IMGT-ONTOLOGY for immunogenetics knowledge management and biocuration in IMGT®

Joumana Jabado-Michaloud, Fatena Bellahcene, Géraldine Folch, Patrice Duroux, Véronique Giudicelli and Marie-Paule Lefranc

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

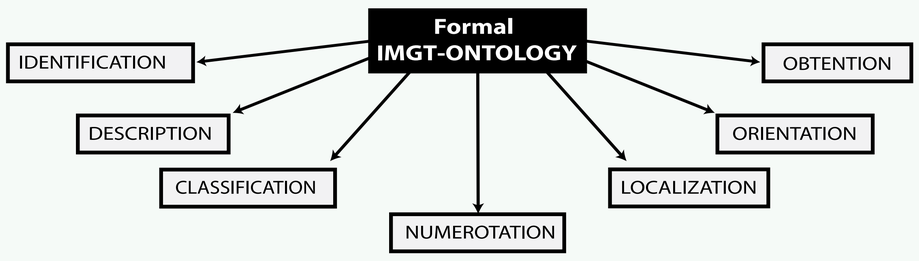
http://www.imgt.org

Immunogenetics is the science that studies the genetics of the immune system and immune responses. The adaptive immune response, acquired during evolution by vertebrates with jaws (gnathostomata), is characterized by an extreme diversity of the specific antigen receptors that comprise the immunoglobulins (IG) or antibodies and the T cell receptors (TR). The potential repertoire of each individual is estimated to comprise about 2 x 10^{12} different IG and TR, and the limiting factor is only the number of B and T cells that an organism is genetically programmed to produce. This huge diversity results from the complex and unique molecular synthesis and genetics of the antigen receptor chains that include DNA molecular rearrangements (combinatorial diversity) in multiple loci (three for IG and four for TR in humans) located on different chromosomes (four in humans), nucleotide deletions and insertions at the rearrangement junctions (or N-diversity), and somatic hypermutations in the IG loci. Owing to the complexity of the biological mechanisms, immunogenetics represents one of the greatest challenges for data interpretation. IMGT®, the international ImMunoGeneTics information system® (http://www.imgt.org) was created in 1989 by Marie-Paule Lefranc (University Montpellier 2 and CNRS) to answer the need of standardization and knowledge management. IMGT® is now acknowledged as the global reference in immunogenetics and immunoinformatics. IMGT® has reached that goal through the building of a unique ontology, IMGT-ONTOLOGY, which represents the first ontology in the domain. The Formal IMGT-ONTOLOGY, or IMGT-Kaleidoscope, includes seven axioms: "IDENTIFICATION", "DESCRIPTION", "CLASSIFICATION", "NUMEROTATION", "LOCALIZATION", "ORIENTATION" and "OBTENTION". These axioms have led to the generation of the concepts of IMGT-ONTOLOGY, and based on these concepts, to the IMGT Scientific chart rules. Thus for examples, the concepts of identification have led to the IMGT® standardized keywords, the concepts of description to the IMGT® standardized labels, the concepts of classification to the IMGT® standardized nomenclature (IMGT® gene and allele names approved by HGNC and endorsed by WHO/IUIS) and the concepts of numerotation to the IMGT unique numbering (for V, C and G domains) and to the IMGT Colliers de Perles. IMGT-ONTOLOGY is at the core of biocuration by human experts and of annotation by IMGT® automated resources. IMGT-ONTOLOGY is also key in the building of IMGT-Choreography (In Silico Biology, 2005) and in the evolution and content extension of the IMGT® system. In 2012, IMGT® is a high-quality integrated knowledge resource, specialized in the IG, TR and major histocompatibility (MH) proteins of humans and other vertebrates, proteins of the immunoglobulin superfamily (IgSF) and MH superfamily (MhSF), related proteins of the immune system (RPI) of vertebrates and invertebrates, therapeutic monoclonal antibodies (mAbs), fusion proteins for immune applications (FPIA), and composite proteins for clinical applications (CPCA). IMGT® provides a common access to standardized data from genome, proteome, genetics, two-dimensional (2D) and three-dimensional (3D) structures. IMGT® comprises 7 databases (sequence, gene, structure and specialist databases), 17 online tools and more than 15,000 pages of web resources.

