# IMGT® biocuration of IG and TR in IMGT/LIGM-DB and IMGT/GENE-DB

Géraldine Folch*, Joumana Michaloud*, Marine Peralta, Mélanie Arrivet, Imène Chentli, Mélissa Cambon, Pascal Bento, Souphatta Sasorith, Typhaine Paysan-Lafosse, Patrice Duroux, Véronique Giudicelli, Sofia Kossida* and Marie-Paule Lefranc*

*Equal contribution

Université Montpellier and CNRS, Laboratoire d'ImmunoGénétique Moléculaire (LIGM), Institut de Génétique Humaine (IGH), UPR CNRS 1142, Montpellier (France)

IMGT® Information system®
http://www.imgt.org

IMGT®, the international ImMunoGeneTics information system, http://www.imgt.org, is the global reference in immunogenetics and immunoinformatics [1]. By managing the extreme diversity and complexity of the antigen receptors of the adaptive immune response, the immunoglobulins (IG) or antibodies and the T cell receptors (TR) [2,3] ($2.10^{12}$ different specificities per individual), IMGT® is at the origin of immunoinformatics, a science at the interface between immunogenetics and bioinformatics [4]. IMGT® is based on the concepts of IMGT-ONTOLOGY [5] and these concepts are used for expert annotation and standardized knowledge in IMGT/LIGM-DB, the IMGT® database of IG and TR nucleotide sequences from human and other vertebrate species and in IMGT/GENE-DB, the IMGT® gene and allele database. The IMGT/LIGM-DB biocuration pipeline of IG and TR sequences includes IMGT/LIGMotif, for the analysis of large genomic DNA sequences, and IMGT/Automat, for the automatic annotation of rearranged cDNA sequences. Analysis results are checked for consistency, both manually and by using IMGT® tools (IMGT/NtiToVald, IMGT/V-QUEST, IMGT/BLAST, etc.). The annotated sequences are integrated in IMGT/LIGM-DB and include the sequence identification (IMGT® keywords), the gene and allele classification (IMGT® nomenclature), the constitutive and specific motif description (IMGT® labels in capital letters, no plural), the translation of the coding regions (IMGT® unique numbering) [4,5]. For genomic IMGT/LIGM-DB sequences containing either an IG or TR variable (V), diversity (D) or joining (J) gene in germline configuration or a constant (C) gene, the gene and allele information is entered in IMGT/GENE-DB. In parallel, the IMGT® Repertoire is updated (Locus representations, Gene tables and Protein displays (for new genes), Alignments of alleles (for new and/or confirmatory alleles)) and the IMGT® reference directory [1,4] is completed (sequences used for gene and allele comparison and assignment in IMGT® tools (IMGT/V-QUEST, IMGT/HighV-QUEST for next generation sequencing (NGS), IMGT/DomainGapAlign) and databases (IMGT/2Dstructure-DB, IMGT/3Dstructure-DB). An IMGT/GENE-DB entry also provides information on the rearranged cDNA and gDNA entries (with links to IMGT/LIGM-DB) and on the three-dimensional structures (with links to IMGT/3Dstructure-DB). IMGT/GENE-DB is the official repository of IG and TR genes and alleles. IMGT® gene names were approved by HGNC and endorsed by WHO-IUIS, the World Health Organization (WHO)-International Union of Immunological Societies (IUIS) Nomenclature Subcommittee for IG and TR. Reciprocal links exist between IMGT/GENE-DB and HGNC, NCBI and Vega. The definition of antibodies published by the WHO International Nonproprietary Name (INN) Programme is based on the IMGT® concepts [6], and allows easy retrieval via IMGT/mAb-DB query [1,4]. The IMGT® standardized annotation has allowed to bridge the gaps for IG or antibodies and TR between fundamental and medical research, veterinary research, repertoire analysis, biotechnology related to antibody engineering, diagnostics and therapeutical approaches.

