IMGT® expert biocuration pipeline and IMGT/LIGMotif annotation tool for IG and TR genomic DNA sequences

Fatena Bellahcene, Géraldine Folch, Joumana Michaloud, Jérôme Lane, Amandine Lacan, Véronique Giudicelli, Patrice Duroux and Marie-Paule Lefranc

Université Montpellier 2 and CNRS, Laboratoire d'ImmunoGénétique Moléculaire (LIGM), Institut de Génétique Humaine (IGH), UPR CNRS 1142, Montpellier (France)

ImMunoGeneTics Information system®

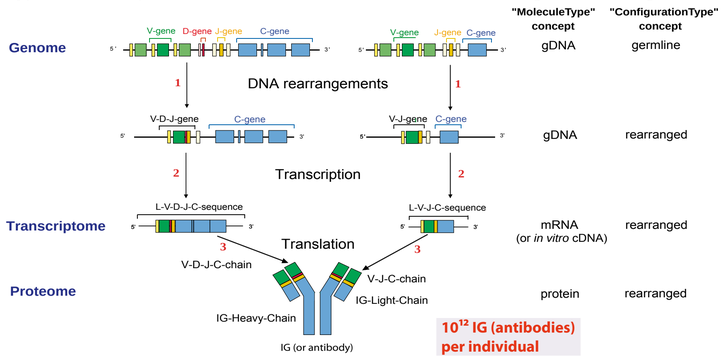
http://www.imgt.org

IMGT®, the international ImMunoGeneTics information system®, http://www.imgt.org, has developed an expert biocuration pipeline and the IMGT/LIGMotif annotation tool, for immunoglobulin (IG) and T cell receptor (TR) genomic DNA sequences. IG and TR genes are organized in the genome in several loci (7 in humans): IGH, IGL, IGK, TRA, TRB, TRD and TRG, that comprise different gene types: variable (V), diversity (D), joining (J) and constant (C) genes. Owing to the particularities of IG and TR gene structures related to their molecular mechanisms of rearrangements, conventional bioinformatic softwares and tools are not adapted to their identification and description in large genomic sequences. In order to answer that need, IMGT® has developed an expert biocuration pipeline for IG and TR genomic sequences from any vertebrate species, from fish to human, that includes IMGT/LIGMotif, a tool for the analysis of genomic sequences, up to 2.5 megabase pairs, of human and mouse (e.g. new haplotypes) and of closely related species (e.g. nonhuman primates, rat, respectively). Both are based on the IMGT standardized rules that are generated from the axioms and concepts of IMGT-ONTOLOGY. The procedure starts with the gene identification, locus assignment and specific standardized keywords (IDENTIFICATION) and orientation in the DNA sequence. The gene identification comprises a search of long motifs by alignment using BLAST. Alignments are selected according to BLAST parameters such as E-value, score, length and identity. In the second step, the identified V, D and J genes are described using the IMGT® standardized labels (DESCRIPTION). The procedure comprises the search of short conserved motifs (e.g. V-HEPTAMER), that are then used as anchors for the delimitation of the longest labels. IMGT/V-QUEST is used for the V-REGION label delimitations. Coding regions are delimited according to the IMGT unique numbering (NUMEROTATION). The third step is the gene functionality assignment with the specific IMGT® qualifier (Functional, ORF or Pseudogene). The functionality is identified based on well defined criteria that include the number of identified labels, the absence or presence of stop codons, the reading frame (open reading frame or frameshift) and the quality of recombination signal or other conserved motifs. The fourth step comprises the gene delimitation and cluster assembly. IMGT/LIGMotif results are checked by biocurators for consistency by comparison with IG and TR gene structure prototypes, in order to detect and revise error positions, organization and lack of IMGT labels assignment. Then, these data are integrated in IMGT/LIGM-DB temporary tables and can be retrieved in order to assign the IMGT nomenclature (CLASSIFICATION). For this purpose, a phylogenetic study (for IMGT subgroup determination), a gene mapping on locus (for IMGT gene determination) and a polymorphism study (for IMGT allele determination) are realized. Only human expertise allows to assign the IMGT gene and allele names. The complete annotation of V, D, J and C genes is followed by the update of IMGT® web resources (IMGT repertoire), IMGT® databases (IMGT/LIGM-DB, IMGT/GENE-DB, IMGT/2Dstructure-DB) and IMGT reference directories of IMGT® tools (IMGT/V-QUEST, IMGT/HighV-QUEST, IMGT/DomainGapAlign).