[1] 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.*

Formal IMGT-ONTOLOGY axioms

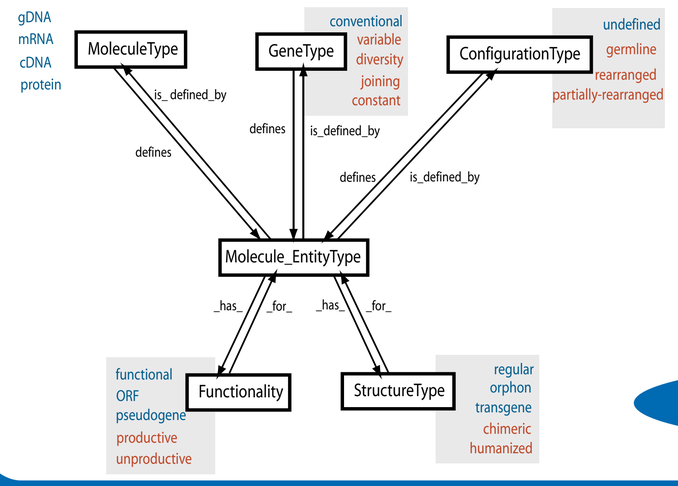
IMGT®, the international ImMunoGeneTics information system® (http://www.imgt.org) is based on IMGT-ONTOLOGY, the first ontology for immunogenetics and immunoinformatics [1]. IMGT-ONTOLOGY manages the immunogenetics knowledge through diverse facets that rely on seven axioms of the formal IMGT-ONTOLOGY 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).

1 The IDENTIFICATION axiom

The Molecular_EntityType concept



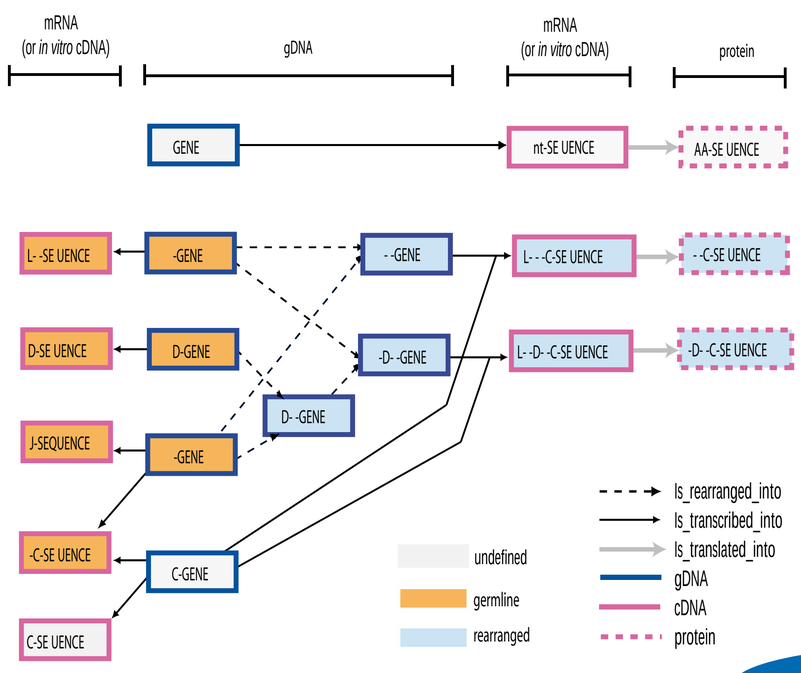
The IDENTIFICATION axiom has generated the concepts of identification which provide the terms and rules to identify an entity, its processes and its relations in IMGT®. They provide the IMGT® standardized keywords.

The "Molecule_EntityType" concept, shown as an example, is defined by the "MoleculeType", "GeneType" and "ConfigurationType" concept and has relations with the "Functionality" and "StructureType" concepts. It includes 38 leaf-concepts (L-V-gene, L-V-D-J-gene...).

