A; *HLA-B,* major histocompatibility complex, class I, B; *HLA-C,* major histocompatibility complex, class I, C.

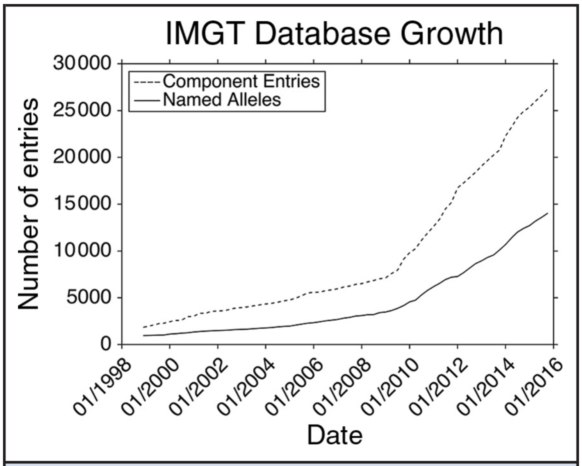

Fig. 1. The number of component entries [user submitted EMBL (European Molecular Biology Laboratory)-Bank/GenBank/DDBJ (DNA Data Bank of Japan) entries] and named alleles (expertly annotated database entries with an assigned WHO nomenclature allele name) within the IMGT/HLA database (release 1.0–3.22.0).

It should be noted that there may be multiple user submitted component entries for a single named allele.

brane region and exons 6–8 encode the cytoplasmic tail of the peptide *(7)*.

Given the need for routine HLA genotyping for histocompatibility testing, the Immunogenetics community has characterized over 14000 named HLA alleles using a variety of techniques including traditional serological approaches and more recent DNA-based methodologies including restriction fragment length polymorphism (RFLP) analysis, PCR-sequence specific oligonucleotide probes (PCR-SSOP), sequence specific primer PCR (SSP-PCR), Sanger sequence based typing (SBT) and most recently next generation sequencing (NGS) *(8–19)*. HLA genotyping by NGS has facilitated the routine characterization of full-length HLA alleles with unprecedented throughput, providing clinicians and researchers alike with sequence annotation beyond the traditionally characterized ARD.

With the advent and growing implementation of routine HLA genotyping by NGS within clinical immunogenetics laboratories, the pace of discovery and annotation of novel HLA alleles continues to grow. The International ImMunoGeneTics information system (IMGT) database *(20–22)* contains annotations for 14015 named HLA alleles (release 3.22), and the pace of discovery continues to increase *(22)*, with over 1600 novel alleles discovered within the last year alone (Fig. 1). Consequently, there is a clear need for a NGS based HLA genotyping method and analytical pipeline that not only can generate accurate HLA genotyping information, but also generate an accurate, fully phased HLA consensus sequence for both genotyped alleles. In the work presented here, we utilize the NGS based Omixon Holotype HLA kit in combination with the Omixon Twin software to genotype 50 samples at class I HLA loci (*HLA-A*, *HLA-B*, and *HLA-C*) and generate a fully phased consensus sequence for each genotyped allele (300 in total) to characterize and annotate novel HLA alleles using a custom developed analytical pipeline.

Materials and Methods

SAMPLE PREPARATION

Genomic DNA for 50 individuals was extracted from peripheral blood using the Qiagen EZ1® DNA extraction platform in conjunction with the Qiagen EZ1® Blood 350 μL Kit. In our experience, the quality of DNA from this platform and kit is consistently appropriate for long range PCR; therefore, the quality (A260/230, A260/280) was not assessed. DNA was quantified with Qubit BR assay and adjusted to a concentration of 30 ng/μl.

Sample DNA was amplified at 3 loci (*HLA-A*, *HLA-B*, and *HLA-C*) by long-range PCR using the Omixon Holotype HLA genotyping kit, generating full-length gene amplicons. Following PCR, amplicons were cleaned with Exo-SAP (Affymetrix), quantified with QuantiFluor dsDNA system (Promega), and normalized to approximately 70 ng/μl.