[1] Lane L., Duroux P. and Lefranc M.-P. BMC Bioinformatics, 11:223 (2010).
[2] IMGT booklet (11 papers), Cold Spring Harb Protocol, 124 pages (2011) (pdf, IMGTReferences, http://www.imgt.org). With generous provision from Cold Spring Harbor (CSH) Protocols.

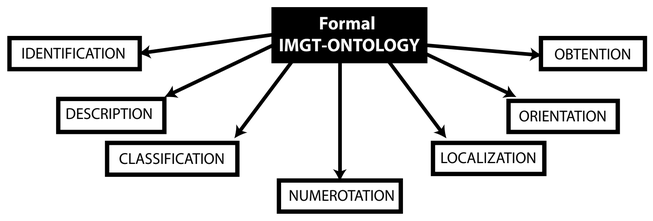
Biological context

The adaptive immune response is characterized by an extreme diversity of the specific antigen receptors that comprise the immunoglobulins (IG) or antibodies and the T cell receptors (TR) (10^{12} different IG and 10^{12} different TR per individual, in humans). The complex molecular mechanisms (DNA rearrangements, N-diversity, and for IG, somatic hypermutations) that occur in B cells and T cells are at the origin of that huge diversity.



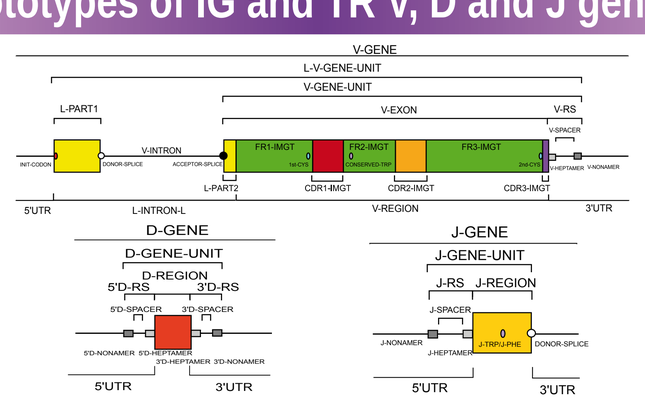
Formal IMGT-ONTOLOGY axioms

IMGT®, the international ImMunoGeneTics information system® (http://www.imgt.org) is based on IMGT-ONTOLOGY, the first ontology for immunogenetics and immunoinformatics [1]. IMGT-ONTOLOGY manages the immunogenetics knowledge through diverse facets that rely on seven axioms of the formal IMGT-ONTOLOGY or IMGT- Kaleidoscope [2]. Each axiom gives rise to a set of concepts. The concepts of identification, description, classification and numerotation are particuraly used for the immunogenetic sequence annotation.



[1] Giudicelli, V. and Lefranc, M-P., Bioinformatics, 15, 1047-1054 (1999).
[2] Duroux, P. et al., Biochimie, 90, 570-583 (2008).

