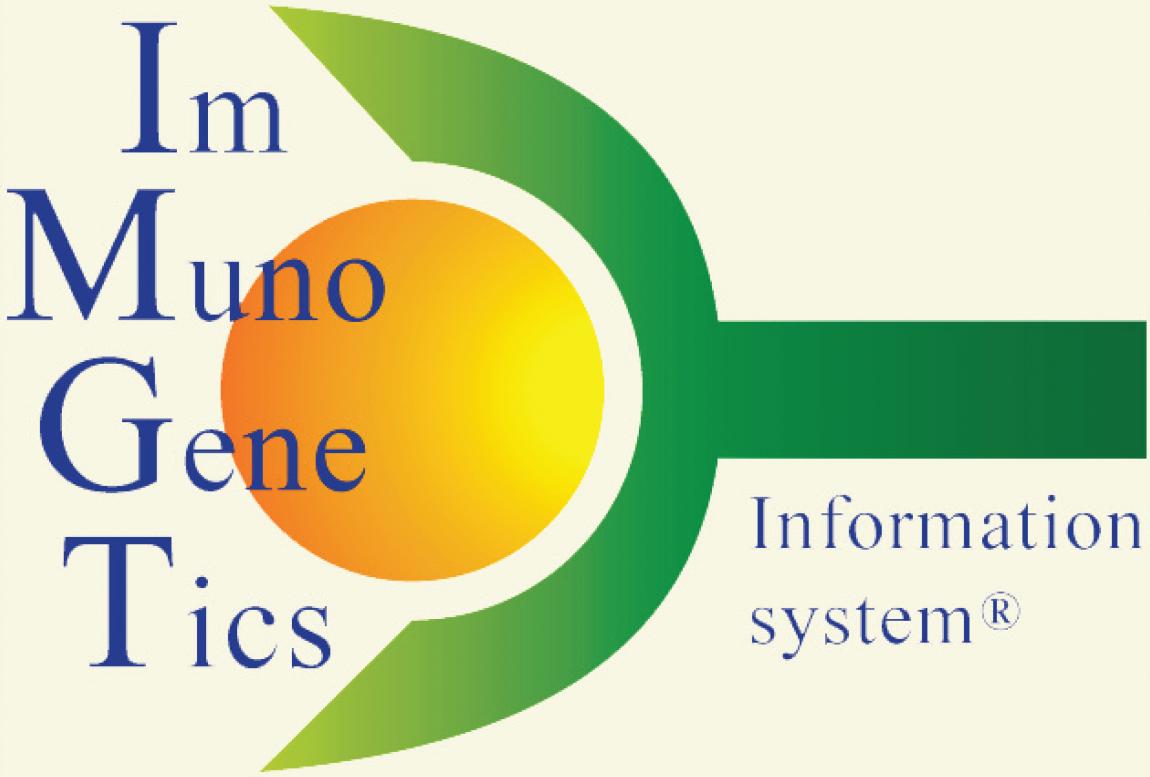
COLD SPRING HARBOR

Protocols



CSH PRESS Cold Spring Harbor Laboratory Press

Cold Spring Harbor Protocols www.cshprotocols.org

Contents

TOPIC INTRODUCTION

INFORMATION PANELS

PROTOCOLS

Topic Introduction

IMGT, the International ImMunoGeneTics Information System

Marie-Paule Lefranc

IMGT, the international ImMunoGeneTics information system, Laboratoire d'ImmunoGénétique Moléculaire LIGM, Université Montpellier 2, Institut de Génétique Humaine IGH, UPR CNRS 1142, 34396 Montpellier cedex 5, France

INTRODUCTION

IMGT, the international ImMunoGeneTics information system (http://www.imgt.org), created in 1989 by the Laboratoire d'ImmunoGénétique Moléculaire LIGM (Université Montpellier 2 and CNRS, Montpellier, France), is the global reference in immunogenetics and immunoinformatics. IMGT-ONTOLOGY concepts, generated from the Formal IMGT-ONTOLOGY or IMGT-Kaleidoscope axioms, have allowed us to set up the IMGT Scientific chart rules and to standardize and manage the complexity of immunogenetics data. IMGT is a high-quality integrated knowledge resource specializing in the immunoglobulins (IG) or antibodies, T cell receptors (TR), and major histocompatibility (MH) of human and other vertebrate species, proteins of the immunoglobulin superfamily (IgSF) and MH superfamily (MhSF), related proteins of the immune systems (RPI) of vertebrates and invertebrates, therapeutic monoclonal antibodies (mAb), and fusion proteins for immune applications (FPIA). IMGT provides a common access to standardized data from genome, proteome, genetics, and three-dimensional (3D) structures. IMGT consists of databases, interactive online tools, and Web resources.

RELATED INFORMATION

IMGT is freely available at http://www.imgt.org, which provides documentation for IMGT databases and tools. IMGT-ONTOLOGY concepts of identification are available at the National Center for Biomedical Ontology (NCBO) BioPortal http://bioportal.bioontology.org/ and at IMGT http://www.imgt.org.

Information is also available in **From IMGT-ONTOLOGY IDENTIFICATION Axiom to IMGT Standardized Keywords: For Immunoglobulins (IG), T Cell Receptors (TR) and Conventional Genes** (Lefranc 2011a), **From IMGT-ONTOLOGY DESCRIPTION Axiom to IMGT Standardized Labels: For Immunoglobulin (IG) and T Cell Receptor (TR) Sequences and Structures** (Lefranc 2011b), **From IMGT-ONTOLOGY CLASSIFICATION Axiom to IMGT Standardized Gene and Allele Nomenclature: For Immunoglobulins (IG) and T Cell Receptors (TR)** (Lefranc 2011c), **IMGT Unique Numbering for the Variable (V), Constant (C), and Groove (G) Domains of IG, TR, MH, IgSF, and MhSF** (Lefranc 2011d), and **IMGT Collier de Perles for the Variable (V), Constant (C), and Groove (G) Domains of IG, TR, MH, IgSF, and MhSF** (Lefranc 2011e).

In addition, protocols are available for **IMGT/V-QUEST: IMGT Standardized Analysis of the Immunoglobulin (IG) and T Cell Receptor (TR) Nucleotide Sequences** (Giudicelli et al. 2011), **IMGT/JunctionAnalysis: IMGT Standardized Analysis of the V-J and V-D-J Junctions of the Rearranged Immunoglobulins (IG) and T Cell Receptors (TR)** (Giudicelli and Lefranc 2011), **IMGT/Collier de Perles: IMGT Standardized Representation of Domains (IG, TR, and IgSF Variable and Constant Domains, MH and MhSF Groove Domains)** (Ehrenmann et al. 2011), **IMGT/DomainGapAlign: IMGT Standardized Analysis of Amino Acid Sequences of Variable, Constant, and Groove Domains (IG, TR, MH, IgSF, MhSF)** (Ehrenmann and Lefranc 2011a), and **IMGT/3Dstructure-DB: Querying the IMGT Database for 3D Structures in Immunology and Immunoinformatics (IG or Antibodies, TR, MH, RPI, and FPIA)** (Ehrenmann and Lefranc 2011b).

BACKGROUND INFORMATION

The amount of genomics, genetics, 3D, and functional data published in the immunogenetics field is growing exponentially and involves fundamental, clinical, veterinary, and pharmaceutical research.

Corresponding author (Marie-Paule.Lefranc@igh.cnrs.fr).
Cite as: Cold Spring Harb Protoc; 2011; doi:10.1101/pdb.top115

The number of potential protein forms of the antigen receptors, IG or antibodies, and TR is almost unlimited. The potential repertoire of each individual is estimated to comprise about 2×10^{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 is inherent to the particularly complex and unique molecular synthesis and genetics of the antigen receptor chains (Fig. 1). This includes biological mechanisms such as 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 (for review, see Lefranc and Lefranc 2001a,b).

IMGT, the international ImMunoGeneTics information system (http://www.imgt.org) (Lefranc et al. 2009), was created in 1989 by Marie-Paule Lefranc, Laboratoire d'ImmunoGénétique Moléculaire LIGM (Université Montpellier 2 and CNRS) at Montpellier, France, in order to standardize and to manage the complexity of immunogenetics data. IMGT has reached that goal through the building of a unique ontology, IMGT-ONTOLOGY (Giudicelli and Lefranc 1999; Lefranc et al. 2004, 2005a, 2008; Duroux et al. 2008), the first ontology in immunogenetics and immunoinformatics.

IMGT-ONTOLOGY

IMGT-ONTOLOGY (Giudicelli and Lefranc 1999; Lefranc et al. 2004, 2005a; 2008; Duroux et al. 2008), the first ontology for immunogenetics and immunoinformatics, has been built to ensure the accuracy and the consistency of the IMGT data as well as the coherence between the components (databases, tools, and Web resources) of IMGT, the international ImMunoGeneTics information system (http://www.imgt.org). IMGT-ONTOLOGY is now acknowledged as the global reference in immunogenetics and immunoinformatics, allowing IMGT to bridge biological and computational spheres in bioinformatics (Lefranc et al. 2008).