[1] Lefranc M-P et al. Nucleic Acids Res 43:413-422 (2015) PMID: 25378316,
[2] Lefranc M-P, Lefranc G. The Immunoglobulin FactsBook (2001),
[3] Lefranc M-P, Lefranc G. The T cell receptor FactsBook (2001),
[4] Lefranc M-P. Front Immunol 5:22 (2014) PMID: 24600447,
[5] Giudicelli V, Lefranc, M-P. Front Genet 3:79 (2012) PMID: 22654892,
[6] Lefranc M-P. mAbs 3(1):1-2 (2011) PMID: 21099347.

## IMGT® Expert Biocuration Pipeline

### IMGT Tools

Internally developped, proprietary IMGT® research tools:

**IMGT/LIGMotif**
Annotation of **genomic sequences** of immunoglobulin (IG) and T cell receptor (TR) loci.

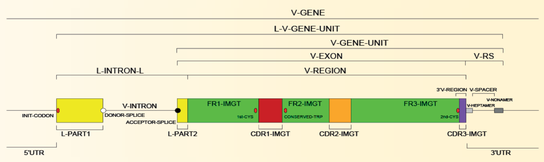
**IMGT/Automat**
Annotation of **cDNA sequences** of immunoglobulin (IG) and T cell receptor (TR) loci.

**IMGT/NtiToVald**
Automatical control of data for **consistency** and **relevance**

*optimization* — *efficiency/accuracy* — **IMGT expert biocurators** — *optimization* — *standardization*

### IMGT-ONTOLOGY & Prototypes

1. IDENTIFICATION
2. DESCRIPTION
3. CLASSIFICATION
4. NUMEROTATION
5. LOCALIZATION
6. ORIENTATION
7. OBTENTION




Left: IMGT-ONTOLOGY axioms and concepts, bridge the gap between genes, sequences and three-dimensional structures. The concepts include the IMGT® standardized keywords (identification), labels (description) and nomenclature (classification), as well as IMGT unique numbering and IMGT Colliers de Perles (numerotation). IMGT-ONTOLOGY includes a controlled vocabulary and annotation rules which are indispensable to ensure accuracy, consistency and coherence in IMGT®.
Right: IMGT standardized labels for the description of prototypic V-GENEs.

## IMGT/LIGM-DB

**177 115 sequences**
**351 species**

**IMGT/LIGM-DB** includes all germline (non-rearranged) and rearranged IG and TR genomic DNA and complementary DNA sequences published in generalist databases. IMGT/LIGM-DB allows searches from the Web interface according to biological and immunogenetic criteria. For a given entry, nine types of display are available including the IMGT flat file, the translation of the coding regions and the analysis by the IMGT/V-QUEST tool.
The annotations hugely enhance the quality and the accuracy of the distributed detailed information.

## IMGT/GENE-DB

**3 927 genes**
**5630 alleles**
**24 species**

**IMGT/GENE-DB Query Page** allows the search of IG/TR genes according to IMGT-ONTOLOGY's seven axioms.
**IMGT/GENE-DB entry** displays accurate gene data related to genome (gene localization), allelic polymorphisms (number of alleles, IMGT reference sequences, functionality, etc.) gene expression (known cDNAs) and proteins structures (IMGT Colliers de Perles, IMGT/3Dstructure-DB). It provides internal links to the IMGT sequence databases and the IMGT Web resources as well as external links to genome and generalist sequence databases.

## Web Resources

### 1 IMGT flat file

1 **IDENTIFICATION: Keywords**
genomic DNA=MoleculeType
germline=ConfigurationType
regular=StructureType
functional=Functionality
Homo sapiens=Taxon
Ig-Heavy-Mu=ChainType
variable=GeneType

2 **DESCRIPTION: Labels**
V-GENE=Entity
V-REGION=CoreRegion
FR1-IMGT=SubRegion
...