Prototypes of IG and TR V, D and J genes

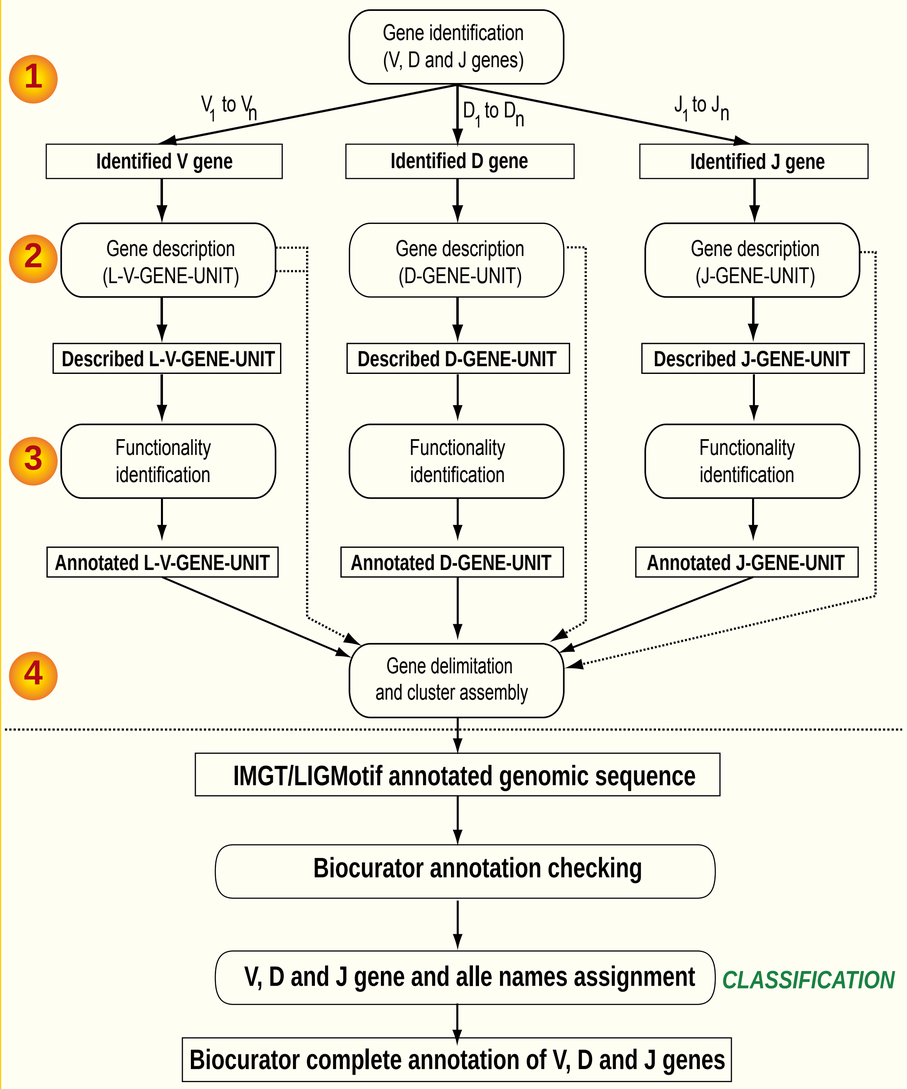


Each instance of the Molecular_EntityPrototype concept (DESCRIPTION) has a graphical representation or prototype.

Genomic sequences containing V, D or J genes can be described in using L-V-GENE-UNIT, D-GENE-UNIT and J-GENE-UNIT prototypes.

Biocuration pipeline overview

IMGT/LIGMotif comprises 4 modules: 'Gene identification', 'Gene description', 'Functionality identification' and 'Gene delimitation and cluster assembly'.



IMGT/LIGMotif annotated genomic sequence results are checked by biocurators (presence of all contitutive labels by comparaison with protytpe, consistency of relations between labels).V, D and J gene and allele names are assigned based on the IMGT concept of CLASIFICATION. Sequences and new genes are entered in IMGT/LIGM-DB and IMGT/GENE-DB respectively.