IMGT-ONTOLOGY KNOWLEDGE DOMAIN

An ontology is a formal representation of a knowledge domain (Gruber 1993; Giudicelli and Lefranc 1999). The knowledge domain covered by IMGT-ONTOLOGY is immunogenetics and immunoinformatics.

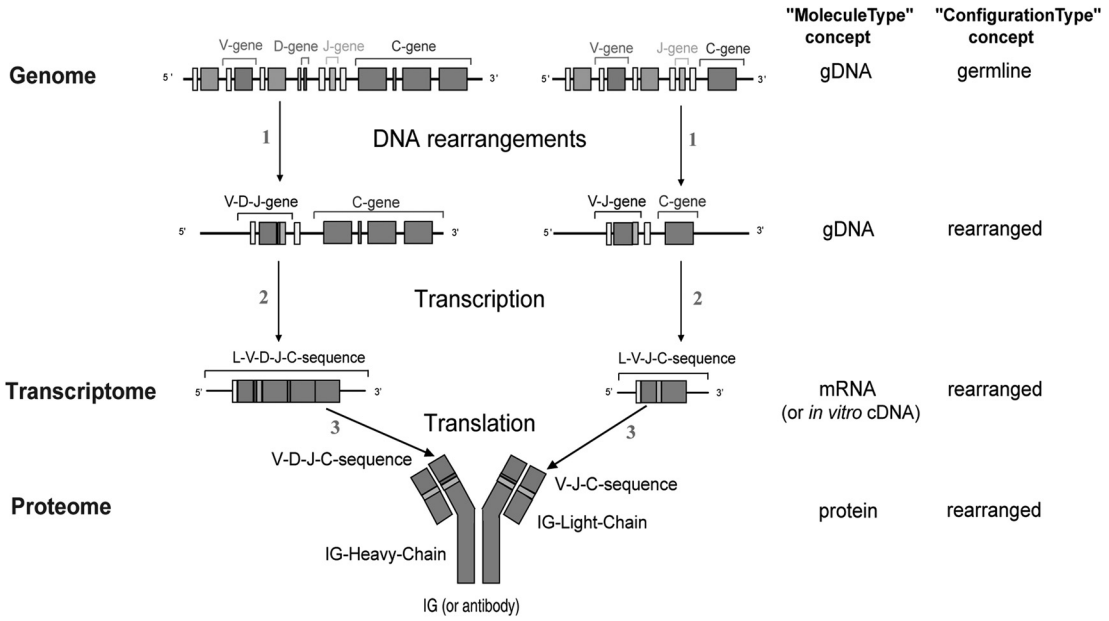


FIGURE 1. Synthesis of an IG or antibody in humans. A human being may potentially synthesize 10^{12} different antibodies (Lefranc and Lefranc 2001a). (1) DNA rearrangements (is_rearranged_into), (2) Transcription (is_transcribed_into), (3) Translation (is_translated_into). The configuration of C-GENE is "undefined" (ConfigurationType) (IMGT Repertoire, http://www/imgt/org).

Immunogenetics studies the genetics of immune responses. In addition to the innate immune response present in all living organisms, vertebrates with jaws (or gnathostomata, from fish to humans) have acquired the adaptative immune response, characterized by an extreme diversity of the specific antigen receptors that comprise the IG or antibodies and the TR (10^{12} different IG and 10^{12} different TR per individual, in humans). Complex molecular mechanisms in B cells (Fig. 1) and in T cells are at the origin of this huge diversity (Lefranc and Lefranc 2001a,b). These mechanisms include, in particular, DNA rearrangements and, for the IG, somatic hypermutations. In addition, there is considerable polymorphism of the MH proteins (human leucocyte antigens [HLA] in humans). These particularities of the adaptive immune system of the vertebrates, and a better knowledge of the innate immune responses found in any species, have made the immune system an excellent model for systems biology.

IMGT-ONTOLOGY AXIOMS AND CONCEPTS

IMGT-ONTOLOGY manages the immunogenetics knowledge through diverse facets that rely on the seven axioms of the Formal IMGT-ONTOLOGY or IMGT-Kaleidoscope: "IDENTIFICATION," "DESCRIPTION," "CLASSIFICATION," "NUMEROTATION," "LOCALIZATION," "ORIENTATION," and "OBTENTION" (Giudicelli and Lefranc 1999; Lefranc et al. 2004, 2005a, 2008; Duroux et al. 2008). These axioms postulate that any object, any process, and any relation has to be identified, described, classified, numbered, localized, and oriented, and that the way it is obtained can be characterized (Duroux et al. 2008; Lefranc et al. 2008) (Fig. 2).

The concepts generated from these axioms have been essential for establishing the IMGT Scientific chart rules for the genome, proteome, genetics, two-dimensional (2D) and 3D structures (Lefranc 2005, 2008b,c). The IMGT-ONTOLOGY axioms, the concepts generated from them, examples of concepts, and IMGT Scientific chart rules are listed in Table 1.

As an example, the NUMEROTATION axiom has generated two major concepts of numerotation:

- the IMGT unique numbering for the variable (V) domain (Lefranc 1997, 1999; Lefranc et al. 2003) and the constant (C) domain (Lefranc et al. 2005c) of the IG, TR, and IgSF, and for the groove (G) domain (Lefranc et al. 2005b) of the MH and MhSF.
- the IMGT Colliers de Perles for V, C, and G domains or graphical 2D representations of the domains that bridge the gap between sequences and 3D structures (Ruiz and Lefranc 2002; Garapati and Lefranc 2007; Kaas and Lefranc 2007; Kaas et al. 2007).

The IMGT Scientific chart rules based on the concepts of identification, description, classification, and numerotation have been essential in providing the standards currently used in immunogenetics and immunoinformatics: standardized keywords, prototypes and standardized labels, official nomenclature of the IG and TR genes and alleles (Lefranc 2000a,b; Lefranc and Lefranc 2001a,b), framework region (FR)-IMGT and complementarity determining region (CDR)-IMGT delimitations, contact analysis and paratope/epitope characterization, etc. (see Table 2).

IMGT DATABASES, TOOLS, AND WEB RESOURCES

Based on the IMGT-ONTOLOGY concepts, IMGT is a high-quality integrated knowledge resource, specializing in the IG, TR, and MH of human and other vertebrates, proteins of the immunoglobulin

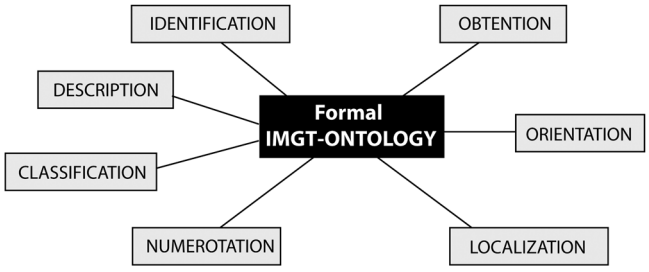


FIGURE 2. The seven axioms of the Formal IMGT-ONTOLOGY or IMGT-Kaleidoscope. (Modified from Duroux et al. 2008 and reprinted with permission from Elsevier © 2008.)

Table 1. Axioms of the formal IMGT-ONTOLOGY (or IMGT-Kaleidoscope), concepts generated from them, examples of concepts, and IMGT Scientific chart rules

Formal IMGT-ONTOLOGY axioms	Concepts generated from the axioms	Examples of IMGT-ONTOLOGY concepts	IMGT Scientific chart rules[a]
IDENTIFICATION axiom	Concepts of identification	ChainType[b] ConfigurationType DomainType FunctionalityType FunctionType GeneType LocationType Molecule_EntityType MoleculeType MoleculeUnit ReceptorType[b] SpecificityType StructureType TaxonRank[b]	Standardized keywords
DESCRIPTION axiom	Concepts of description	Core Molecule_EntityPrototype Recombination signal (RS)	Prototypes Standardized labels
CLASSIFICATION axiom	Concepts of classification	Group Subgroup Gene Allele	Gene and allele nomenclature
NUMEROTATION axiom	Concepts of numerotation	IMGT_unique_numbering (IMGT unique numbering) for V, C and G IMGT_Collier_de_Perles (IMGT Collier de Perles) for V, C and G domains	FR-IMGT and CDR-IMGT delimitations Contact analysis Paratope/epitope characterization Conserved positions
LOCALIZATION axiom	Concepts of localization	Locus Chromosome Cell compartment Organ part Organ	Locus representation Chromosomal localization Standardized keywords
ORIENTATION axiom	Concepts of orientation	DNA_strand_orientation Genomic_orientation[b]	IMGT Index
OBTENTION axiom	Concepts of obtention	Origin Methodology	Standardized keywords

[a]The corresponding controlled vocabulary and rules are available in the IMGT Scientific Chart at http://www.imgt.org.
[b]Indicates a high concept, that is, a concept with different levels of granularity.

superfamily (IgSF) and MH superfamily (MhSF), RPI of vertebrates and invertebrates, therapeutic mAbs and FPIA (Lefranc et al. 2009). IMGT provides a common access to standardized data from genome, proteome, genetics, and 3D structures (Lefranc 2005, 2008b,c). The IMGT information system manages immunogenetics knowledge according to three main IMGT biological approaches: genomic, genetic, and structural (Fig. 3). For each approach, IMGT provides databases, interactive online tools, and Web resources (Table 2).

