The IMGT/HLA database

James Robinson1, Jason A. Halliwell1, Hamish McWilliam2, Rodrigo Lopez2, Peter Parham3 and Steven G. E. Marsh1,4,*  

1HLA Informatics Group, Anthony Nolan Research Institute, Royal Free Hospital, Pond Street, Hampstead, London NW3 2QG, 2External Services, EMBL Outstation - European Bioinformatics Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge CB10 1SD, UK, 3Department of Structural Biology, Stanford University School of Medicine, Sherman Fairchild Research Building, Stanford, CA 94305-5136, USA and 4UCL Cancer Institute, University College London, Royal Free Campus, Pond Street, Hampstead, London NW3 2QG, UK

Received September 14, 2012; Accepted September 19, 2012

ABSTRACT

It is 14 years since the IMGT/HLA database was first released, providing the HLA community with a searchable repository of highly curated HLA sequences. The HLA complex is located within the 6p21.3 region of human chromosome 6 and contains more than 220 genes of diverse function. Of these, 21 genes encode proteins of the immune system that are highly polymorphic. The naming of these HLA genes and alleles and their quality control is the responsibility of the World Health Organization Nomenclature Committee for Factors of the HLA System. Through the work of the HLA Informatics Group and in collaboration with the European Bioinformatics Institute, we are able to provide public access to these data through the website http://www.ebi.ac.uk/imgt/hla/. Regular updates to the website ensure that new and confirmatory sequences are dispersed to the HLA community and the wider research and clinical communities. This article describes the latest updates and additional tools added to the IMGT/HLA project.

INTRODUCTION

The IMGT/HLA database was established to provide a locus-specific database (LSDB) for the allelic sequences of the genes in the HLA system, also known as the human major histocompatibility complex (MHC). The MHC is one of the most complex and polymorphic regions of the human genome, with excess of 220 genes (1). The core genes of interest in the HLA system are 21 highly polymorphic HLA genes, found within the 6p21.3 region of the short arm of human chromosome 6, whose protein products mediate human responses to infectious disease and influence the outcome of cell and organ transplants. Three distinct regions have been identified within the MHC. The class I region is located at the telomeric end of the MHC and encodes the genes for the HLA class I molecules, HLA-A, -B and -C. These are co-dominantly expressed on the cell surface and responsible for presenting intracellularly derived peptides to CD8-positive T cells. The class II region lies at the centromeric end of the MHC and encodes HLA class genes HLA-DRA, -DRB1, -DRB3, -DRB4, -DRB5, -DQA1, -DQB1, -DPA1 and -DPB1. HLA class II expression is limited to cells involved in immune responses, where these molecules present extracellularly derived peptides to CD4-positive T cells. Located between the class I and class II regions lies the class III region where a number of non-HLA genes with immune function are located. With a nomenclature covering more than 50 genes and 8000 alleles, there is an obvious need for a curated LSDB to manage these highly polymorphic variants. The first public release of the IMGT/HLA database was made on the 16 December 1998 (2). Since then the database has been updated every 3 months, in a total of 55 releases, to include all the publicly available sequences officially named by the WHO Nomenclature Committee at the time of release.

The naming of new HLA genes and allele sequences and their quality control is the responsibility of the WHO Nomenclature Committee for Factors of the HLA System, which first met in 1968. This committee meets regularly to discuss the issues of nomenclature and has published 19 major reports (3–21) initially documenting the serologically defined HLA antigens and more recently the genes and alleles defined by nucleotide sequences. The IMGT/HLA database provides the nomenclature committee with the online tools necessary for its task. The dissemination of new allele names and sequences is of paramount importance in the clinical transplant setting, because the variation that distinguishes

*To whom correspondence should be addressed. Tel: +44 20 7284 8321; Fax: +44 20 7284 8331; Email: steven.marsh@ucl.ac.uk

© The Author(s) 2012. Published by Oxford University Press. This is an Open Access article distributed under the terms of the Creative Commons Attribution License (http://creativecommons.org/licenses/by-nc/3.0/), which permits non-commercial reuse, distribution, and reproduction in any medium, provided the original work is properly cited. For commercial re-use, please contact journals.permissions@oup.com.
HLA alleles can have a critical impact on the outcome of a haematopoietic stem cell transplant (22,23). The identification, verification and publication of the sequences of these variants through a centralized resource are necessary for accurate identification of HLA alleles in a clinical setting. Sequencing of HLA alleles began in the late 1970’s, predominantly using protein-based techniques to determine the sequences of HLA class I allotypes. The first complete HLA class I allotype sequence, B7.2, now known as B*07:02:01, was published in 1979 (24). The first HLA class II allele, DRA*01:01, was defined by protein sequencing and later in 1982 by DNA sequencing (25–27). The first HLA DNA sequences or alleles were named by the WHO Nomenclature Committee for Factors of the HLA System (10) in 1987. At that time, 12 class I alleles and 9 class II alleles were named: in the first 8 months of 2012, the WHO Nomenclature Committee was able to assign names to 1163 alleles (Figure 1).