LIBRARY PREPARATION AND SEQUENCING

Sequencing libraries were generated for each sample using the Omixon Holotype HLA Genotyping Kit as previously described *(18)*. In brief, libraries from individual HLA amplicons were prepared by enzymatic fragmentation, end repair, adenylation, and ligation of indexed adaptors. The indexed libraries were pooled and concentrated with Ampure XP beads (Beckman Coulter) before fragment size selection using a PippinPrep™ (Sage Science), selecting a range of fragments between 650 and 1300 bp. The size-selected library pool was quantified by quantitative PCR (qPCR; Kapa Biosystems) and adjusted to 2 nmol/L. The library was then denatured with NaOH and diluted to a final concentration of 8 pmol/L for optimal cluster density and 600 μL was loaded into the MiSeq reagent cartridge (v2 500 cycle kit). The reagent cartridge and flow cell were placed on the Illumina MiSeq (Illumina) for cluster generation and 2 × 250 bp paired-end sequencing. Samples were demultiplexed on the instrument and the resulting FASTQ files were used for further analysis.

COMPUTATIONAL ANALYSIS

All 50 samples were genotyped for the 3 class I loci using Twin^{TM} version 1.1.1 (Omixon, Inc.) and IMGT/HLA

Table 1. Frequencies of observed alleles and novel alleles for each genotyped allele at the *HLA-A* locus.[a]

	HLA-A				
	Allele	Complete IMGT/HLA allele	Allele frequency	Novel allele frequency	Polymorphism location
	A*02:01:01:01	Yes	18	1	Intron 3; 1419 T>C
	A*24:02:01:01	Yes	8	3	Intron 3; 1248 C>T intron 3; 1565 G>T intron 5; 2410 C>G
	A*11:01:01:01	Yes	7	0	
	A*03:01:01:01	Yes	6	2	Intron 6; 2605 C>T intron 6; 2605 C>T
	A*32:01:01	Yes	6	0	
	A*02:06:01:01	Yes	5	0	
	A*31:01:02:01	Yes	5	0	
	A*01:01:01:01	Yes	4	1	Intron 1; 161 C>G
	A*23:01:01	Yes	4	0	
	A*33:03:01	Yes	4	0	
	A*68:02:01:01	Yes	4	0	
	A*02:05:01	Yes	3	0	
	A*26:01:01:01	Yes	3	0	
	A*74:01	Yes	3	0	
	A*25:01:01	Yes	2	0	
	A*29:02:01:01	Yes	2	0	
	A*33:01:01	Yes	2	0	
	A*02:02:01	Yes	1	0	
	A*02:04	No (complete CDS)	1	0	
	A*02:07:01	Yes	1	0	
	A*03:02:01	Yes	1	0	
	A*24:03:01	Yes	1	1	Intron 6; 2627 C>T
	A*29:01:01:01	Yes	1	0	
	A*29:02:01:02	Yes	1	0	
	A*30:01:01	Yes	1	0	
	A*30:02:01:03	Yes	1	0	
	A*30:04:01	Yes	1	0	
	A*66:01:01	Yes	1	0	
	A*68:01:01:02	Yes	1	0	
	A*68:01:02:01	Yes	1	0	
	A*68:01:02:02	Yes	1	0	
Total	31	30	100	8	

[a] Fifty samples were genotyped at the *HLA-A* locus, comprising 31 unique *HLA-A* alleles including one allele (A*02:04) with an incomplete full-length gene sequence annotation (IMGT/HLA release 3.20.0). The absolute frequency of observed alleles and novel alleles of each genotyped allele is tabulated along with the base position of the observed polymorphism within novel alleles (polymorphism location is calculated from the first base of exon 1).

database version 3.20.0, using the default settings. In the case that an allele could not be fully phased using the default set of 4000 read-pairs, the number of read-pairs was increased to 16000. Fully phased consensus sequences were generated for each allele and exported from Twin^{TM} in FASTA format. The consensus sequences for each allele were aligned to their respective genotyped IMGT/HLA reference alleles using the Needleman–Wunsch semiglobal alignment algorithm, as implemented within MATLAB (R2014b). In cases in which the full-length reference sequence is not annotated within the IMGT/HLA database, consensus sequences were aligned to the available annotated sequence for the given genotyped allele [complete or partial CDS (coding

Table 2. Frequencies of observed alleles and novel alleles for each genotyped allele at the *HLA-B* locus.[a]