3 **CLASSIFICATION: Nomenclature**
IGHV=Group
IGHV3=Subgroup
IGHV3-66=Gene
IGHV3-66*04=Allele

4 **NUMEROTATION**
[8.7.2]=V-REGION CDR lengths
1 to 26, AA 10 is missing=AA IMGT numbering

```
ID   X70208; SV 1; linear; gDNA; STD; HUM; 830 BP.                          1
XX
AC   X70208; X68844;
XX
DT   15-MAY-1995 (Rel. 199520-1, arrived in LIGM-DB)
DT   31-MAR-2015 (Rel. 201514-2, Last updated, Version 19)
XX
DE   H.sapiens DNA for IgM heavy chain (MTGLa)                               1
XX
KW   antigen receptor; Immunoglobulin superfamily (IgSF); immunoglobulin (IG);
KW   IG-Heavy; IG-Heavy-Mu; variable; lymphoma; IMGT reference sequence;
KW   regular; gDNA; germline; functional; V-gene.
-------------------------------------------------------------------------------
CC   IMGT/LIGM-DB annotation level: by annotators
XX
FH   Key                  Location/Qualifiers
FH
FT   V-GENE           2   1..830
FT                        /IMGT_allele="IGHV3-66*04"
FT                        /IMGT_gene="IGHV3-66"                              3
FT                        /db_xref="taxon:9606"
FT                        /mol_type="genomic DNA"
FT                        /organism="Homo sapiens"
FT   5'UTR                1..291
FT   L-PART1              292..337
FT                        /translation="MEFGLSWVFLVAILK"
FT   INIT-CODON           292..294
FT   DONOR-SPLICE         337..339
FT   V-INTRON             338..438
FT   ACCEPTOR-SPLICE      436..440
FT   V-EXON               439..742
FT                        /codon_start=3
FT                        /translation="VQCEVQLVESGGGLVQPGGSLRLSCAASGFTVSSNYMSWVR
FT                        QAPGKGLEWVSVIYSGGSTYYADSVKGRFTISRDNSKNTLYLQMNSLRAEDTAVY
FT                        YCAR"
FT   L-PART2              439..449
FT                        /codon_start=3
FT                        /translation="VQC"
FT   V-REGION             450..742
FT                        /translation="EVQLVESGGGLVQPGGSLRLSCAASGFTVSSNYMSWVRQAP
FT                        GKGLEWVSVIYSGGSTYYADSVKGRFTISRDNSKNTLYLQMNSLRAEDTAVYYCA
FT                        R"
FT                        /CDR_length="[8.7.2]"                              4
FT                        /IMGT_allele="IGHV3-66*04"
FT                        /IMGT_gene="IGHV3-66"
FT   FR1-IMGT             450..524
FT                        /translation="EVQLVESGGGLVQPGGSLRLSCAAS"
FT                        /AA_IMGT="1 to 26, AA 10 missing"                  4
-------------------------------------------------------------------------------
FT   3'UTR                741..830
XX
SQ   Sequence 830 BP; 203 A; 202 C; 237 G; 188 T; 0 other;
     cctaaatgaa tatcaaggca caatcacata atataaaatt atattttcct gaatgatagg        60
     aatactacca atctcccccc aggagacttt catctgcact gggcccggcc ctctcctcag       120
     ctgtgtatta ctgtgcgaga cacacagtga gggaaggcca ttgtgcgccc agacacaaac       780
     cctcccctgca ggaacgctgg ggaaatcagc ggcagggggc gctcaggagc                 830
//
```