**IMGT/HLA DATA SOURCES**

The IMGT/HLA database receives submissions from laboratories across the world. These submissions are curated and analysed, and if they meet the strict requirements, an official allele designation is assigned. The IMGT/HLA database is the official repository for the WHO Nomenclature Committee for factors of the HLA system and is the only way of receiving an official allele designation for a sequence. The sequence is then incorporated into the next 3-monthly release of the database. Since its release in December 1998, the database has received over 14 000 submissions. These submissions come from a variety of sources; the majority are from laboratories involved in clinical HLA typing, for hospitals or donor registries, or commercial organizations performing contract HLA typing for large haematopoietic stem cell donor registries. Further data have been submitted following large-scale genome sequencing projects (1,28). All submissions must meet strict acceptance criteria before the sequence receives an official designation. These minimum standards cover the methodologies used to define the sequence, the length of sequence submitted and the source of the sequence; the full list of the minimum criteria can be seen at http://www.ebi.ac.uk/imgt/hla/subs/submit.html. Around 3% of the submissions received fail to meet these criteria and are rejected. In addition, all the submissions received by the IMGT/HLA database are also available from the International Nucleotide Sequence Database Collaboration (INSDC) (29). The INSDC consists of DNA DataBank of Japan (Japan), GenBank (USA) and the EMBL-European Nucleotide Archive (ENA) (UK) (30–32). The ENA entries also contain database cross-references to the IMGT/HLA entries. The cross-references to the IMGT/HLA database are also included in ENSEMBL (33) and vertebrate genome annotation (VEGA) entries (34).

---

**Figure 1.** The number of HLA alleles named each year and included in the IMGT/HLA database. The recent surge in the number of submissions received by the database is clearly shown.
TOOLS AVAILABLE AT IMGT/HLA

The IMGT/HLA database provides a diversity of tools for the analysis of HLA sequences. Some of these tools were custom written for the IMGT/HLA database, and others were incorporated from the existing set of tools provided on the European Bioinformatics Institute’s (EBI) website (35,36). The website (Figure 2) includes tools for producing user-defined sequence alignments at the protein, cDNA and gDNA level. The user is also able to perform queries for particular HLA alleles; the output provides access to detailed information on any HLA allele, including information on the ethnic origin of the source, database cross-references and seminal publications. This information is also available through integration with the Sequence Retrieval System (SRS) service at EBI (37).

Tools have also been developed to support the laboratories that sequence HLA. The use of sequence-based typing (SBT) as a method for defining the HLA type is well documented (38,39); most SBT typing strategies currently employed use the exon 2 and exon 3 sequences for HLA class I analysis and exon 2 alone for HLA class II analysis. Because of the heterozygous nature of the SBT analysis, the combinations of many pairs of alleles may give an ambiguous typing result; currently, there are over 60,000 recognized ambiguous combinations. The IMGT/HLA maintains and regularly updates a listing of these ambiguous allele combinations. The document also includes a list of all alleles that are identical over exons 2+3 for HLA class I and exon 2 for HLA class II.

Where possible, sequence data, both nucleotide and protein, from the IMGT/HLA database is incorporated into the EBI’s suite of search tools including FASTA (40) and BLAST (41) and downloadable from the EBI’s
LATEST DEVELOPMENTS