		HLA-B		
Allele	**Complete IMGT/HLA allele**	**Allele frequency**	**Novel allele frequency**	**Polymorphism location**
B*07:02:01	Yes	7	0	
B*51:01:01:01	Yes	5	0	
B*58:02	No (complete CDS)	5	0	
B*14:02:01	Yes	4	0	
B*15:10:01	Yes	4	0	
B*45:01:01	Yes	4	0	
B*53:01:01	Yes	4	0	
B*15:03:01	Yes	3	3	Intron 3; 1467 C>T intron 3; 1467 C>T intron 3; 1467 C>T
B*44:03:01	Yes	3	1	Intron 3; 1259 G>C
B*48:01:01	Yes	3	0	
B*50:01:01	Yes	3	0	
B*14:01:01	Yes	2	0	
B*14:03	No (annotated exons 2, 3, and 4)	2	0	
B*15:16:01	Yes	2	2	Intron 2; 666 CGGGG Intron 2; 666 CGGGG
B*15:18:01	Yes	2	0	
B*27:05:02	Yes	2	0	
B*38:01:01	Yes	2	0	
B*39:05:01	Yes	2	0	
B*40:01:02	Yes	2	0	
B*40:02:01	Yes	2	0	
B*44:02:01:01	Yes	2	0	
B*55:01:01	Yes	2	0	
B*57:01:01	Yes	2	0	
B*58:01:01:01	Yes	2	0	
B*07:05:01	Yes	1	0	
B*08:01:01	Yes	1	0	
B*13:02:01	Yes	2	0	
B*15:01:01:01	Yes	1	0	
B*15:02:01	Yes	1	0	
B*15:07:01	Yes	1	0	
B*15:08:01	No (complete CDS)	1	0	
B*15:30	No (annotated exons 2,3, and 4)	1	0	
B*18:01:01:01	Yes	1	0	
B*27:02:01	Yes	1	0	
B*35:01:01:01	Yes	1	0	
B*35:02:01	Yes	1	0	
B*35:03:01	Yes	1	0	
B*35:04:01	No (complete CDS)	1	0	

Continued on page XXX

Table 2. Frequencies of observed alleles and novel alleles for each genotyped allele at the *HLA-B* locus.[a]
(Continued from page XXX)

	HLA-B				
	Allele	Complete IMGT/HLA allele	Allele frequency	Novel allele frequency	Polymorphism location
	B*35:43:01	No (complete CDS)	1	0	
	B*37:01:01	Yes	1	0	
	B*39:01:01:02L	Yes	1	0	
	B*39:09:01	No (complete CDS)	1	0	
	B*40:06:01:01	Yes	1	0	
	B*41:01:01	Yes	1	0	
	B*44:02:01:03	Yes	1	0	
	B*46:01:01	Yes	1	0	
	B*47:01:01:01	Yes	1	1	Intron 6; 2542 C>T; intron 6; 2552 A>G
	B*49:01:01	Yes	1	0	
	B*52:01:01:01	Yes	1	0	
	B*56:01:01:02	Yes	1	0	
	B*51:08:01	No (Complete CDS)	1	0	
	B*57:03:01	Yes	1	1	Intron 5; 2259 T>C
Total	52	44	100	8	

[a] Fifty samples were genotyped at the *HLA-B* locus, comprising 52 unique *HLA-B* alleles including eight alleles (B*58:02, B*14:03, B*15:08:01, B*15:30, B*35:04:01, B*35:43:01, B*39:09:01, B*51:08:01) with an incomplete full-length gene sequence annotation (IMGT/HLA release 3.20.0). The absolute frequency of observed alleles and novel alleles of each genotyped allele is tabulated along with the base position of the observed polymorphism within novel alleles (polymorphism location is calculated from the first base of exon 1).

sequences)]. A custom MATLAB program was developed to analyze the alignments, facilitating the detection and annotation of novel alleles and alleles with incomplete reference sequences in the IMGT/HLA database.

Sequence variability across HLA genes was quantified by calculating the Shannon entropy *(23, 24)* at each position of the multiple sequence alignment (MSA) of *HLA-A*, *HLA-B*, and *HLA-C* alleles respectively. Annotated, full-length HLA gene sequences for each locus were obtained from the IMGT/HLA database (release 3.22.0) and aligned using Clustal Omega *(25)*. For each respective locus, Shannon entropy was calculated at each position of the MSA (Equation 1) using a 4 character alphabet (ACGT) and plotted as a function of the position within each MSA using MATLAB (R2014b). Average Shannon entropy for each genomic feature [exon, intron, or untranslated region (UTR)] of *HLA-A*, *HLA-B*, and *HLA-C* was calculated by dividing the sum of Shannon entropy over the feature divided by the total length of the genomic feature.