Genomic Approach

The IMGT genomic approach is gene-centered and mainly oriented toward the study of the genes within their loci and on the chromosomes.

IMGT/GENE-DB (Giudicelli et al. 2005), the IMGT genome database, is the official repository of all of the IG and TR genes and alleles approved by the World Health Organization (WHO)/International Union

Table 2. IMGT databases, tools, and Web resources for genomic, genetic, and structural approaches

Approaches	Databases	Tools	Web resources[a]
Genomic	IMGT/GENE-DB	IMGT/GeneView	IMGT Repertoire "Locus and genes" section:
		IMGT/LocusView	Chromosomal localizations
		IMGT/CloneSearch	Locus representations
		IMGT/GeneSearch	Locus description, etc.
		IMGT/GeneInfo	Gene tables
		IMGT/LIGMotif	Potential germline repertoires
		IMGT/GeneFrequency	Lists of genes
			Correspondence between nomenclatures, etc.
Genetic	IMGT/LIGM-DB	IMGT/V-QUEST	IMGT Repertoire "Proteins and alleles" section:
	IMGT/MHC-DB	IMGT/JunctionAnalysis	Alignments of alleles
	IMGT/PRIMER-DB	IMGT/HighV-QUEST	Tables of alleles
	IMGT/	IMGT/BlastSearch	Allotypes
	2Dstructure-DB	IMGT/Allele-Align	Isotypes
		IMGT/PhyloGene	Protein displays,
		IMGT/DomainDisplay	etc.
Structural	IMGT/	IMGT/DomainGapAlign	IMGT Repertoire "2D and 3D structures" section:
	3Dstructure-DB	IMGT/Collier de Perles	IMGT Colliers de Perles
		IMGT/	FR-IMGT and CDR-IMGT lengths
		DomainSuperimpose	IMGT classes for amino acid characteristics
		IMGT/StructuralQuery	IMGT Colliers de Perles reference profiles
			3D representations,
			etc.

[a]Only Web resources examples from the IMGT Repertoire sections are shown.

of Immunological Societies (IUIS) Nomenclature Subcommittee for IG and TR (Lefranc 2007, 2008a). Reciprocal links exist between IMGT/GENE-DB and the Human Genome Nomenclature Committee (HGNC) database and Gene at the National Center for Biotechnology Information (NCBI).

The IMGT genomics tools manage the locus organization and gene location and provide the display of physical maps for the human and mouse IG, TR, and MH loci. They allow one to view genes in a locus (IMGT/GeneView, IMGT/LocusView), to search for clones (IMGT/CloneSearch), to search for genes in a locus based on IMGT gene names, functionality, or localization on the chromosome (IMGT/GeneSearch, IMGT/GeneInfo) (Baum et al. 2004, 2006), to identify and describe IG and TR genes in large genomic sequences (IMGT/LIGMotif) (Lane et al. 2010), or to provide a graphical representation of the numbers of available sequences containing rearranged IG and TR genes (IMGT/GeneFrequency).

The IMGT genomics resources include chromosomal localizations, locus representations, locus description, gene positions, gene exon/intron organization, gene exon/intron splicing sites, gene tables, potential germline repertoires, lists of genes, and correspondence between nomenclatures (IMGT Repertoire "Locus and genes" section).

Genetic Approach

The IMGT genetic approach refers to the study of the genes in relation to their sequence polymorphisms and mutations, their expression, their specificity, and their evolution.
The IMGT sequences databases include:

1. IMGT/LIGM-DB (Giudicelli et al. 2006), created in 1989 and on the Web since July 1995. It is the first and the largest IMGT database, containing IG and TR nucleotide sequences and the translation of annotated sequences from human and other vertebrate species (more than 150,000 sequences from 261 species in December 2010).
2. IMGT/MH-DB, hosted at EBI, for the MH sequences of human (HLA) and other vertebrates.

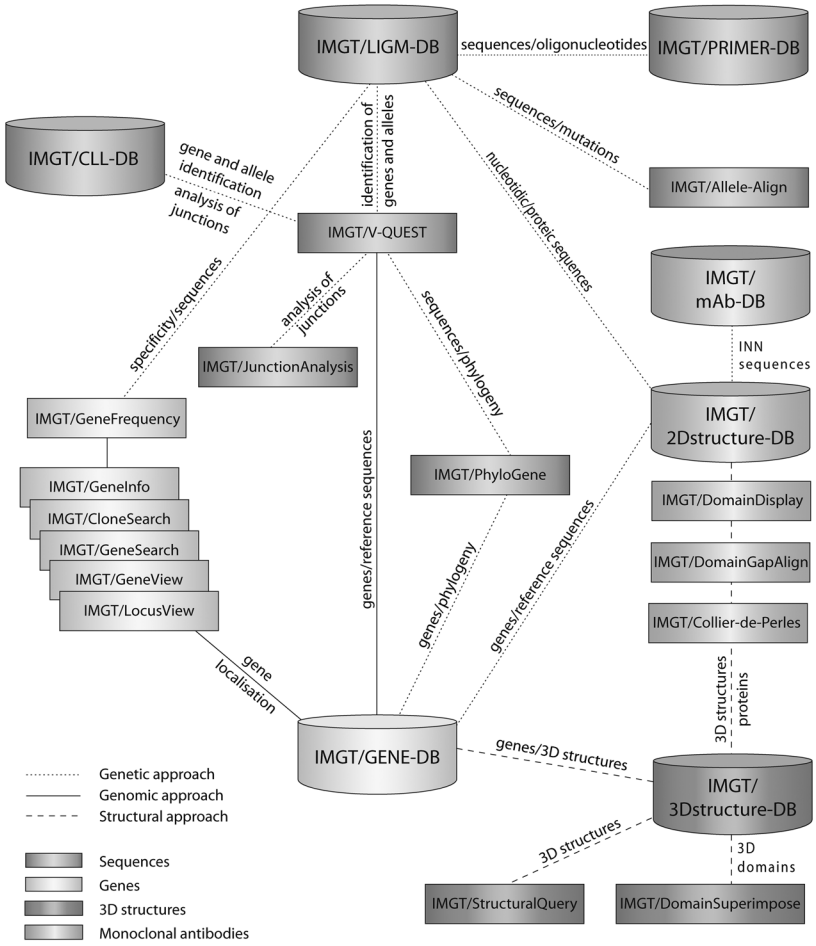


FIGURE 3. IMGT (http://www.imgt.org) comprises databases (shown as cylinders), tools (shown as rectangles), and Web resources (data not shown, see Table 2 for examples). Interactions between databases and tools in the genomic, genetic, and structural approaches are represented with continuous, dotted, and broken lines. (Modified from Lefranc et al. 2009 and reprinted with permission from Oxford University Press © 2009.)

3. IMGT/PRIMER-DB that contains oligonucleotide sequences used as primers for combinatorial library constructions, single-chain Fragment variable (scFv), phage display, or microarray technologies.
4. IMGT/2Dstructure-DB that contains amino acid sequences of Kabat IG (those only known at the amino acid level) and of therapeutic mAbs and FPIA of the WHO/International Nonproprietary Name (INN) programme (Lefranc 2011f).

The IMGT sequence analysis tools include:

1. IMGT/V-QUEST (Giudicelli et al. 2004; Brochet et al. 2008) for the identification of the variable (V), diversity (D), and joining (J) genes and the description of their mutations in rearranged IG and TR sequences. IMGT/V-QUEST has been approved by the European Research Initiative on chronic lymphocytic leukaemia CLL (ERIC) for the analysis of the IGHV gene mutational status in CLL (Ghia et al. 2007; Giudicelli and Lefranc 2009).
2. IMGT/JunctionAnalysis (Yousfi Monod et al. 2004) for a comprehensive analysis of the V-J and V-D-J junctions that confer the antigen receptor specificity (Bleakley et al. 2006, 2008).
3. IMGT/HighV-QUEST, a high-throughput version of IMGT/V-QUEST, for the analysis of the IG and TR sequences, generated from Next Generation Sequencing (NGS) or deep sequencing, by batches of up to 150,000 sequences (Alamyar et al. 2010).

