Streamlining submission to the IPD-IMGT/HLA Database: The Sequence Feature Annotation Tool

Dominic J Barker[1, 2], Richard Natarajan[1], James Robinson[1,2], Steven GE Marsh[1,2]

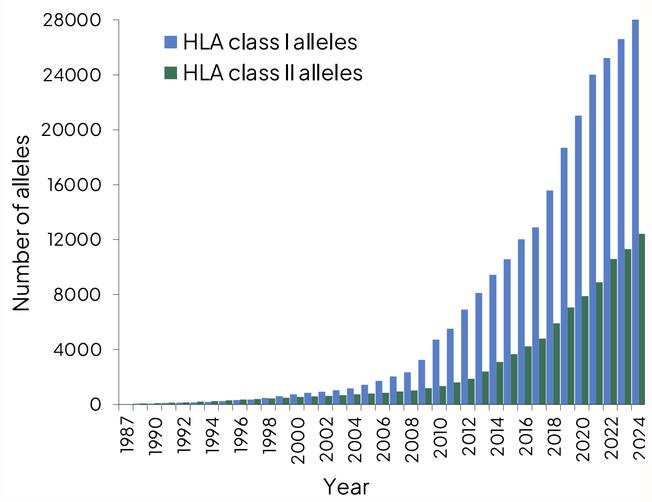
1. Anthony Nolan Research Institute, Royal Free Hospital, Pond Street, London, NW3 2QG, UK
2. UCL Cancer Institute, University College London (UCL), Royal Free Campus, Pond Street, London, NW3 2QG, UK

IPD-IMGT/HLA Database

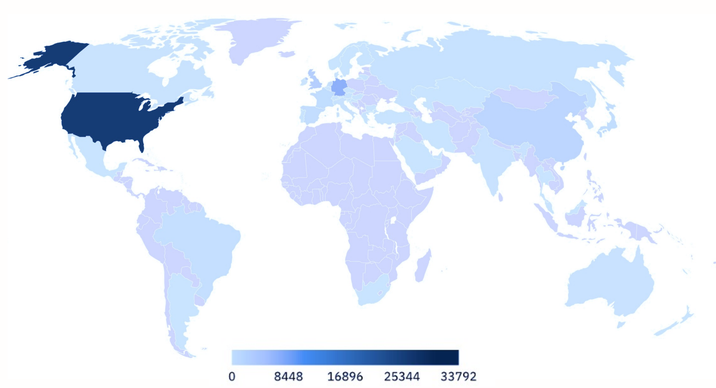
The **IPD-IMGT/HLA Database** is the global reference database for polymorphisms in the HLA system. It contains all recognized allelic variants officially recognized by the **WHO Nomenclature Committee for Factors of the HLA System**.

Since its initial release in **1998** with **932** alleles the database has continued to expand now reaching **41,433** distinct HLA sequences[1].

Over the last 26 years the database has continued to develop new tools and ways of representing this data to meet the needs of the HLA community[1,2].



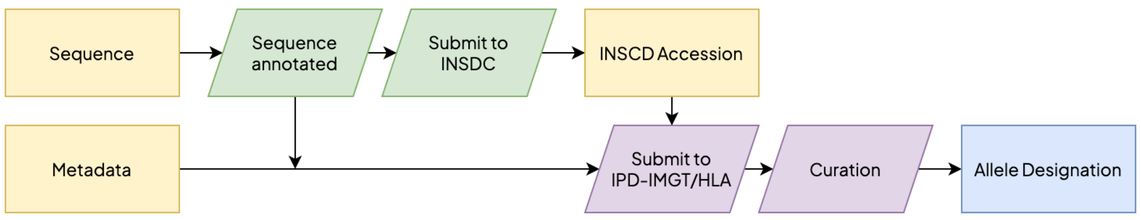
Submissions to IPD-IMGT/HLA



The IPD-IMGT/HLA Database receives submissions from around the world, both through direct submissions from approved bulk submitters and standard submissions using our online **HLA Submission Tool**. These submissions are curated, analyzed and sent to the WHO Nomenclature Committee for new allele assignment.