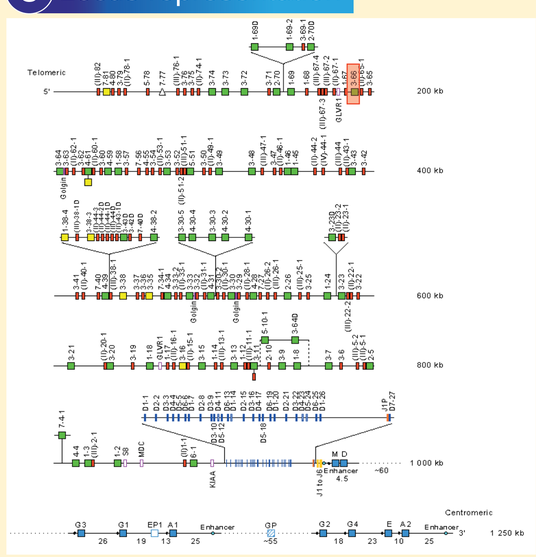
*Extracts from the IMGT Flat File X70208 with the IGHV3-66 gene and the IGHV3-66*04 allele.*

### 2 Gene table

| IMGT/LIGM-DB reference sequences | | | | | | | | | | | IMGT/LIGM-DB sequences from the literature | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| IMGT subgroup | IMGT gene name | IMGT allele name | Fct | Chromosomal localization | R | T | Pr | Positions in the locus | Clone names | Accession numbers | Positions in the sequence (from INIT-CODON to the end of V-REGION) | Secondary accession numbers | Clone names | Accession numbers | Positions in the sequence (from INIT-CODON to the end of V-REGION) |
| IGHV3 | IGHV3-66 | IGHV3-66*01 | F | 14q32.33 | + | + | + | | 8-1B MAP | X92218 [5] | | | | | |
| | | IGHV3-66*02 | F | 14q32.33 | + | | | | YAC-5 MAP | Z27504 [12] | | | | | |
| | | IGHV3-66*03 | F | 14q32.33 | | | | 155001-165000 | | AB019437 [61] | 153482-158510 | | DP-86 | Z27455 [65] | |
| | | | | | | | | | | | | | DA-9 | Z29984 [12] | |
| | | IGHV3-66*04 | F | 14q32.33 | | | | | | X70208 [50] | | | | | |

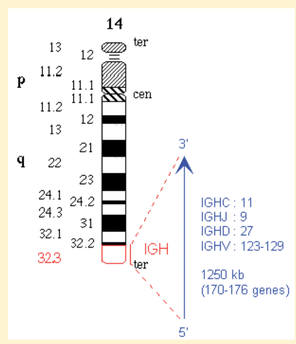
*Extract from IMGT Gene Table: human (Homo sapiens) IGHV. View on the IGHV3-66 gene and the IGHV3-66*04 allele.*