4. IMGT/BlastSearch for a sequence search against the following databases: IMGT/LIGM-DB, IMGT/GENE-DB reference sequences, IMGT/3Dstructure-DB sequences, IMGT/3Dstructure-DB domain sequences, or IMGT/DomainDisplay reference sequences.
5. IMGT/Allele-Align for the detection of allelic polymorphisms.
6. IMGT/Phylogene (Elemento and Lefranc 2003) for IG and TR gene evolution analyses.
7. IMGT/DomainDisplay (Ehrenmann et al. 2010a,b) for the display of amino acid sequences from the IMGT domain directory.

The IMGT genetics resources include alignments of alleles, tables of alleles, allotypes (Jefferis and Lefranc 2009), isotypes, and protein displays (IMGT Repertoire "Proteins and alleles" section).

Structural Approach

The IMGT structural approach refers to the study of the 2D and 3D structures and to the antigen- or ligand-binding characteristics in relationship to the protein functions, polymorphisms, and evolution.

IMGT/3Dstructure-DB (Kaas et al. 2004; Ehrenmann et al. 2010b), the IMGT 3D structure database, comprises IG, TR, MH, IgSF, MhSF, RPI, mAb, and FPIA with known 3D structures. The IMGT/3Dstructure-DB cards provide chain details with IMGT annotations (receptor, chain and domain description, IMGT gene and allele names), domain delimitations and IMGT Collier de Perles, contact analysis, downloadable renumbered IMGT/3Dstructure-DB flat files, visualization tools (Jmol and QuickPDB), and external links. IMGT Residue@Position cards provide detailed information on the inter- and intradomain contacts at each residue position, based on the IMGT unique numbering. IMGT/3Dstructure-DB entries containing IG/antigen (IG/Ag) or TR/peptide/MH (TR/pMH) complexes can be queried for standardized analysis of the interactions between paratope/epitope (Kaas and Lefranc 2005; Kaas et al. 2008; Lefranc 2009; Ehrenmann et al. 2010a,b).

The IMGT structural tools include:

1. IMGT/DomainGapAlign (Ehrenmann et al. 2010b) for the analysis of amino acid sequences per domain, including those of the IG and TR V domains.
2. IMGT/Collier de Perles to make your own IMGT Collier de Perles.
3. IMGT/DomainSuperimpose to superimpose two domain 3D structures from IMGT/3Dstructure-DB.
4. IMGT/StructuralQuery (Kaas et al. 2004) to retrieve the IMGT/3Dstructure-DB entries containing a V domain, based on specific structural characteristics of the intramolecular interactions: e.g., ϕ and ψ angles, distance in angstrom between amino acids, CDR-IMGT lengths.

The IMGT structural resources include IMGT Colliers de Perles (2D representations on one layer or two layers), FR-IMGT and CDR-IMGT lengths, IMGT classes for amino acid characteristics, and profiles based on the 11 IMGT amino acid physicochemical classes (Pommié et al. 2004) (IMGT Repertoire "2D and 3D structures" section).

Other IMGT Databases and Web Resources

IMGT/CLL-DB (Brochet 2008) contains IG sequences of CLL patients, analyzed by IMGT/V-QUEST. IMGT/mAb-DB (Poiron et al. 2010) provides information on therapeutic mAbs and FPIA. In addition to the IMGT Scientific chart and IMGT Repertoire, IMGT Web resources comprise IMGT Index, IMGT Bloc-notes, IMGT Education (e.g., IMGT Lexique, Aide-Mémoire, Tutorials), The IMGT Medical page, The IMGT Veterinary page, The IMGT Biotechnology page, The IMGT Immunoinformatics page, and IMGT other accesses.

IMGT USERS

Since July 1995 IMGT has been available on the Web at the IMGT Home page http://www.imgt.org. The knowledge provided by the IMGT information system is of much value to clinicians and biological scientists in general. IMGT databases, tools, and Web resources are extensively queried and used by scientists from both academic and research laboratories from pharmaceutical companies, who are equally

distributed between the United States, Europe, and the remaining world. IMGT is used in very diverse domains: (1) basic research, (2) medical research (vaccine design, repertoire analysis of the IG antibody sites and of the TR recognition sites in normal and pathological situations such as autoimmune diseases, infectious diseases, AIDS, leukemias, lymphomas, and myelomas), (3) veterinary research (immune repertoires in domestic and wildlife species, animal models), (4) genomics (study of the genome diversity and evolution of the adaptive immune response), (5) structural biology (evolution of the domains of the IgSF and MhSF proteins), (6) biotechnology related to antibody engineering and antibody humanization (construction and analysis of scFv, phage displays, combinatorial libraries, chimeric, humanized, and human antibodies), (7) diagnostics (clonalities, detection, and follow-up of minimal residual diseases), and (8) therapeutic approaches (grafts, immunotherapy, and vaccinology).

ACKNOWLEDGMENTS

We are grateful to Gérard Lefranc and to the IMGT team for their motivation and expertise.

REFERENCES

Alamyar E, Giudicelli V, Duroux P, Lefranc M-P. 2010. IMGT/HighV-QUEST: A high-throughput system and Web portal for the analysis of rearranged nucleotide sequences of antigen receptors - High-throughput version of IMGT/V-QUEST. JOBIM 2010 Poster 60. http://www.jobim2010.fr/?q=fr/node/55.

Baum TP, Pasqual N, Thuderoz F, Hierle V, Chaume D, Lefranc M-P, Jouvin-Marche E, Marche N, Demongeot J. 2004. IMGT/GeneInfo: Enhancing V(D)J recombination database accessibility. *Nucl Acids Res* **32**: D51–D54.

Baum TP, Hierle V, Pascal N, Bellahcene F, Chaume D, Lefranc M-P, Jouvin-Marche E, Marche PN, Demongeot J. 2006. IMGT/GeneInfo: T cell receptor gamma TRG and delta TRD genes in database give access to all TR potential V(D)J recombinations. *BMC Bioinformatics* **7**: 224. doi: 10.1186/1471-2105-7-224.

Bleakley K, Giudicelli V, Wu Y, Lefranc M-P, Biau G. 2006. IMGT standardization for statistical analyses of T cell receptor junctions: The TRAV-TRAJ example. *In Silico Biol* **6**: 573–588.

Bleakley K, Lefranc MP, Biau G. 2008. Recovering probabilities for nucleotide trimming processes for T cell receptor TRA and TRG V-J junctions analyzed with IMGT tools. *BMC Bioinformatics* **9**: 408. doi: 10.1186/1471-2105-9-408.

Brochet X. 2008. Conception et intégration d'un système d'information dédié à l'analyse et à la gestion des séquences réarrangées des récepteurs d'antigènes au sein d'IMGT: application à la Leucémie Lymphoïde Chronique, *PhD Thesis*, Université Montpellier 1, Montpellier, France.

Brochet X, Lefranc M-P, Giudicelli V. 2008. IMGT/V-QUEST: The highly customized and integrated system for IG and TR standardized V-J and V-D-J sequence analysis. *Nucl Acids Res* **36**: W503–W508.

Duroux P, Kaas Q, Brochet X, Lane J, Ginestoux C, Lefranc M-P, Giudicelli V. 2008. IMGT-Kaleidoscope, the Formal IMGT-ONTOLOGY paradigm. *Biochimie* **90**: 570–583.

Ehrenmann F, Lefranc M-P. 2011a. IMGT/DomainGapAlign: IMGT standardized analysis of amino acid sequences of variable, constant and groove domains (IG, TR, MH, IgSF, MhSF). *Cold Spring Harb Protoc* doi: 10.1101/pdb.prot5636.

Ehrenmann F, Lefranc M-P. 2011b. IMGT/3Dstructure-DB: Querying the IMGT database for 3D structures in immunology and immunoinformatics (IG or antibodies, TR, MH, RPI, and FPIA). *Cold Spring Harb Protoc* doi: 10.1101/pdb.prot5637.

Ehrenmann F, Duroux P, Giudicelli V, Lefranc M-P. 2010a. Standardized sequence and structure analysis of antibody using IMGT. In *Antibody engineering* 2nd ed. (ed. R Kontermann, S Dübel), Vol 2, pp. 11–31. Springer-Verlag, Berlin/Heidelberg, Germany.

Ehrenmann F, Kaas Q, Lefranc M-P. 2010b. IMGT/3Dstructure-DB and IMGT/DomainGapAlign: A database and a tool for immunoglobulins or antibodies, T cell receptors, MHC, IgSF and MhcSF. *Nucl Acids Res* **38**: D301–D307.

Ehrenmann F, Giudicelli V, Brochet X, Lefranc M-P. 2011. IMGT/Collier de Perles: IMGT standardized representation of domains (IG, TR, and IgSF variable and constant domains, MH and MhSF groove domains). *Cold Spring Harb Protoc* doi: 10.1101/pdb.prot5635.