1 Gene identification — IDENTIFICATION

Comprises:

- Search of motifs by alignment using BLAST :

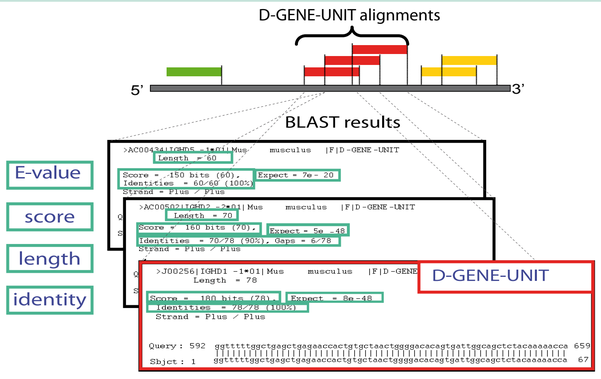
16 motifs for V-GENE		
V-GENE-UNIT	V-EXON	FR3-IMGT
L-PART1	V-REGION	CDR3-IMGT
L-INTRON-L	FR1-IMGT	V-RS
V-INTRON	CDR1-IMGT	V-SPACER
L-V-GENE-UNIT	FR2-IMGT	
L-PART2	CDR2-IMGT	

6 motifs for D-GENE	
D-GENE-UNIT	5'D-SPACER
D-REGION	3'D-RS
5'D-RS	3'D-SPACER

4 motifs for J-GENE
J-GENE-UNIT
J-REGION
J-RS
J-SPACER

- Alignment selection

Alignments are selected according to BLAST parameters such as E-Value, score, length and identity.

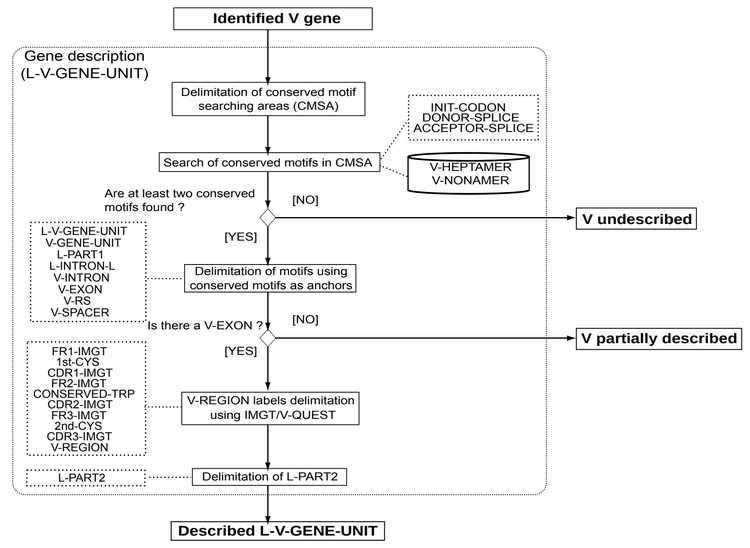


This allows to define, for each gene unit, the gene type (V, D or J gene), its localization and orientation in the DNA sequence.

2 Gene description — DESCRIPTION

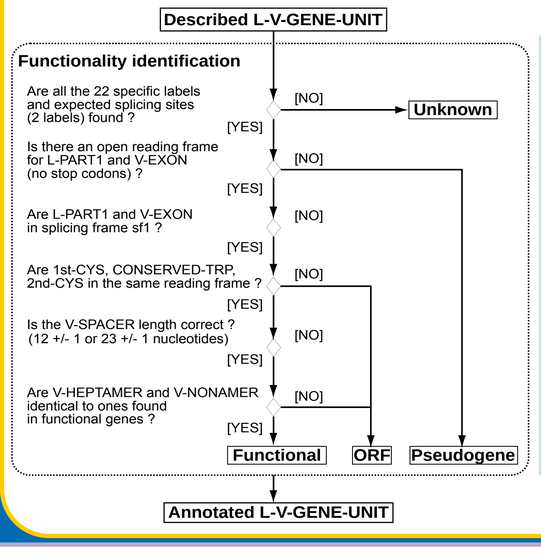
Gene description comprises :

- Delimitation of longest motifs using of short labels as anchors
- Delimitation of V-REGION labels using IMGT/V-QUEST
- IMGT standardized labels (DESCRIPTION)
- IMGT unique numbering (NUMEROTATION)



