Human Immunology 77 (2016) 233–237

Contents lists available at ScienceDirect

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

# The IPD-IMGT/HLA Database – New developments in reporting HLA variation

James Robinson [a,b], Anup R. Soormally [a], James D. Hayhurst [a], Steven G.E. Marsh [a,b,*]

[a] Anthony Nolan Research Institute, Royal Free Hospital, Pond Street, Hampstead, London NW3 2QG, UK
[b] UCL Cancer Institute, University College London, Royal Free Campus, Pond Street, Hampstead, London NW3 2QG, UK

ARTICLE INFO

*Article history:*
Received 26 June 2015
Revised 18 December 2015
Accepted 23 January 2016
Available online 27 January 2016

*Keywords:*
Immunogenetics
Database
Polymorphism
Variation
Sequence
Allele
MHC
HLA

ABSTRACT

IPD-IMGT/HLA is a constituent of the Immuno Polymorphism Database (IPD), which was developed to provide a centralised system for the study of polymorphism in genes of the immune system. The IPD project works with specialist groups of nomenclature committees who provide and curate individual sections before they are submitted to IPD for online publication. The primary database within the IPD project is the IPD-IMGT/HLA Database, which provides a locus-specific database for the hyper-polymorphic allele sequences of the genes in the HLA system, also known as the human Major Histocompatibility Complex. The IPD-IMGT/HLA Database was first released over 17 years ago, building on the work of the WHO Nomenclature Committee for Factors of the HLA system that was initiated in 1968. The IPD-IMGT/HLA Database enhanced this work by providing the HLA community with an online, searchable repository of highly curated HLA sequences. Many of the genes encode proteins of the immune system and are hyper polymorphic, with some genes currently having over 4000 known allelic variants. Through the work of the HLA Informatics Group and in collaboration with the European Bioinformatics Institute we are able to provide public access to this data through the website, http://www.ebi.ac.uk/ipd/imgt/hla.


## 1. Introduction

The Immuno Polymorphism Database (IPD) is a set of specialist databases related to the study of polymorphic genes in the immune system. The primary database within the IPD project is the IPD-IMGT/HLA Database, which provides a locus-specific database for the hyper polymorphic allelic sequences of the genes in the HLA system, also known as the human Major Histocompatibility Complex (MHC). Within the MHC is the HLA complex, which is a region of ~4 mb on the short arm of chromosome 6 (6p21) containing >220 genes, many of which contribute to immune defence against infection [1]. The core genes of interest in the HLA system are 21 highly polymorphic HLA genes, whose protein products mediate human responses to infectious disease and influence the outcome of cell and organ transplants. The MHC is one of the most complex and polymorphic regions of the human genome [1]. These highly complex genes encode the hyper-polymorphic HLA class I (*HLA-A*, *-B* and *-C*) and class II molecules (*HLA-DP*, *-DQ*, and *-DR*) which diversify immunity within human populations and form the major genetic barriers to clinical transplantation of cells, tissues and organs. The ontology of the HLA alleles, has been continuously developed since 1968 [2], when the initial work in this area was initiated. Thirty years later, the IPD-IMGT/HLA Database was established to serve this purpose [3]. The IPD-IMGT/HLA Database continued the work of an initial phase of development when HLA class I and II sequences were published by the WHO Nomenclature Committee [4–8]. In the 17 years since the IPD-IMGT/HLA Database was first released [9], over 21,000 submissions have been accepted for inclusion in the database, leading to the definition of 14,000 HLA alleles by the end of November 2015. The first public release of the IPD-IMGT/HLA Database was made in December 1998. Since its inception, the database has been updated every three months, with over 66 major releases, to include all the publicly available sequences officially named by the WHO Nomenclature Committee.

A driving force behind the development and continued success of the IPD-IMGT/HLA Database is its use by the transplantation community. The HLA molecules, whose polymorphic variants are stored in the database, play a key role in transplantation, with the success of kidney and bone marrow transplantation correlated with the degree to which donors and recipient are HLA matched. It

* Corresponding author at: Anthony Nolan Research Institute, Royal Free Hospital, Pond Street, Hampstead, London NW3 2QG, UK.
*E-mail address:* steven.marsh@ucl.ac.uk (S.G.E. Marsh).

http://dx.doi.org/10.1016/j.humimm.2016.01.020
0198-8859/

has been shown that HLA matching is recognised as a critical determinant of outcome for patients receiving unrelated donor hematopoietic stem cells for haematological disorders [10]. This has led to progressive improvements in the level of resolution achieved by HLA class I and II typing methods. HLA alleles are now routinely characterised using molecular typing methods and, moving forward, next-generation sequencing techniques [11–13]. The typing of HLA focuses on distinguishing differences at both the synonymous and non-synonymous level for the nucleotide sequences encoding the protein domains of HLA class I and II, which bind peptides and interact with variable lymphocyte receptors. The consequence of these improvements has required the development of a nucleotide sequence database that is both accurate and comprehensive.