$$H = -\sum_{i=1}^{n} p_i \log_2 (p_i) \qquad (1)$$

In information theory, the Shannon entropy (H) defines the "randomness" or variability of a system by quantifying the information content within a given message, where p_i is the frequency of a random variable i, with n states (n = 4, ATCG). Shannon entropy was calculated at each position of the MSA such that if every base occurred with a frequency of ¼, the Shannon entropy would be maximal with a value of 2 bits, whereas if a specific position of the MSA contained the same letter for each allele (no variability), the Shannon entropy would be equal to zero.

Results

HLA genotyping results and fully phased consensus sequences were successfully generated for all 50 participants using the Omixon Twin algorithm (300 total alleles) and were found to be concordant with SBT/SSP genotyping results (2 field resolution, data not shown). From the HLA genotyping results, our sample set was found to contain 31 unique *HLA-A* (Table 1), 52 *HLA-B* (Table 2), and 31 *HLA-C* alleles (Table 3). Our analysis reveals that 7.7% (23/300) of our characterized class I alleles contain a previously uncharacterized polymorphism, and another 4.7% (14/300) of the genotyped alleles contain incomplete sequence annotations within IMGT (release 3.22.0). In total our analysis identified 17 novel class I HLA allele sequences (7 *HLA-A*, 5 *HLA-B*,

Table 3. Frequencies of observed alleles and novel alleles for each genotyped allele at the *HLA-C* locus.[a]

	HLA-C				
	Allele	Complete IMGT/HLA allele	Allele occurrences	Novel allele frequency	Polymorphism location
	C*06:02:01:01	Yes	12	0	
	C*04:01:01:01	Yes	8	2	Intron 5; 2206 A>G intron 5; 2206 A>G
	C*07:02:01:03	Yes	8	0	
	C*08:02:01:01	Yes	6	0	
	C*15:02:01:01	Yes	5	0	
	C*01:02:01	Yes	4	0	
	C*02:02:02:01	Yes	4	0	
	C*03:04:02	Yes	4	0	
	C*16:01:01	Yes	4	0	
	C*02:10	Yes	3	1	Intron 1; 114 G>C
	C*03:03:01	Yes	3	0	
	C*03:04:01:01	Yes	3	0	
	C*06:02:01:02	Yes	3	0	
	C*07:01:01:01	Yes	3	1	Intron 6; 2616 T>C intron 6; 2619 T>C
	C*07:02:01:01	Yes	3	0	
	C*07:04:01	Yes	3	0	
	C*08:01:01	Yes	3	0	
	C*12:03:01:01	Yes	3	0	
	C*05:01:01:02	Yes	2	0	
	C*06:02:01:03	Yes	2	0	
	C*08:02:01:02	Yes	2	0	
	C*08:03:01	Yes	2	0	
	C*14:02:01	Yes	2	2	Intron 3; 1117 G>C intron 5; 2110 G>T
	C*03:02:02:01	Yes	1	0	
	C*03:05	Yes	1	0	
	C*07:01:02	Yes	1	0	
	C*07:18	Yes	1	0	
	C*12:02:02	Yes	1	0	
	C*15:05:02	Yes	1	0	
	C*16:02:01	Yes	1	0	
	C*17:01:01:02	Yes	1	1	Intron 7; 2902 G>T
Total	31	31	100	7	

[a] Fifty samples were genotyped at the *HLA-C* locus, comprising 31 unique *HLA-C* alleles. The absolute frequency of observed alleles and novel alleles of each genotyped allele is tabulated along with the base position of the observed polymorphism within novel alleles (polymorphism location is calculated from the first base of exon 1). Two individuals were found to have an identical consensus sequence for the genotyped C*14:02:01 allele, which harbors 2 distinct polymorphisms, resulting in the reporting of only a single novel allele sequence.

and 5 *HLA-C*). All identified novel alleles were found to harbor intronic polymorphisms. Examining the set of 7 unique identified novel *HLA-A* alleles, polymorphisms were found within intron 1 (n = 1), intron 3 (n = 3), intron 5 (n = 1), and intron 6 (n = 2). For *HLA-B*, the set of 5 unique identified novel alleles were found to harbor polymorphisms within intron 3 (n = 2), intron 5 (n = 1), and intron 6 (n = 2). For *HLA-C*, the set of 5 unique identified novel alleles were found to harbor polymorphisms within intron 1 (n = 1), intron 5 (n = 1), intron 6 (n = 1), intron 7 (n = 1), and 1 novel allele with polymorphisms within both intron 5 and intron 3 (n = 1).