The HLA Submission Tool ensures data is received in a standardized format and that all required data is submitted. In the last 26 years **62,612** submissions have processed by the IPD-IMGT/HLA Database.

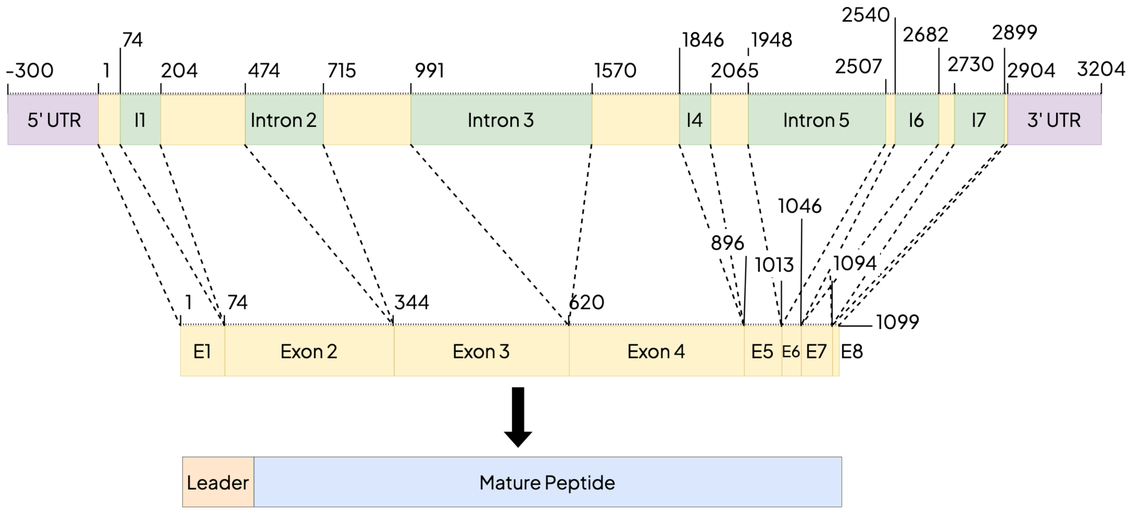
Submitting to the IPD-IMGT/HLA Database has several steps. First a sequence must be deposited in a public databank belonging to the **International Nucleotide Sequence Database Collaboration (INSDC)** such as GenBank. This ensures the longevity and traceability of sequences in the database. Once accessioned by INSDC the **sequence**, **metadata** and **INSDC accession** can be submitted to IPD-IMGT/HLA.



This process can be both challenging and time-consuming. To simplify and accelerate it, we developed the **Sequence Feature Annotation Tool (SFAT)** to assist with submissions to the INSDC database and IPD-IMGT/HLA. SFAT generates sequence annotations in multiple formats, streamlining submissions to IPD-IMGT/HLA.

Sequence Annotation

Sequences are annotated to mark the positions of **exons** and **introns** in a **genomic** sequence. These coordinates can then be used to extract the **coding** exonic sequence. This **coding** sequence can then be translated into a **protein** and used for assigning correct nomenclature.

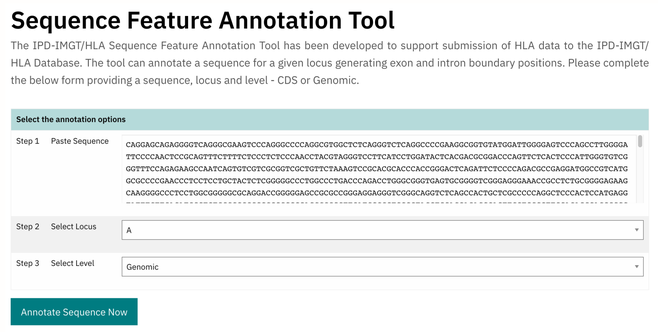


Manual annotation is extremely difficult, which is why the HLA Submission Tool has the option for **automated feature annotation**. However automated annotation previously only existed within the HLA Submission Tool and could not be used to submit to GenBank. To help streamline submissions to GenBank and the IPD-IMGT/HLA Database we have developed the **Sequence Feature Annotation Tool**.