### 3 Locus representation



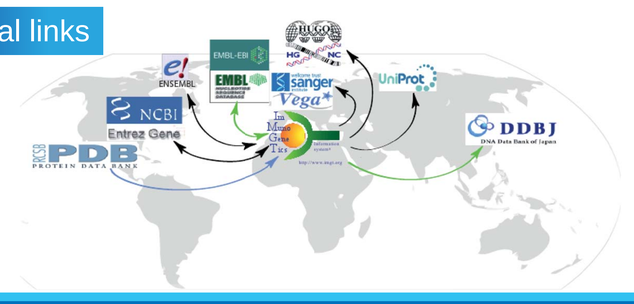

*Localization of the IGHV3-66 gene on human (Homo sapiens) IGH locus*

### 4 Chromosomal localization

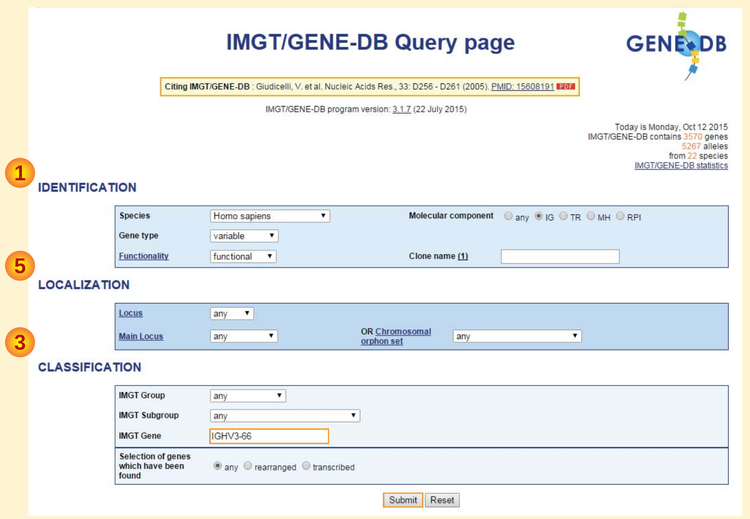



*Localization of the human (Homo sapiens) IGH locus on chromosome 14 (14q32.33)*

### 5 External links

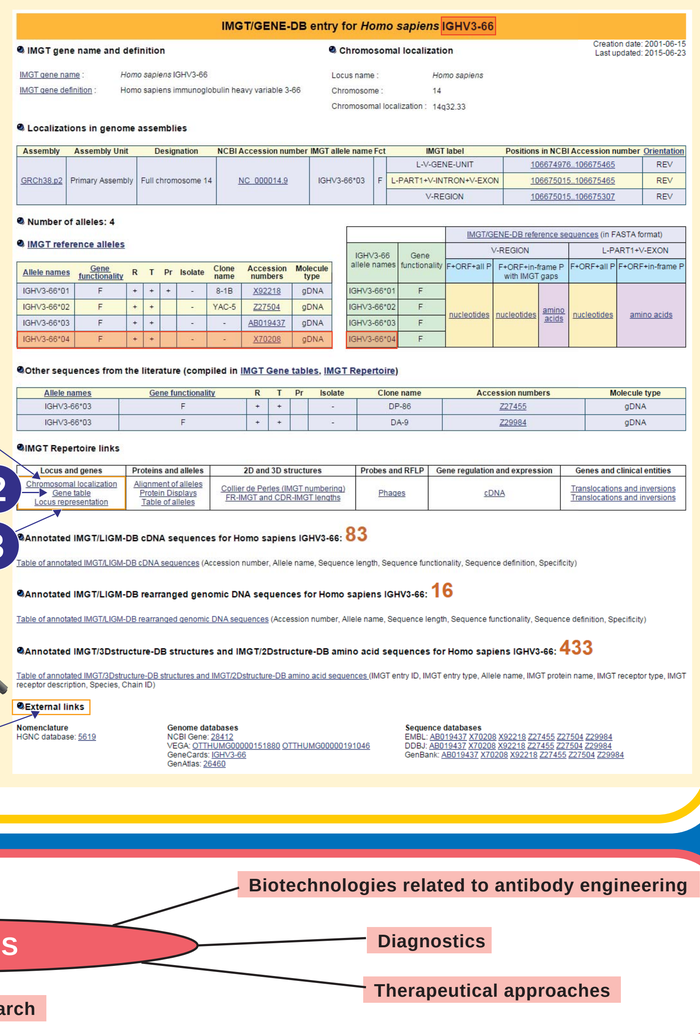



## INTERFACE WEB








## APPLICATIONS

- Fundamental and medical research
- Genome diversity and genome evolution
- Structural evolution of the IgSF and MhcSF
- Veterinary research
- Biotechnologies related to antibody engineering
- Diagnostics
- Therapeutical approaches

IMGT® director: Sofia Kossida (Sofia.Kossida@igh.cnrs.fr)
IMGT® founder and executive director emeritus: Marie-Paule Lefranc (Marie-Paule.Lefranc@igh.cnrs.fr)
Bioinformatics manager: Véronique Giudicelli (Veronique.Giudicelli@igh.cnrs.fr)
Computer manager: Patrice Duroux (Patrice.Duroux@igh.cnrs.fr)