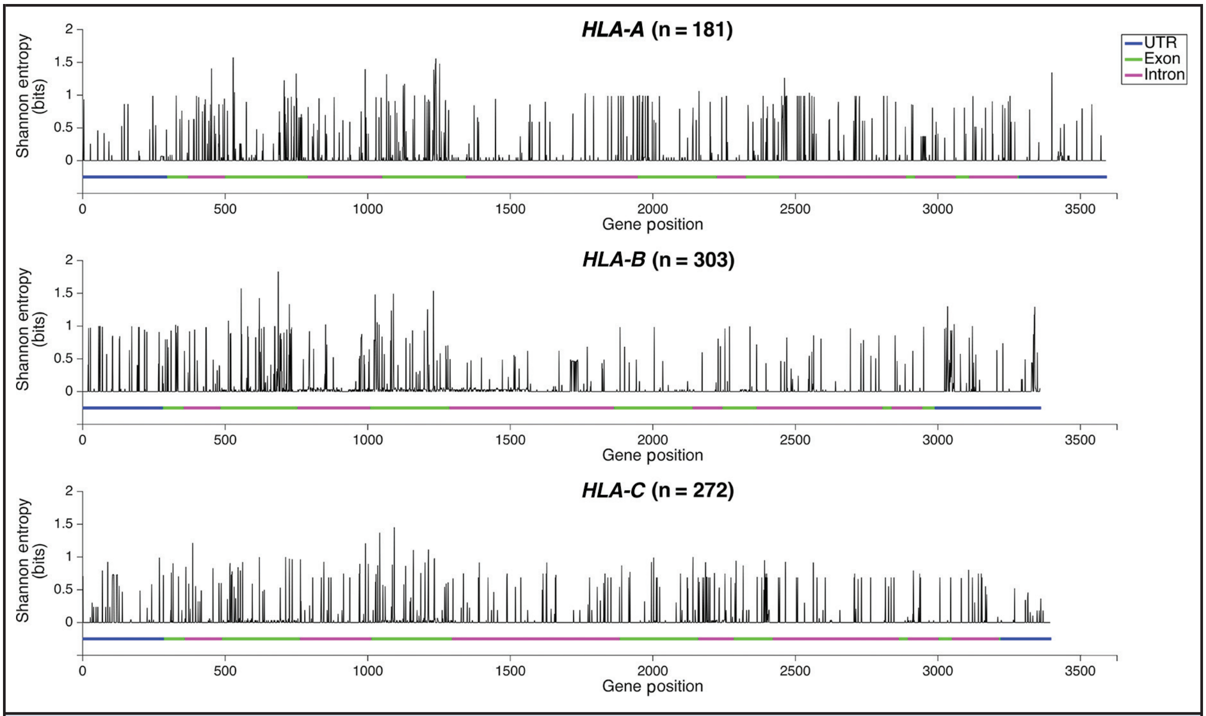

Fig. 2. HLA variability across class I HLA genes.

Shannon entropy was calculated from the MSA of every annotated full-length HLA gene for *HLA-A* (n = 181), *HLA-B* (n = 303), and *HLA-C* (n = 272) loci respectively (IMGT/HLA database release 3.22.0). All gene annotations are shown from 5′->3′ with the 5′ UTR shown in blue from the left end, followed by exon 1, intron 1, et al.

Additionally the set of 50 samples included 14 individuals with 9 incompletely characterized HLA allele sequences (1 *HLA-A* and 8 *HLA-B*, Table 1 and 2), (IMGT/HLA release 3.22.0).

The sequence variability of class I HLA loci was assessed to quantify the relative occurrence and location of polymorphisms within fully characterized alleles (IMGT/HLA release 3.22.0). An MSA for each HLA locus was generated and the Shannon entropy (Equation 1) was calculated at each position of the MSA (Fig. 2). The average Shannon entropy was calculated over each genomic feature (UTRs, exons, and introns) for each gene (Table 4) to better assess the average variability of specific genomic features and make comparisons between the different HLA genes. Our analysis revealed that the highest average sequence variability was within exon 3 of *HLA-A* and *HLA-C* genes, whereas exon 2 of *HLA-B* was found to have the highest average sequence variability, followed closely by exon 3. For *HLA-A* and *HLA-C*, intron 1 and 4 were found to have the second highest average sequence variability respectively. The region with the lowest average variability was found to be exon 8 of *HLA-A* and exon 6 for both *HLA-B* and *HLA-C*.