References

1. Robinson J, Barker DJ, Marsh SGE. 25 years of the IPD-IMGT/HLA Database. HLA (2024) 103(6):e15549
2. Barker DJ, Maccari G, Georgiou X, *et al*. The IPD-IMGT/HLA Database. Nucleic Acids Research (2023) 51:D1053-60

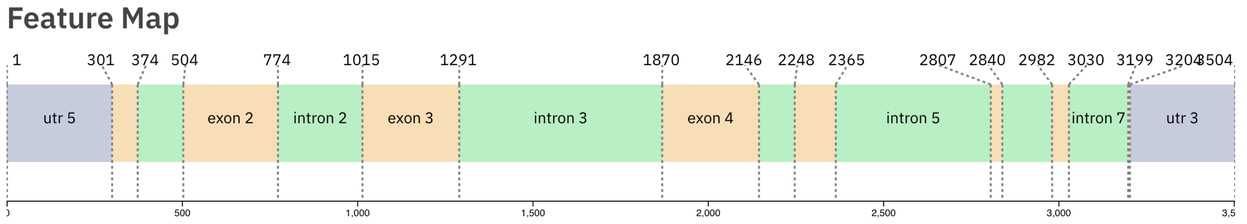
Sequence Feature Annotation Tool



The **Sequence Feature Annotation Tool (SFAT) quickly** and **accurately** annotates the exon intron boundaries of any HLA sequence. SFAT is available on the EBI website. To use the tool just provide a **sequence**, select a **locus** and **resolution** (protein, coding or genomic) and click annotate.

Results are returned as a graphic, table and in standard formats suitable for submitting to IPD-IMGT/HLA, EMBL-ENA or GenBank. These results include extracted coding sequence and translated protein, where appropriate. In **validation** SFAT was able to correctly annotate **99.7%** of submissions made to the IPD-IMGT/HLA Database. Sequences **incorrectly** annotated by SFAT contain **splice site polymorphisms** causing aberrant splicing which cannot be predicted from the reference.

Feature Map



SFAT has been developed as an **Application Program Interface (API)** which allows it to be used **manually** via our **web-interface** or **integrated** into existing **bioinformatics pipelines** for **automated** annotation. SFAT has been incorporated into our new **HLA Submission Tool API**.

New HLA Submission Tool API

The IPD-IMGT/HLA Database team have developed a new **HLA Submission Tool API** to facilitate submissions to the database. This API will allow approved submitters to **submit** large numbers of sequences to the IPD-IMGT/HLA Database **programmatically**.

The HLA Submission Tool API is available on the EBI website. To become an approved API submitter please contact: ipdsubs@anthonynolan.org.

The HLA Submission Tool API provides upfront **validation** for submitted data, **reducing** time for curation. In addition to supporting programmatic submissions the API will power a new web-based user interface for manual submissions including **validation**, **automated** feature **annotation** using SFAT and more **intuitive data entry**.

The API has now successfully made **347** submissions in **September 2024**. The web-based user interface is scheduled for release in **December 2024**. In addition to an improved user-interface the new HLA Submission Tool is planned to **broker** submissions to the **INSDC** meaning sequences don't have to first be submitted to INSDC prior to submitting to IPD-IMGT/HLA.

Conclusion

- **The IPD-IMGT/HLA Database continues to be an important resources for the HLA community.**
- **Its continued relevance is reliant on submissions made to the database.**
- **To facilitate submissions to IPD-IMGT/HLA, EMBL-ENA or GenBank we have developed the Sequence Feature Annotation Tool (SFAT)**
- **Automated feature annotation supports submission of data through the new HLA Submission Tool API**
- **A new web-based submission tool is under development streamlining the process for non-bioinformaticians**
- **In future this will include brokering with INSDC database so submissions can be made directly to IPD and deposited in ENA.**