Standardized keywords

2 The DESCRIPTION axiom

The Molecular_EntityPrototype concept

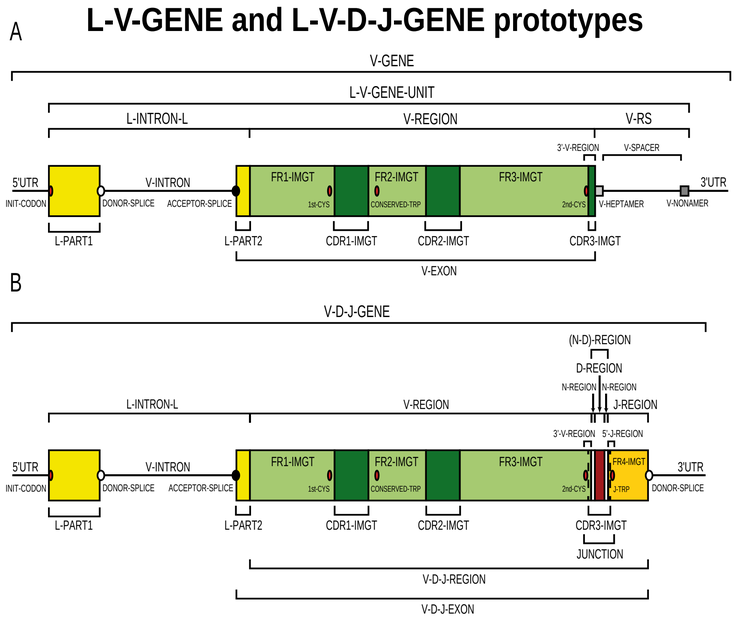


The DESCRIPTION axiom has generated the concepts of description which allow the description of any instance in IMGT®. The instances of the concepts of description correspond to IMGT® standardized labels. They are more than 560 standardizerd labels (available in the IMGT Scientific chart), 277 for the nucleotide sequences and 285 for the 3D structures.

Three concepts "GENE", "nt-SEQUENCE" and "AA-SEQUENCE" correspond to conventional genes while the 16 other concepts are specific of the IG and TR. The concepts for mRNA are also valid for in vitro cDNA. The first column correspond to "sterile transcript" concepts.

Standardized labels

A L-V-GENE and L-V-D-J-GENE prototypes



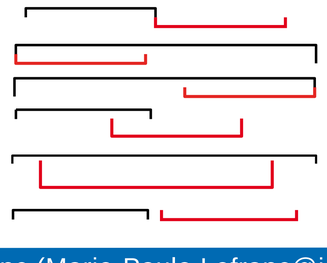
Each Molecular_EntityPrototype leaf-concept has a graphical representation or prototype. The L-V-GENE and L-V-D-J-GENE prototypes are shown.

Thirty-nine labels are necessary and sufficient for a complete description of the V-GENE and V-D-J-GENE prototypes (27 and 33 labels, respectively, 20 of them being shared by the two prototypes)

Relations between IMGT® labels

Relation	Reciprocal relation
"adjacent_in_its_3_prime_to"	"adjacent_in_its_5_prime_to"
"includes_with_same_5_prime",	"is_included_with_same_5_prime_in",
"includes_with_same_3_prime",	"is_included_with_same_3_prime_in",
"overlaps_in_its_3_prime_with"	"overlaps_in_its_5_prime_with"
"includes"	"is_included_in"
"is_in_3_prime_of"	"is_in_5_prime_of"

Relations between IMGT labels



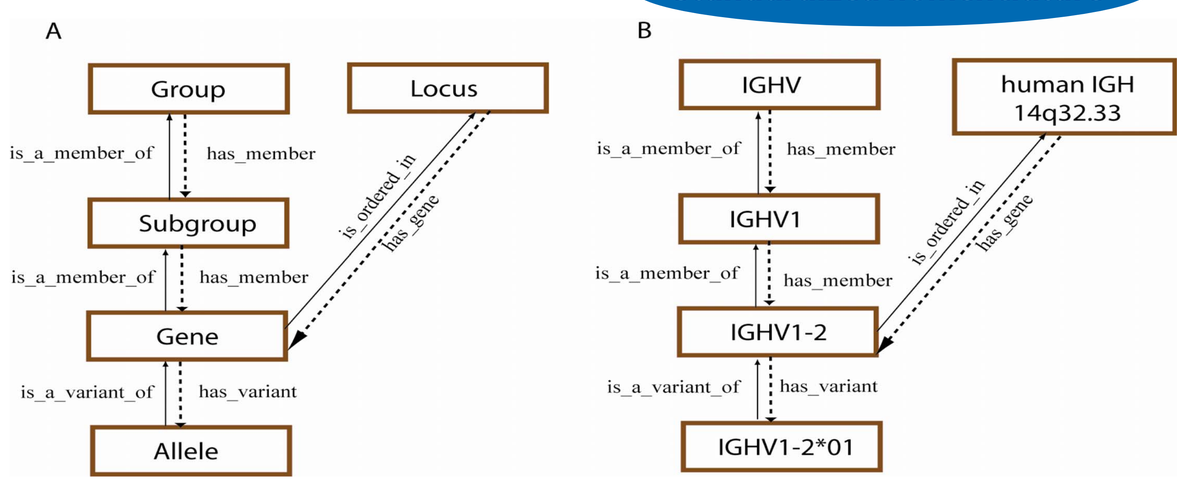
Twelve relations between labels are necessary and sufficient for a complete description of any Molecular_EntityPrototype .