The complete, fully characterized sequence of each identified novel HLA gene, stretching from the 5′ end of exon 1 through the 3′ end of the last exon (exon 7 of *HLA-B* and exon 8 of *HLA-A* and *HLA-C*) including all encompassed introns were submitted to GenBank and the WHO HLA Nomenclature Committee for factors of the HLA system. The WHO Nomenclature Committee officially assigned every allele name in April of 2016 following the policy outlined in the most recent Nomenclature Report *(26)*. Assigned allele names (WHO nomenclature) and GenBank accession numbers for the 17 identified novel alleles (7 *HLA-A*, 5 *HLA-B*, and 5 *HLA-C*) and the 9 completed alleles as well as the race and ethnicity information for each sample in which a novel allele was identified is provided in Table 1 in the Data Supplement that accompanies the online version of this article at http://www.clinchem.org/content/vol62/issue12.

Discussion

HLA genotyping by NGS facilitates high-resolution HLA genotyping with unprecedented throughput. However, despite this important advancement within the field of clinical immunogenetics, the ability to characterize and set phase across the length of the gene using Illumina sequencing technology has yet to be fully realized. The

Table 4. Mean Shannon entropy (bits/base-pair) calculated for each genomic feature of *HLA-A*, *HLA-B*, and *HLA-C* as depicted in Figure 2.[a]

Genomic feature	*HLA-A*	*HLA-B*	*HLA-C*
5′ UTR	0.030	0.065	0.056
Exon 1	0.045	0.114	0.051
Intron 1	0.105	0.058	0.048
Exon 2	0.087	0.126	0.064
Intron 2	0.055	0.068	0.044
Exon 3	0.112	0.125	0.082
Intron 3	0.039	0.040	0.032
Exon 4	0.062	0.019	0.047
Intron 4	0.024	0.022	0.081
Exon 5	0.068	0.037	0.075
Intron 5	0.079	0.026	0.031
Exon 6	0.054	0.000	0.005
Inrtron 6	0.046	0.024	0.029
Exon 7	0.025	0.022	0.029
Intron 7	0.063	N/A	0.036
Exon 8	0.000	N/A	0.014
3′ UTR	0.025	0.068	0.017

[a] Data values are shaded in with varying opacity to visualize those regions with the highest mean sequence variation (shown in darker grey) from those of lower mean sequence variation (shown in white).

Omixon Twin software as well as our implemented custom downstream analytical framework addresses this growing gap, allowing full-length gene read phasing, consensus sequence generation and the characterization of novel class I HLA alleles.

Because all novel alleles were found to exclusively harbor intronic polymorphisms, we sought to characterize HLA variability by genomic feature (UTR, intron, and exon) to determine whether or not intronic variations are more or less common than polymorphisms within other regions of the HLA genes. Because the sample set used to calculate sequence variability only contains completely characterized HLA alleles (IMGT/HLA release 3.22.0), we do not under or oversample for any particular region relative to others, with each MSA containing equal numbers of alleles characterized at each distinct genomic feature. However, the analyzed sequences are those reported in IMGT, which have been generated as the Immunogenetics community identifies new sequences primarily in the exons and only incidentally in the introns. Thus, the collection of data is driven by the exonic differences and not the unbiased sequence of every HLA allele. As a result, sequence variation within different genomic features outside of the traditionally characterized exons is likely underrepresented in the current database.