Elemento O, Lefranc MP. 2003. IMGT/PhyloGene, an online software package for phylogenetic analysis of immunoglobulin and T cell receptor genes. *Dev Comp Immunol* **27**: 763–779.

Garapati VP, Lefranc M-P. 2007. IMGT Colliers de Perles and IgSF domain standardization for T cell costimulatory activatory (CD28, ICOS) and inhibitory (CTLA4, PDCD1 and BTLA) receptors. *Dev Comp Immunol* **31**: 1050–1072.

Ghia P, Stamatopoulos K, Belessi C, Moreno C, Stilgenbauer S, Stevenson F, Davi F, Rosenquist R, European Research Initiative on CLL (ERIC). 2007. ERIC recommendations on IGHV gene mutational status analysis in chronic lymphocytic leukemia. *Leukemia* **21**: 1–3.

Giudicelli V, Lefranc M-P. 1999. Ontology for Immunogenetics: The IMGTONTOLOGY. *Bioinformatics* **12**: 1047–1054.

Giudicelli V, Lefranc M-P. 2009. IMGT standardized analysis of immunoglobulin rearranged sequences. In *Immunoglobulin gene analysis in chronic lymphocytic leukemia* (ed. P Ghia, R Rosenquist, F Davi), pp. 33–52. Wolters Kluwer Health Italy Ltd, Milan, Italy.

Giudicelli V, Lefranc M-P. 2011. IMGT/JunctionAnalysis: IMGT standardized analysis of the V-J and V-D-J junctions of the rearranged immunoglobulins (IG) and T cell receptors (TR). *Cold Spring Harb Protoc* doi: 10.1101/pdb.prot5634.

Giudicelli V, Chaume D, Lefranc M-P. 2004. IMGT/V-QUEST, an integrated software for immunoglobulin and T cell receptor V-J and V-D-J rearrangement analysis. *Nucl Acids Res* **32**: W435–W440.

Giudicelli V, Chaume D, Lefranc M-P. 2005. IMGT/GENE-DB: A comprehensive database for human and mouse immunoglobulin and T cell receptor genes. *Nucl Acids Res* **33**: D256–D261.

Giudicelli V, Duroux P, Ginestoux C, Folch G, Jabado-Michaloud J, Chaume D, Lefranc M-P. 2006. IMGT/LIGM-DB, the IMGT comprehensive database of immunoglobulin and T cell receptor nucleotide sequences. *Nucl Acids Res* **34**: D781–D784.

Giudicelli V, Brochet X, Lefranc M-P. 2011. IMGT/V-QUEST: IMGT standardized analysis of the immunoglobulin (IG) and T cell receptor (TR) nucleotide sequences. *Cold Spring Harb Protoc* doi: 10.1101/pdb.prot5633.

Gruber TR. 1993. A translation approach to portable ontologies. *Knowledge Acquisition* **5**: 199–220.

Jefferis R, Lefranc M-P. 2009. Human immunoglobulin allotypes: Possible implications for immunogenicity. *MAbs* **1**: 332–338.

Kaas Q, Lefranc M-P. 2005. T cell receptor/peptide/MHC molecular characterization and standardized pMHC contact sites in IMGT/3Dstructure-DB. *In Silico Biol* **5**: 505–528.

Kaas Q, Lefranc M-P. 2007. IMGT Colliers de Perles: Standardized sequencestructure representations of the IgSF and MhcSF superfamily domains. *Current Bioinformatics* **2**: 21–30.

Kaas Q, Ruiz M, Lefranc M-P. 2004. IMGT/3Dstructure-DB and IMGT/StructuralQuery, a database and a tool for immunoglobulin, T cell receptor and MHC structural data. *Nucl Acids Res* **32**: D208–D210.

Kaas Q, Ehrenmann F, Lefranc M-P. 2007. IG, TR, MHC, IgSf and MhcSF: What do we learn from the IMGT Colliers de Perles? *Brief. Funct Genomic Proteomic* **6**: 253–264.

Kaas Q, Duprat E, Tourneur G, Lefranc M-P. 2008. IMGT standardization for molecular characterization of the T cell receptor/peptide/MHC complexes. In *Immunoinformatics* (ed. C Schoenbach, S Ranganathan, V Brusic), pp. 19–49. Immunomics Reviews, Series of Springer Science and Business Media LLC, Springer, New York.

Lane L, Duroux P, Lefranc MP. 2010. From IMGT-ONTOLOGY to IMGT/LIGMotif: The IMGT standardized approach for immunoglobulin and T cell receptor gene identification and description in large genomic sequences. *BMC Bioinformatics* **11**: 223. doi: 10.1186/1471-2105-11-223.

Lefranc M-P. 1997. Unique database numbering system for immunogenetic analysis. *Immunol Today* **18**: 509. doi: 10.1016/S0167-5699(97)01163-8.

Lefranc M-P. 1999. The IMGT unique numbering for Immunoglobulins, T cell receptors and Ig-like domains. *The Immunologist* **7**: 132–136.

Lefranc M-P. 2000a. Nomenclature of the human immunoglobin genes. In *Current protocols in immunology* (ed. JE Coligan et al.), pp. A.1P.1–A.1P.37. John Wiley and Sons Inc, Hoboken, NJ.

Lefranc M-P. 2000b. Nomenclature of the human T cell receptor genes. In *Current protocols in immunology* (ed. JE Coligan et al.), pp. A.1O.1–A.1O.23. John Wiley and Sons Inc, Hoboken, NJ.

Lefranc M-P. 2005. IMGT, the international ImMunoGeneTics information system: A standardized approach for immunogenetics and immunoinformatics. *Immunome Res* **1**: 3. doi: 10.1186/1745-7580-1-3.

Lefranc M-P. 2007. WHO-IUIS Nomenclature Subcommittee for immunoglobulins and T cell receptors report. *Immunogenetics* **59**: 899–902.

Lefranc M-P. 2008a. WHO-IUIS Nomenclature Subcommittee for immunoglobulins and T cell receptors report August 2007, 13th International Congress of Immunology, Rio de Janeiro, Brazil. *Dev Comp Immunol* **32**: 461–463.

Lefranc M-P. 2008b. IMGT, the international ImMunoGeneTics information system for immunoinformatics. Methods for querying IMGT databases, tools and Web resources in the context of immunoinformatics. *Mol Biotechnol* **40**: 101–111.

Lefranc M-P. 2008c. IMGT-ONTOLOGY, IMGT databases, tools and Web resources for Immunoinformatics. In *Immunoinformatics* (ed. C Schoenbach, S Ranganathan, V Brusic), pp. 1–18. Immunomics Reviews, Series of Springer Science and Business Media LLC, Springer, New York.

Lefranc M-P. 2009. Antibody database and tools: The IMGT experience. In *Therapeutic monoclonal antibodies: from Bench to Clinic* (ed. A Zhiqiang), pp. 91–114. John Wiley Sons, Inc, Hoboken, NJ.

Lefranc M-P. 2011a. From IMGT-ONTOLOGY IDENTIFICATION axiom to IMGT standardized keywords: For immunoglobulins (IG), T cell receptors (TR), and conventional genes. *Cold Spring Harb Protoc* doi: 10.1101/pdb.ip82.

Lefranc M-P. 2011b. From IMGT-ONTOLOGY DESCRIPTION axiom to IMGT standardized labels: For immunoglobulin (IG) and T cell receptor (TR) sequences and structures. *Cold Spring Harb Protoc* doi: 10.1101/pdb.ip83.

Lefranc M-P. 2011c. From IMGT-ONTOLOGY CLASSIFICATION axiom to IMGT standardized gene and allele nomenclature: For immunoglobulins (IG) and T cell receptors (TR). *Cold Spring Harb Protoc* doi: 10.1101/pdb.ip84.

Lefranc M-P. 2011d. IMGT unique numbering for the variable (V), constant (C), and groove (G) domains of IG, TR, MH, IgSF, and MhSF. *Cold Spring Harb Protoc* doi: 10.1101/pdb.ip85.

Lefranc M-P. 2011e. IMGT Collier de Perles for the Variable (V), Constant (C), and Groove (G) Domains of IG, TR, MH, IgSF, and MhSF. *Cold Spring Harb Protoc* doi: 10.1101/pdb.ip86.

Lefranc M-P. 2011. Antibody nomenclature: From IMGT-ONTOLOGY to INN definition. *MAbs* **3**: 1–2.

Lefranc M-P, Lefranc G. 2001a. *The Immunoglobulin FactsBook*, 1–458. Academic Press, London, UK.

Lefranc M-P, Lefranc G. 2001b. *The T cell receptor FactsBook*, 1–398. Academic Press, London, UK.

Lefranc M-P, Pommié C, Ruiz M, Giudicelli V, Foulquier E, Truong L, Thouvenin-Contet V, Lefranc G. 2003. IMGT unique numbering for immunoglobulin and T cell receptor variable domains and Ig superfamily V-like domains. *Dev Comp Immunol* **27**: 55–77.