3 The CLASSIFICATION axiom

The CLASSIFICATION axiom generates the concepts of classification, they allow to classify and name the genes and their alleles. The genes which code the IG and TR belong to highly polymorphic multigenic families. A major contribution of IMGT-ONTOLOGY was to set the principles of their classification and to propose a standardized nomenclature[1,2].

[1] Lefranc, M.-P. and Lefranc, G., The Immunoglobulin FactsBook (2001)
[2] Lefranc, M.-P. and Lefranc, G., The T cell receptor FactsBooks (2001)

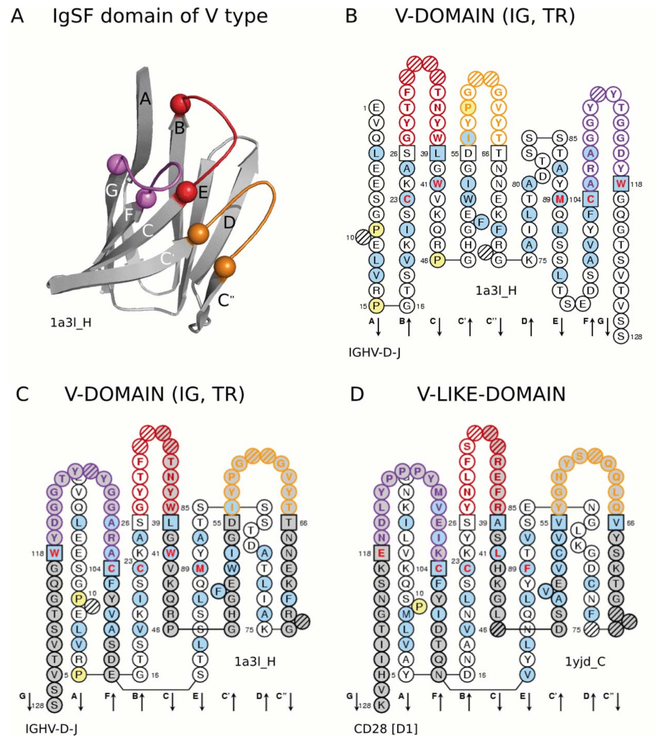
Standardized nomenclature



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, Ensembl (European Bioinformatics Institute, EBI), and Vega (Wellcome Trust Sanger Institute).

4 The NUMEROTATION axiom



IMGT Collier de Perles for V domain. (A) Ribbon representation of a V-DOMAIN. (B) and (C) V-DOMAIN on one layer and on two layers, respectively (*Mus musculus* VH [8.8.12]). (D) V-LIKE-DOMAIN on two layers (*Homo sapiens* CD28 [9.9.13]).

IMGT unique numbering

The NUMEROTATION axiom and the concepts of numerotation determine the principles of a unique numbering for a domain (sequences and 3D structures). The "IMGT_unique_numbering" concept is illustrated by the "IMGT_Collier_de_Perles" concept which allows graphical representation in two dimensions (2D) of amino acid sequences of varible (V) [1], constant (C) [2] or groove (G) [3] domains.

[1] Lefranc, M.-P. et al., Dev. Comp. Immunol., 27, 55-77 (2003). [2] Lefranc, M.-P. et al., Dev. Comp. Immunol., 29, 185-203 (2005) [3]Lefranc, M.-P. et al., Dev. Comp. Immunol., 29, 917-938 (2005)

The conserved amino acids always have the same position, for instance cysteine 23 (1st-CYS), tryptophan 41 (CONSERVED-TRP), hydrophobic amino acid 89, cysteine 104 (2nd-CYS), phenylalanine or tryptophan 118 (J-PHE or J-TRP). The IMGT unique numbering provides a standardized delimitation of the framework regions (FR1-IMGT: positions 1 to 26, FR2-IMGT: 39 to 55, FR3-IMGT: 66 to 104 and FR4-IMGT: 118 to 128) and of the complementarity determining regions: CDR1-IMGT: 27 to 38, CDR2-IMGT: 56 to 65 and CDR3-IMGT: 105 to 117. CDR-IMGT lengths (shown between brackets and separated by dots, e.g. [8.8.12]) are crucial information for antibody engineering and antibody humanization.

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)