Our results, using completely characterized allele sequences from IMGT, indicate differential sequence variability across various genomic features within and across class I HLA alleles. Exon 3, which encodes the highly variable $\alpha2$ domain, was found to have the highest average sequence diversity across the set of fully characterized *HLA-A* and *HLA-C* IMGT alleles (release 3.22.0) analyzed, whereas exon 2 was found to be the most variable region of *HLA-B* alleles (Table 4). Exon 8 was found to have the lowest sequence variability within *HLA-A*, whereas exon 6 was found to have the lowest sequence variability within *HLA-B* and *HLA-C*. Intronic sequences are also quite variable, whereby intron 1 of *HLA-A* is the most variable. However, our results indicate that there may be a level of intronic variation that corresponds to the same exonic sequences, which has not been previously appreciated. Because we have been using NGS technology for HLA genotyping we have discovered that intronic regions may also be significantly polymorphic. Considering that HLA typing by traditional legacy methods provides sequencing information for only a few of the exons, we were previously unaware of the extent of polymorphisms within introns. As additional alleles are fully characterized by NGS, the number of polymorphisms within each genomic feature, particularly within introns, may be substantially altered. Is it possible that as we perform full gene sequence characterization we will find that alleles with the same exonic sequences have different intronic sequences or otherwise have genes coding the same protein but differ in their intronic sequences? If so, what are the implications for the physiology of the cell? Do intronic regions play any role?

The Omixon Holotype HLA genotyping kit combined with the Omixon Twin software and our downstream analysis pipeline provides a robust framework for detecting and characterizing novel, full-length HLA alleles and paves the way for elucidating the role of HLA polymorphism in transplant medicine and autoimmune disease. As the use of HLA genotyping by NGS continues to increase within clinical immunogenetics laboratories, the number of novel alleles will only continue to increase at an expedited rate. Although there is currently little research on the physiological implications of polymorphisms within noncoding regions and coding regions beyond the ARD, research indicates that intron 4 of *HLA-B* harbors a miRNA, HSA-miR-6891, which is formed following exon splicing *(27, 28)*. Furthermore, research demonstrates that UTRs as well as coding and intronic regions are targeted by miRNA for translational suppression of mRNA transcripts *(29–31)*. With a 7.7% novel allele discovery rate for class I HLA genes, there are likely to be numerous yet undiscovered alleles of unknown significance. Taken together, full-length gene characterization is paramount for unambiguous HLA genotyping and facilitates a deeper understanding of HLA gene polymorphisms and the eventual role they may play in the immune response and the overall physiology of the cell.

Author Contributions: *All authors confirmed they have contributed to the intellectual content of this paper and have met the following 3 requirements: (a) significant contributions to the conception and design, acquisition of data, or analysis and interpretation of data; (b) drafting or revising the article for intellectual content; and (c) final approval of the published article.*

Authors' Disclosures or Potential Conflicts of Interest: *Upon manuscript submission, all authors completed the author disclosure form. Disclosures and/or potential conflicts of interest:*

Employment or Leadership: T. Vágó, Omixon Biocomputing; D. Monos, Omixon.
Consultant or Advisory Role: D. Monos, Omixon.
Stock Ownership: D. Monos, Omixon.
Honoraria: None declared.
Research Funding: None declared.
Expert Testimony: None declared.
Patents: None declared.
Other Remuneration: J.L. Duke, Omixon; D. Ferriola, Omixon.

Role of Sponsor: No sponsor was declared.