Lefranc M-P, Giudicelli V, Ginestoux C, Bosc N, Folch G, Guiraudou D, Jabado-Michaloud J, Magris S, Scaviner D, Thouvenin V, et al. 2004. IMGT-ONTOLOGY for Immunogenetics and Immunoinformatics. *In Silico Biol* **4**: 17–29.

Lefranc M-P, Clément O, Kaas Q, Duprat E, Chastellan P, Coelho I, Combres K, Ginestoux C, Giudicelli V, Chaume D, et al. 2005a. IMGT-Choreography for Immunogenetics and Immunoinformatics. *In Silico Biol*, **5**: 45–60.

Lefranc M-P, Duprat E, Kaas Q, Tranne M, Thiriot A, Lefranc G. 2005b. IMGT unique numbering for MHC groove G-DOMAIN and MHC superfamily (MhcSF) G-LIKE-DOMAIN. *Dev Comp Immunol* **29**: 917–938.

Lefranc M-P, Pommié C, Kaas Q, Duprat E, Bosc N, Guiraudou D, Jean C, Ruiz M, Da Piedade I, Rouard M, et al. 2005c. IMGT unique numbering for immunoglobulin and T cell receptor constant domains and Ig superfamily C-like domains. *Dev Comp Immunol* **29**: 185–203.

Lefranc M-P, Giudicelli V, Regnier L, Duroux P. 2008. IMGT, a system and an ontology that bridge biological and computational spheres in bioinformatics. *Brief Bioinform* **9**: 263–275.

Lefranc M-P, Giudicelli V, Ginestoux C, Jabado-Michaloud J, Folch G, Bellahcene F, Wu Y, Gemrot E, Brochet X, Lane J, et al. 2009. IMGT, the international ImMunoGeneTics information system. *Nucl Acids Res* **37**: D1006–D1012.

Poiron C, Wu Y, Ginestoux C, Ehrenmann F, Duroux P, Lefranc M-P. 2010. IMGT/mAb-DB: The IMGT database for therapeutic monoclonal antibodies. JOBIM 2010 Poster 13. http://www.jobim2010.fr/?q=fr/node/55.

Pommié C, Levadoux S, Sabatier R, Lefranc G, Lefranc M-P. 2004. IMGT standardized criteria for statistical analysis of immunoglobulin V-REGION amino acid properties. *J Mol Recognit* **17**: 17–32.

Ruiz M, Lefranc M-P. 2002. IMGT gene identification and Colliers de Perles of human immunoglobulin with known 3D structures. *Immunogenetics* **53**: 857–883.

Yousfi Monod M, Giudicelli V, Chaume D, Lefranc M-P. 2004. IMGT/JunctionAnalysis: The first tool for the analysis of the immunoglobulin and T cell receptor complex V-J and V-D-J JUNCTIONs. *Bioinformatics* **20**: i379–i385.

Information Panel

From IMGT-ONTOLOGY IDENTIFICATION Axiom to IMGT Standardized Keywords: For Immunoglobulins (IG), T Cell Receptors (TR), and Conventional Genes

Marie-Paule Lefranc[1]

IMGT, the international ImMunoGeneTics information system, Laboratoire d'ImmunoGénétique Moléculaire LIGM, Université Montpellier 2, Institut de Génétique Humaine IGH, UPR CNRS 1142, 34396 Montpellier cedex 5, France

INTRODUCTION

IMGT standardized keywords have been defined for the identification of immunoglobulins (IG) and T cell receptors (TR) and conventional genes. These keywords are based on the concepts of identification, generated from the IDENTIFICATION axiom of the Formal IMGT-ONTOLOGY or IMGT-Kaleidoscope that postulates that any molecule, cell, tissue, organ, organism, or population, any process and any relation, has to be identified. Most of the IMGT standardized keywords used for the identification of sequences and structures are names of IMGT-ONTOLOGY leafconcepts (a leafconcept is a concept that corresponds to the finest level of granularity). IMGT standardized keywords have been essential for the data entries into the IMGT databases and tools, and particularly into IMGT/LIGM-DB, the first and the largest IMGT database that contains IG and TR nucleotide sequences and the translation of annotated sequences. IMGT standardized keywords for IG, TR, and conventional genes are defined in the context of the IMGT-ONTOLOGY concepts of identification.

RELATED INFORMATION

IMGT standardized keywords for sequences of IG, TR, and conventional genes that are used in IMGT/LIGM-DB (Giudicelli et al. 2006) are available at the database query page (IMGT Home page http://www.imgt.org). IMGT standardized keywords for amino acid sequences and structures are available in the IMGT Scientific chart at http://www.imgt.org/textes/IMGTScientificChart/SequenceDescription/IMGT3Dkeywords.html. IMGT-ONTOLOGY concepts of identification are available at the National Center for Biomedical Ontology (NCBO) BioPortal http://bioportal.bioontology.org/ and at IMGT http://www.imgt.org.

A detailed description of IMGT is provided in **IMGT, The International ImMunoGeneTics Information System** (Lefranc 2011a). Information is also available on **From IMGT-ONTOLOGY DESCRIPTION Axiom to IMGT Standardized Labels: For Immunoglobulin (IG) and T Cell Receptor (TR) Sequences and Structures** (Lefranc 2011b), **From IMGT-ONTOLOGY CLASSIFICATION Axiom to IMGT Standardized Gene and Allele Nomenclature: For Immunoglobulins (IG) and T Cell Receptors (TR)** (Lefranc 2011c), **IMGT Unique Numbering for the Variable (V), Constant (C), and Groove (G) Domains of IG, TR, MH, IgSF, and MhSF** (Lefranc 2011d), and **IMGT Collier de Perles for the Variable (V), Constant (C), and Groove (G) Domains of IG, TR, MH, IgSF, and MhSF** (Lefranc 2011e).

In addition, protocols are available for **IMGT/V-QUEST: IMGT Standardized Analysis of the Immunoglobulin (IG) and T Cell Receptor (TR) Nucleotide Sequences** (Giudicelli et al. 2011), **IMGT/JunctionAnalysis: IMGT Standardized Analysis of the V-J and V-D-J Junctions of the Rearranged Immunoglobulins (IG) and T Cell Receptors (TR)** (Giudicelli and Lefranc 2011), **IMGT/Collier de Perles: IMGT Standardized Representation of Domains (IG, TR, and IgSF Variable and Constant Domains, MH and MhSF Groove Domains)** (Ehrenmann et al. 2011), **IMGT/DomainGapAlign: IMGT Standardized Analysis of Amino Acid Sequences of Variable, Constant, and Groove Domains (IG, TR, MH, IgSF, MhSF)** (Ehrenmann and Lefranc 2011a), and

[1]Corresponding author (Marie-Paule.Lefranc@igh.cnrs.fr).
Cite as: Cold Spring Harb Protoc; 2011; doi:10.1101/pdb.ip82

IMGT/3Dstructure-DB: Querying the IMGT Database for 3D Structures in Immunology and Immunoinformatics (IG or Antibodies, TR, MH, RPI, and FPIA) (Ehrenmann and Lefranc 2011b).

BACKGROUND INFORMATION

The IDENTIFICATION axiom is an axiom of the Formal IMGT-ONTOLOGY or IMGT-Kaleidoscope (Giudicelli and Lefranc 1999; Lefranc et al. 2004, 2005a, 2008; Duroux et al. 2008). The IDENTIFICATION axiom postulates that any molecule, cell, tissue, organ, organism, or population, any process and any relation, has to be identified. The IDENTIFICATION axiom has generated the concepts of identification of IMGT-ONTOLOGY, the global reference in immunogenetics and immunoinformatics, built by IMGT, the international ImMunoGeneTics information system (http://www.imgt.org; Lefranc et al. 2009). The IDENTIFICATION axiom is one of seven axioms of the Formal IMGT-ONTOLOGY, the others being "CLASSIFICATION axiom," "DESCRIPTION axiom," "NUMEROTATION axiom," "LOCALIZATION axiom," "ORIENTATION axiom," and "OBTENTION axiom." Keywords used in the IMGT databases and tools mostly derive from the concepts of identification generated from the IDENTIFICATION axiom, and among them, most particularly the following: "MoleculeType," "GeneType," "ConfigurationType," "FunctionalityType," "Molecule_EntityType," "StructureType," "LocationType," "ChainType," "DomainType," and "ReceptorType."

CONCEPTS OF IDENTIFICATION

"MoleculeType"

"MoleculeType" is a concept of identification that allows one to identify the type of molecule, based on the type of the constitutive elements and on the concepts of obtention.