Details of the IPD-IMGT/HLA Database have been published previously and a more detailed description of the database its structure and content are available elsewhere [14], here we report some of the new additions to be the database.

## 2. Material and methods

### 2.1. IPD data sources

IPD 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 IPD-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. Sequence submissions come from a variety of sources; the majority are from laboratories involved in clinical HLA typing for hospitals and donor registries, or commercial organisations performing contract HLA typing for large haematopoietic stem cell donor registries. Further data has been submitted following large-scale genome sequencing projects [1,15]. For all projects the 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 online. Within the IPD-IMGT/HLA Database, around 3% of the submissions received failed to meet these criteria and have been rejected. In addition all the submissions received by the IPD are also available from the International Nucleotide Sequence Database Collaboration (INSDC) [16]. The INSDC consists of DNA DataBank of Japan (DDBJ) (Japan), GenBank (USA) and the EMBL-European Nucleotide Archive (ENA) (UK) [17–19]. The ENA entries also contain database cross-references to the IPD entries. Additionally, cross-references to the IPD-IMGT/HLA Database are also included in ENSEMBL [20] and VEGA entries [21]. The use of a generalist sequence database for storing these sequences has been shown to be problematic due to lack of expert curation, failure to correct sequence errors and inaccurate assignment of allele names. With the success of a transplant potentially being effected by a single SNP, it is vital that any repository storing HLA sequences has high standards of both quality control and curation, this need is met through the expert curation within the IPD-IMGT/HLA Database. Sequences that do not meet these standards are not accepted, although they may still be available in the generalist databases of the INSDC.

## 3. Results

### 3.1. Tools available at IPD-IMGT/HLA

The project provides a large number of tools for the analysis of HLA sequences. These tools are either custom written for the database or are incorporated into existing tools on the European Bioinformatics Institute (EBI) website [22,23], see Table 1. We are constantly adding new tools and resources to the database, and list here a number of recent developments.

### 3.2. DPB1 T-Cell Epitope Algorithms

The IPD-IMGT/HLA Database also collaborates with clinicians to provide web-based versions of published algorithms, which have a clinical impact on transplant outcome in unrelated haematopoietic stem cell transplantation (HSCT). Recent data has suggested that certain HLA mismatches may be permissive and do not result in a poor clinical outcome, while others are non-permissive (do result in a poor clinical outcome) [24]. The classification of HLA-DPB1 mismatches based on T-cell-epitope (TCE) groups has been shown to identify permissive mismatches and non-permissive mismatches for HLA-DPB1 after unrelated-donor HSCT. With the strong clinical data showing a survival disadvantage in patients who receive a transplant from a non-permissive HLA-DPB1 TCE mismatched donor (defined on the basis of functional data) matching of DPB1 TCE groups can be routinely included in the donor selection process [24–27]. The IPD-IMGT/HLA Database provides an online, freely available tool, which was developed to help those selecting donors to predict the immunogenicity of any given patient–donor HLA-DPB1 types [28]. The aim of the tool is to provide a web interface to predict HLA-DPB1 immunogenicity based on the published algorithms. The predicted immunogenicity of the HLA-DPB1 matching for each patient–donor pair is provided. The tool also allows for labelling the patient and donors with user-defined identification numbers. The results can therefore be printed and stored. The web tool is hosted on the IPD-IMGT/HLA Database website and can be accessed at http://www.ebi.ac.uk/ipd/IPD-IMGT/HLA/dpb.html. The underlying algorithm for this tool has recently been updated to incorporate the work of Crivello et al. [29] by using a predictive functional distance to assign TCE

**Table 1**

| Tools | Description |
|---|---|
| Sequence alignments | Access to alignment tool, which filters pre-generated alignments to the users' specification. Provides alignments at the protein, cDNA and gDNA level |
| Allele queries | Access to detailed information on any allele, including information on database cross-references and seminal publications |
| Donor selection tools | Access to tools to aid in the selection of unrelated donors, this includes the DPB1 T-Cell Epitope Algorithms |
| Downloads | Access to a FTP directory containing all the data from the current and previous releases of the IPD-IMGT/HLA Database in a variety of commonly used formats like FASTA, MSF and PIR |
| Allele status checker | Allows users to determine which alleles contain only partial sequences and which alleles have not been independently confirmed |
| Cell queries | Search tool, which allows complex queries on source cells |
| Deleted alleles | List of deleted allele names, with reason for deletion |
| Ethnic origins | Search tool, which displays the known ethnic origins for HLA Alleles |
| Polymorphism search tool | Search tool to identify polymorphic positions in HLA sequences |
| Probe and primer search | Search tool for generating probe and primer hit tables |
| Sequence submission tool | Online submission of sequences to the WHO Nomenclature Committee for Factors of the HLA System |
| Sequence similarity search tools | Integration into EBI's suite of search tools including FASTA [39] and BLAST [40] |