References

1. Trowsdale J. Genetic and functional relationships between MHC and NK receptor genes. Immunity 2001; 15:363–74.
2. Chinen J, Buckley RH. Transplantation immunology: solid organ and bone marrow. J Allergy Clin Immunol 2010;125:S324–35.
3. Sheldon S, Poulton K. HLA typing and its influence on organ transplantation. Methods Mol Biol 2006;333: 157–74.
4. Clark PM, Kunkel M, Monos DS. The dichotomy between disease phenotype databases and the implications for understanding complex diseases involving the major histocompatibility complex. Int J Immunogenet 2015;42:413–22.
5. Howell WM. HLA and disease: guilt by association. Int J Immunogenet 2014;41:1–12.
6. Dyer P, McGilvray R, Robertson V, Turner D. Status report from 'double agent HLA': health and disease. Mol Immunol 2013;55:2–7.
7. Malissen M, Malissen B, Jordan BR. Exon/intron organization and complete nucleotide sequence of an HLA gene. Proc Natl Acad Sci U S A 1982;79:893–7.
8. Monos D, Maiers MJ. Progressing towards the complete and thorough characterization of the HLA genes by NGS (or single-molecule DNA sequencing): consequences, opportunities and challenges. Hum Immunol 2015;76: 883–6.
9. Mayor NP, Robinson J, McWhinnie AJ, Ranade S, Eng K, Midwinter W, et al. HLA Typing for the Next Generation. PLoS One 2015;10:e0127153.
10. Lind C, Ferriola D, Mackiewicz K, Heron S, Rogers M, Slavich L, et al. Next-generation sequencing: the solution for high-resolution, unambiguous human leukocyte antigen typing. Hum Immunol 2010;71:1033–42.
11. Gabriel C, Furst D, Fae I, Wenda S, Zollikofer C, Mytilineos J, Fischer GF. HLA typing by next-generation sequencing–getting closer to reality. Tissue Antigens 2014;83:65–75.
12. Erlich H. HLA DNA typing: past, present, and future. Tissue Antigens 2012;80:1–11.
13. Eng HS, Leffell MS. Histocompatibility testing after fifty years of transplantation. J Immunol Methods 2011; 369:1–21.
14. Dunn PP. Human leucocyte antigen typing: techniques and technology, a critical appraisal. Int J Immunogenet 2011;38:463–73.
15. Dunckley H. HLA typing by SSO and SSP methods. Methods Mol Biol 2012;882:9–25.
16. De Santis D, Dinauer D, Duke J, Erlich HA, Holcomb CL, Lind C, et al. 16(th) IHIW: review of HLA typing by NGS. Int J Immunogenet 2013;40:72–6.
17. Bontadini A. HLA techniques: typing and antibody detection in the laboratory of immunogenetics. Methods 2012;56:471–6.
18. Duke JL, Lind C, Mackiewicz K, Ferriola D, Papazoglou A, Gasiewski A, et al. Determining performance characteristics of an NGS-based HLA typing method for clinical applications. HLA 2016;87:141–52.
19. Duke JL, Lind C, Mackiewicz K, Ferriola D, Papazoglou A, Derbeneva O, et al. Towards allele-level human leucocyte antigens genotyping–assessing two next-generation sequencing platforms: Ion Torrent Personal Genome Machine and Illumina MiSeq. Int J Immunogenet 2015;42:346–58.
20. Lefranc MP. IMGT, the International ImMunoGeneTics Information System. Cold Spring Harb Protoc 2011; 2011:595–603.
21. Lefranc MP, Giudicelli V, Ginestoux C, Jabado-Michaloud J, Folch G, Bellahcene F, et al. IMGT, the international ImMunoGeneTics information system. Nucleic Acids Res 2009;37:D1006–12.
22. Robinson J, Halliwell JA, McWilliam H, Lopez R, Parham P, Marsh SG. The IMGT/HLA database. Nucleic Acids Res 2013;41:D1222–7.
23. Shannon CE. A Mathematical Theory of Communication. Bell System Tech J 1948;27:379–423.
24. Kabat EA, Wu TT, Bilofsky H. Unusual distributions of amino acids in complementarity-determining (hypervariable) segments of heavy and light chains of immunoglobulins and their possible roles in specificity of antibody-combining sites. J Biol Chem 1977;252: 6609–16.
25. Sievers F, Wilm A, Dineen D, Gibson TJ, Karplus K, Li W, et al. Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. Mol Syst Biol 2011;7:539.
26. Marsh SG, Albert ED, Bodmer WF, Bontrop RE, Dupont B, Erlich HA, et al. Nomenclature for factors of the HLA system, 2010. Tissue Antigens 2010;75:291–455.
27. Ladewig E, Okamura K, Flynt AS, Westholm JO, Lai EC. Discovery of hundreds of mirtrons in mouse and human small RNA data. Genome Res 2012;22:1634–45.
28. Clark PM, Chitnis N, Johnson BF, Kamoun M, Monos D. Elucidating the targetome of the *HLA-B* intron 4 derived miRNA, miR-6891 and allele specific miRNA isoforms. Hum Immunol 2015;76:S8.
29. Clark PM, Loher P, Quann K, Brody J, Londin ER, Rigoutsos I. Argonaute CLIP-Seq reveals miRNA targetome diversity across tissue types. Sci Rep 2014;4:5947.
30. Helwak A, Kudla G, Dudnakova T, Tollervey D. Mapping the human miRNA interactome by CLASH reveals frequent noncanonical binding. Cell 2013;153:654–65.
31. Li JH, Liu S, Zhou H, Qu LH, Yang JH. starBase v2.0: decoding miRNA-ceRNA, miRNA-ncRNA and protein-RNA interaction networks from large-scale CLIP-Seq data. Nucleic Acids Res 2014;42:D92–7.