The "MoleculeType" concept comprises four major leafconcepts (keywords in the IMGT databases and tools):

1. **"gDNA"** identifies genomic DNA, a polynucleotide sequence made of a, t, c, g (representing the bases adenine [a], thymine [t] , cytosine [c], and guanine [g], attached to the 2-deoxyribose/phosphate backbone), obtained from a genome, or by extension, synthetic DNA having the characteristics of genomic DNA.
2. **"mRNA"** identifies messenger RNA, a polynucleotide sequence made of a, u, c, g (representing the bases adenine [a], uracil [u], cytosine [c], and guanine [g], attached to the ribose/phosphate backbone), and obtained by transcription of gDNA.
3. **"cDNA"** identifies complementary DNA, a polynucleotide sequence made of a, t, c, g (representing the bases adenine [a], thymine [t] , cytosine [c], and guanine [g], attached to the 2-deoxyribose/phosphate backbone), and obtained in vitro by reverse transcription of mRNA.
4. **"protein"** identifies a sequence made of amino acids, obtained by translation of mRNA, or by in vitro synthesis.

"GeneType"

"GeneType" is a concept of identification that allows one to identify the type of gene.

The "GeneType" concept comprises six leafconcepts (keywords in the IMGT databases and tools):

- Two leafconcepts **"conventional-with-leader"** and **"conventional-without-leader"** identify "conventional" genes, that is any (coding or noncoding) gene other than the IG or TR genes, with or without leader L region (or signal peptide), respectively. A conventional gene has by definition the characteristics of a classical gene (initiation codon and stop codon in the same gene unit).
- Four leafconcepts identify the IG and TR genes of the vertebrate adaptive immune response and are specific to immunogenetics. Three of these leafconcepts, **"variable"** (V), **"diversity"** (D), and **"joining"** (J), identify the IG and TR genes that rearrange at the DNA level in the B and T cells and code the V, D, and J regions, respectively, of the IG and TR variable domains; the fourth leafconcept, **"constant"** (C), identifies the IG and TR genes that code the C region of the IG and TR chains (Lefranc and Lefranc 2001a,b). IG and TR genes have special characteristics compared to the conventional

genes: They have a recombination signal (RS) in 3′ (for the V genes), in 5′ and 3′ (for the D genes), and in 5′ (for the J genes) and require DNA rearrangements (V-J or V-D-J) during the biosynthesis of the IG and TR chains (Lefranc and Lefranc 2001a,b).

"ConfigurationType"

"ConfigurationType" is a concept of identification that allows one to identify the type of configuration of a gene, and by extension, the type of configuration of the Molecule_EntityType leafconcepts that contain it.

The "ConfigurationType" concept comprises four leafconcepts (keywords in the IMGT databases and tools):

1. **"undefined"** identifies, whatever the molecule type, the configuration of the conventional genes and that of the IG and TR C genes (Lefranc and Lefranc 2001a,b), and by extension, the configuration of the Molecule_EntityType leafconcepts that only contain genes in undefined configuration.
2. **"germline"** identifies, whatever the molecule type, the configuration of the IG and TR V, D, and J genes **before** DNA rearrangements (Lefranc and Lefranc 2001a,b), and by extension, the configuration of the Molecule_EntityType leafconcepts that contain germline genes (with or without C genes in undefined configuration).
3. **"rearranged"** identifies, whatever the molecule type, the configuration of the IG and TR V, D, and J genes **after** DNA rearrangements, and by extension, the configuration of the Molecule_EntityType leafconcepts that contain rearranged genes with, if present, completely rearranged D genes (with or without C genes in undefined configuration).
4. **"partially-rearranged"** identifies, whatever the molecule type, the configuration of partially rearranged IG or TR D genes, and by extension, the configuration of Molecule_EntityType leafconcepts that contain at least one partially rearranged D gene (with another D or with rearranged V or J genes [with or without C genes in undefined configuration]).

"FunctionalityType"

"FunctionalityType" is a concept of identification that allows one to identify, whatever the molecule type (gDNA, cDNA, mRNA, or protein), the type of functionality of a Molecule_EntityType leafconcept (Giudicelli and Lefranc 1999; Lefranc et al. 2004, 2005a, 2008; Duroux et al. 2008). The "FunctionalityType" concept comprises five leafconcepts (keywords in the IMGT databases and tools), divided into two categories, according to the configuration type of the Molecule_EntityType leafconcept:

Three leafconcepts, **"functional"**, **"ORF"** (open reading frame), and **"pseudogene,"** identify the functionality of Molecule_EntityType leafconcepts in **undefined** configuration (conventional genes and IG and TR C genes), or in **germline** configuration (IG and TR V, D, and J genes **before** DNA rearrangements).

Two leafconcepts, **"productive"** and **"unproductive"**, identify the functionality of Molecule_EntityType leafconcepts in **rearranged** or **partially-rearranged** configuration (IG and TR entities **after** DNA rearrangements, and by extension, fusion entities resulting from translocations, and hybrid entities obtained by biotechnology molecular engineering).

Thus:

1. **"functional"** identifies, whatever the molecule type, the functionality of Molecule_EntityType leafconcepts in **undefined** or **germline** configuration, whose coding region has an ORF without stop codon, and for which there is no described defect in the splicing sites, recombination signals, and/or regulatory elements.
2. **"ORF"** identifies, whatever the molecule type, the functionality of Molecule_EntityType leafconcepts in **undefined** or **germline** configuration, whose coding region has an ORF:
    - but alterations have been described in the splicing sites, recombination signals, and/or regulatory elements.
    - and/or changes of conserved amino acids have been suggested by the authors to lead to incorrect folding.
    - and/or the entity is an orphon.

3. **"Pseudogene"** identifies, whatever the molecule type, the functionality of Molecule_EntityType leafconcepts in **undefined** or **germline** configuration, whose coding region has stop codon(s) and/or frameshift mutation(s), and/or for which a mutation affects the initiation codon (for a conventional or a V gene).

4. **"Productive"** identifies, whatever the molecule type, the functionality of Molecule_EntityType leafconcepts in **rearranged** or **partially-rearranged** configuration, whose coding region has an ORF without stop codon, whose junction is in-frame (for IG and TR), and for which there is no described defect in the initiation codon, splicing sites, and/or regulatory elements.

5. **"Unproductive"** identifies, whatever the molecule type, the functionality of Molecule_EntityType leafconcepts in **rearranged** or **partially-rearranged** configuration, whose coding region has stop codon(s) and/or frameshift mutation(s), and/or whose junction is out-of-frame (for IG and TR), and/or for which a mutation affects the initiation codon, and/or for which there are defects in the splicing sites and/or in the regulatory element(s), and/or there are unusual features (e.g., translocation, gene fusion), and/or changes of conserved amino acids demonstrated as leading to incorrect folding.

"Molecule_EntityType"

"Molecule_EntityType" is a major concept of identification that allows one to identify the molecule entity type. The "Molecule_EntityType" concept is defined by the "MoleculeType," "GeneType," and "ConfigurationType" concepts and has properties identified in the "FunctionalityType" and "StructureType" concepts (Fig. 1).

The "Molecule_EntityType" concept comprises 38 leafconcepts (keywords in the IMGT databases and tools) that, based on the "MoleculeType" ("gDNA", "mRNA", "cDNA," or "protein"), identify the molecule entities of the four major "MoleculeUnit" leafconcepts that are genes (10), transcripts (11), cDNA sequences (11), and chains (6) (as indicated by the suffix).

The 10 leafconcepts that are the most classical representatives of IG and TR data identification in the IMGT information system are shown as examples (Table 1). They are also illustrated in Immunoglobulin synthesis (Fig. 2).

"V-gene" identifies, for gDNA, molecule entities with a germline V gene. V-gene is in germline configuration.

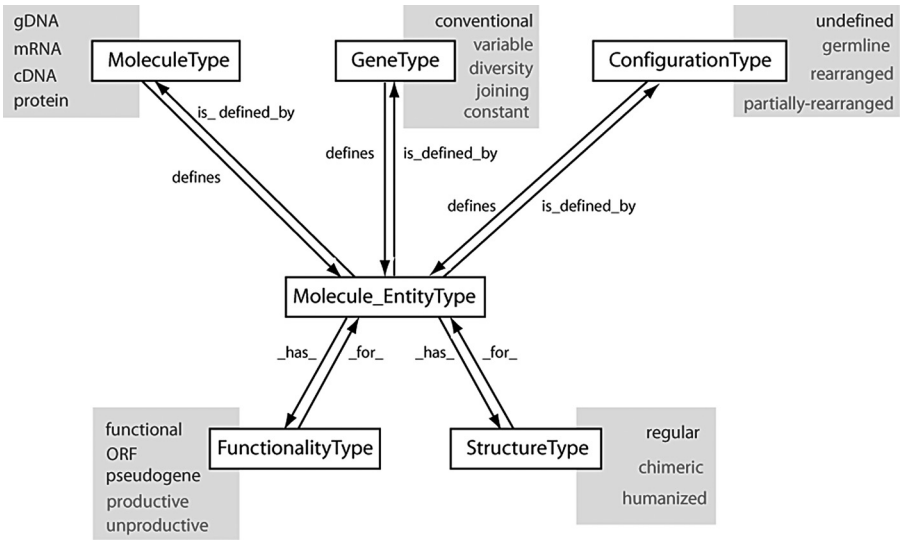


FIGURE 1. The "Molecule_EntityType" concept. The "Molecule_EntityType" concept is defined by the "MoleculeType," "GeneType," and "ConfigurationType" concepts of identification and has properties identified in the "FunctionalityType" and "StructureType" concepts (IDENTIFICATION axiom). Arrows indicate reciprocal relations "is_defined_by" and "defines," "_has_" and "_for_." Leafconcepts are general (in blue) or specific of the IG and TR (in red) The "Molecule_EntityType" concept has 38 leafconcepts (or keywords in the IMGT databases and tools). Only a few examples of the "StructureType" leafconcepts (or keywords in the IMGT databases and tools) are shown.