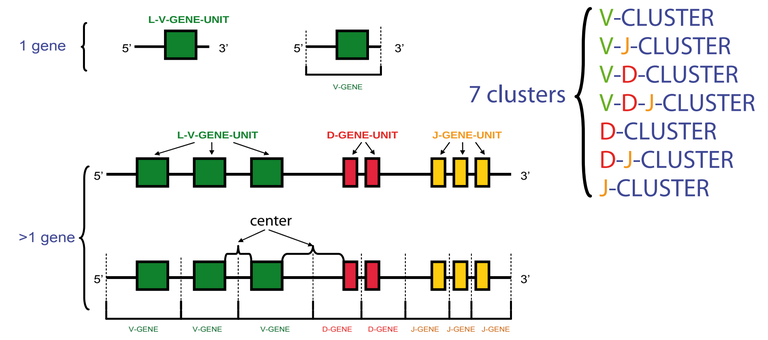
This allows to describe with standardized labels and using the IMGT unique numbering (NUMEROTATION) for coding regions of the L-V-GENE-UNIT, the D-GENE-UNIT and the J-GENE-UNIT.

3 Functionality identification

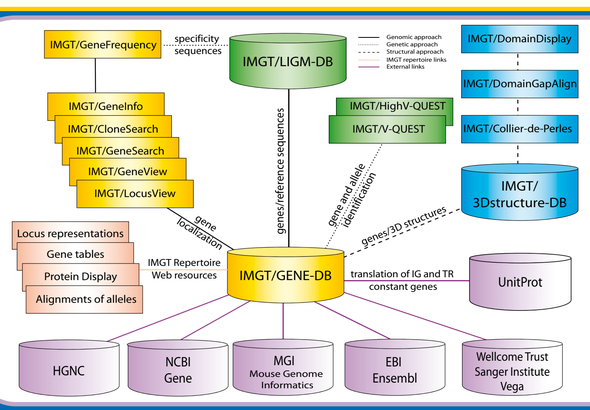


The functionality identification (Functional (F), ORF (Open Reading Frame) and Pseudogene (P)) is based on criteria that include the number of identified labels, the absence or presence of stop codons, the reading frame (ORF or frameshift) and the quality of IMGT labels asignment.

4 Gene delimitation and cluster assembly



IMGT/LIGMotif can identify 7 different clusters consisting of V, D and/or J genes.
IMGT cluster labels have been entered in Sequence Ontology (SO).



The IMGT® gene nomenclature for human IG and TR genes was approved by the Human Genome Organisation (HUGO) Nomenclature Committee (HGNC) in 1999 and endorsed by the World Health Organization-International Union of Immunological Societies (WHO-IUIS).

IMGT® IG and TR gene names have been entered in IMGT/GENE-DB, Human Genome Database (GDB), LocusLink (National Center for Biotechnology Information, NCBI), NCBI Entrez Gene when this gene database superseded LocusLink, NCBI Gene and MapViewer, Mouse Genome Informatics (MGI) in 2002 Ensembl (European Bioinformatics Institute, EBI), and Vega (Wellcome Trust Sanger Institute). The translation of the IG and TR constant genes was provided to Uniprot in 2008.

IMGT/GENE-DB provides links to IMGT® nucleotide sequence database (IMGT/LIGM-DB) and 3D structure database (IMGT/3Dstructure-DB), to IMGT® tools for analysis of nucleotide (IMGT/V-QUEST, IMGT/HighV-QUEST) and amino acid (IMGT/DomainGapAlign, IMGT/DomainDisplay) sequences, genomic (IMGT/GeneFrequency) and structural data, and to IMGT® Repertoire Web resources (Chromosomal localizations, Locus representations, Gene tables).

IMGT® founder and director: 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)
Webmaster: Chantal Ginestoux (Chantal.Ginestoux@igh.cnrs.fr)