In 2012, the IMGT/HLA database added an Extensible Markup Language (XML) export to the data formats available. XML is a simple but flexible language that defines a set of rules for encoding documents in a format that is both human and machine readable. Designed to meet the challenges of large-scale electronic publishing, XML is playing an increasingly important role in the exchange of scientific data. The data format has been developed in a collaborative project between the HLA Informatics Group of the Anthony Nolan Research Institute and the Bioinformatics Department of the National Marrow Donor Programme (NMDP). The NMDP Bioinformatics group has previous success in developing an XML format for electronically communicating HLA typing data, the Histoimmunogenetics Markup Language file format (42). This experience facilitated the collaboration to develop a similar project for publishing the data contained within each release of the IMGT/HLA database. The new format combines the data present in the multiple files of each quarterly IMGT/HLA release into a single file. The IMGT/HLA database provides an FTP site for the retrieval of sequences in a number of pre-formatted files. The sequences are provided as FASTA, PIR and MSF formats, as well as an archive of the sequence alignments and an ENA flat file like formatted copy of the database. The NMDP Bioinformatics Department has also developed a suite of tools for importing data into different database schema, both open source and proprietary, allowing incorporation into different laboratory systems (Figure 3). Additional XML exports are being developed for other sections of the IMGT/HLA database. Further developmental work on a suite of tools for integrating the XML into laboratory systems used by HLA-typing laboratories is underway.

HLA matching is a critical factor when considering potential donors for patients receiving allogeneic transplants for haematological disorders (22,23). The most recent development on the IMGT/HLA website is an online tool to implement the T-cell epitope matching algorithm described by Zino et al. (43–45) and updated by Fleischhauer and Shaw (46). This algorithm classifies the HLA-DPB1 alleles into a number of groups based on functional studies and protein motifs. Predictive analysis of the HLA-DPB1 mismatches between patient and donor based on T-cell epitope (TCE) groups has the potential to distinguish between mismatches that are tolerated (permissive) from those that increase the risks of poor clinical outcome (non-permissive). This tool allows the user to enter the HLA-DPB1 of a prospective patient and donor pair and view the predicted TCEs and resulting prediction of the effect of mismatching when selecting appropriate donors for HSCT recipients. Any allele that does not have a TCE group ‘protein’ is analysed for a motif match to particular protein motifs of those alleles with known TCE group. If the tool needs to predict the TCE group for an allele, then a warning is issued within the output to the user, to ensure that the lack of functional studies is acknowledged. The implementation of an easy to use online tool makes it simple for all those staff involved with selecting donors for transplantation to factor in DPB1 mismatches into their own search algorithms and procedures.
FUTURE DEVELOPMENTS

A major challenge for the database is to keep up with the increasing number of allele sequences that are being submitted. In recent years, the number of sequences in the database increased on average by 29% each year. The database must develop new tools for the visualization of sequences while maintaining the high standards set in the presentation and quality of the HLA sequences and nomenclature to the research community. The database aims to continually develop new tools and refine existing tools to meet this challenge.

CONCLUSIONS

The IMGT/HLA database provides a centralized resource for everybody interested, clinically or scientifically, in the HLA system. The database and accompanying tools allow the study of HLA alleles from a single site on the World Wide Web. It aids in the management and development of HLA nomenclature, providing a continuing and updated resource for the WHO Nomenclature Committee. The challenges for the database are to keep up with this increase in submitted sequences, keep pace with the increasing difficulties in performing analyses on the larger datasets and develop new tools for the visualization of the sequences while maintaining the high standards set in the presentation and quality of the HLA sequences and nomenclature to the research community.

ACKNOWLEDGEMENTS

The authors thank Angie Dahl of the Be The Match Foundation, for her work in securing ongoing funding for the database. They also thank all the individuals and organizations that support the work financially. The authors thank Martin Maiers, Jane Pollack, Adrienne Walts, Joel Schneider, Read Fritsch, Anthony Barber and John Freeman of the Bioinformatics Department of the National Marrow Donor Program for their assistance in developing the XML format.

REFERENCES


FUNDING

Histogenetics; One Lambda Inc.; Conexio; Abbott Molecular Laboratories Inc.; European Federation for Immunogenetics; Gen-Probe; LabCorp; Life Technologies; Olersup SSP; 454 Sequencing; American Society for Histocompatibility and Immunogenetics; Anthony Nolan; Asia-Pacific Histocompatibility and Immunogenetics Association; BAG Healthcare; Be the Match Foundation; DKMS, Inno-train Diagnostik GMBH; National Marrow Donor Program; Rose and Zentrum Knochenmarkspender-Register Deutschland; Imperial Cancer Research Fund (now Cancer Research UK) and a EU Biotech grant [BIO4CT960037]. Funding for open access charge: The publication costs will be met by the Anthony Nolan Research Institute.

Conflict of interest statement. None declared.


**APPENDIX**

**ACCESS AND CONTACT**

- IMGT/HLA Homepage: [http://www.ebi.ac.uk/imgt/hla/](http://www.ebi.ac.uk/imgt/hla/)

Contact: hla@alleles.org