Table 1. Molecule_EntityType concept

Concepts	Leafconcepts	Molecule_EntityType concept									
	Keywords in the IMGT databases and tools	V-gene	D-gene	J-gene	C-gene	V-J-gene	V-D-J-gene	L-V-J-C-sequence	L-V-D-J-C-sequence	V-J-C-chain	V-D-J-C-chain
MoleculeType	gDNA	x	x	x	x	x	x				
	cDNA							x	x		
	protein									x	x
GeneType	variable	x								x	x
	diversity		x				x		x		x
	joining			x		x	x	x	x	x	x
	constant				x			x	x	x	x
ConfigurationType	undefined				x						
	germline	x	x	x							
	rearranged					x	x	x	x	x	x
FunctionalityType	functional	x	x	x	x						
	ORF	x	x	x	x						
	pseudogene	x	x	x	x						
	productive					x	x	x	x	x	x
	unproductive					x	x	x	x	x	x

Ten leafconcepts (keywords in the IMGT databases and tools) are shown as examples. An "x" indicates the leafconcepts (keywords in the IMGT databases and tools) that define them ("MoleculeType," "GeneType," "ConfigurationType") and their possible functionality ("FunctionalityType").

Abbreviations: gDNA, genomic DNA; cDNA, complementary DNA; ORF, open reading frame.

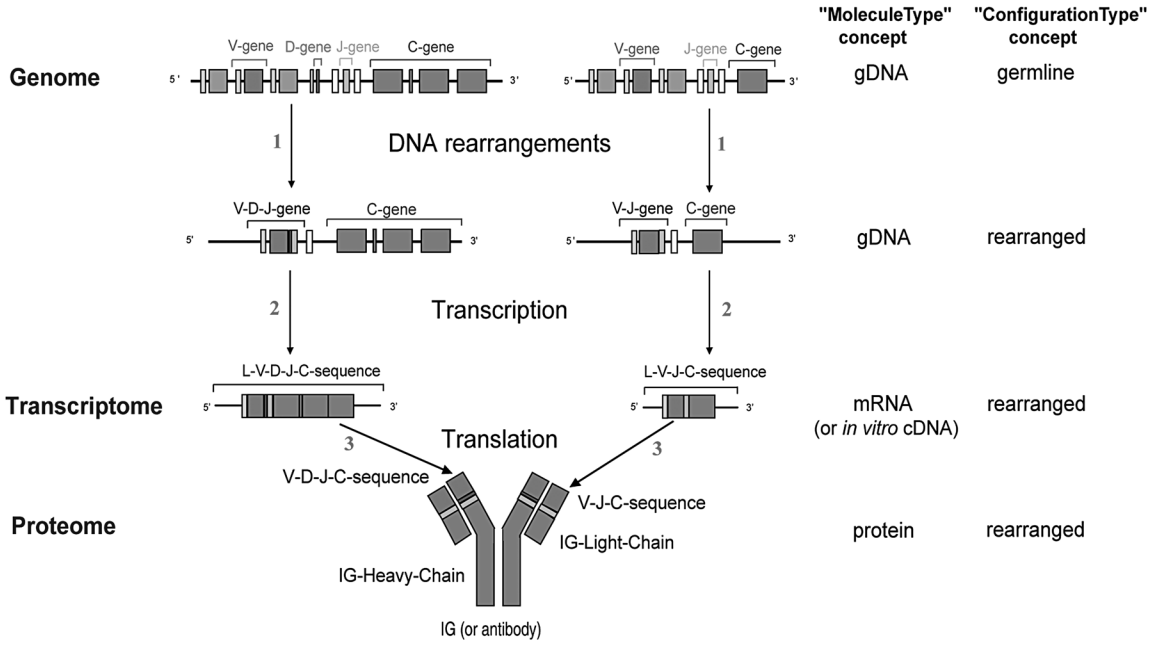


FIGURE 2. Synthesis of an IG or antibody in humans. A human being may potentially synthesize 10^{12} different antibodies (Lefranc and Lefranc 2001a). 1, DNA rearrangements (is_rearranged_into); 2, Transcription (is_transcribed_into); 3, Translation (is_translated_into). The configuration of C-GENE is "undefined" (ConfigurationType) (IMGT Repertoire, http://www.imgt.org).

"**D-gene**" identifies, for gDNA, molecule entities with a germline D gene. D-gene is in germline configuration.

"**J-gene**" identifies, for gDNA, molecule entities with a germline J gene. J-gene is in germline configuration.

"**C-gene**" identifies, for gDNA, molecule entities with a C gene in undefined configuration. C-gene is in undefined configuration.

"**V-J-gene**" identifies, for gDNA, molecule entities with a rearranged V gene and a rearranged J gene. V-J-gene is in rearranged configuration.

"**V-D-J-gene**" identifies, for gDNA, molecule entities with a rearranged V gene, at least one rearranged D gene, and a rearranged J gene. V-D-J-gene is in rearranged configuration.

"**L-V-J-C-sequence**" identifies, for cDNA, molecule entities with a leader L region, a rearranged V region, a rearranged J region, and a C region in undefined configuration. L-V-J-C-sequence is in rearranged configuration.

"**L-V-D-J-C-sequence**" identifies, for cDNA, molecule entities with a leader L region, a rearranged V region, at least one rearranged D region, a rearranged J region, and a C region in undefined configuration. L-V-D-J-C-sequence is in rearranged configuration.

"**V-J-C-chain**" identifies, for protein, molecule entities without a leader L region (mature form) and with a rearranged V region, a rearranged J region, and a C region in undefined configuration. V-J-C-chain is in rearranged configuration.

"**V-D-J-C-chain**" identifies, for protein, molecule entities without a leader L region (mature form) and with a rearranged V region, at least one rearranged D region, a rearranged J region, and a C region in undefined configuration. V-D-J-C-chain is in rearranged configuration.

The 28 other leafconcepts (keywords in the IMGT databases and tools) identify sterile transcripts, partially rearranged entities, partial sequences, or other molecule units, and are not detailed here.

"StructureType"

"**StructureType**" is a concept of identification that allows one to identify, whatever the molecule type (gDNA, cDNA, mRNA, protein), the type of structure of Molecule_EntityType leafconcepts.

The "StructureType" concept comprises 17 leafconcepts (keywords in the IMGT databases and tools) that identify structures with a classical organization ("regular"), from those which have been modified either naturally in vivo ("processed", "unprocessed", "partially-processed", "sterile-transcript", "unspliced", "partially-spliced", "spliced", "membrane", "secreted", "membrane-and-secreted", "mature-form", "immature-form"), or artificially in vitro ("engineered", "humanized", "chimeric", "fusion") ("chimeric" and "fusion" can also be obtained in vivo) (Giudicelli and Lefranc 1999; Lefranc et al. 2004, 2005a, 2008; Duroux et al. 2008).

"LocationType"

"**LocationType**" is a concept of identification that allows one to identify, whatever the molecule type (gDNA, cDNA, mRNA, protein), the Molecule_EntityType leafconcepts, based on their location.

The "LocationType" concept comprises, for instance, the following leafconcepts (keywords in the IMGT databases and tools):

"**orphon**" identifies, whatever the molecule type, a gene that is found in vivo in a different locus from the main locus (either on the same chromosome or on another chromosome).

"**transgene**" identifies, whatever the molecule type, a gene that is artificially introduced into a multicellular organism (e.g., mouse, plant).

"**translocated**" identifies, whatever the molecule type, a gene that results from a translocation (in vivo).

"**transposed**" identifies, whatever the molecule type, a transgene or a retrotransposon that is permanently inserted in a chromosome.

"ChainType"

"**ChainType**" is a concept of identification that allows one to identify the type of chain. It is one of the most important concepts of identification for the standardization of genome, transcriptome, and proteome data in systems biology. Indeed, being able to identify a type of chain means that it is possible to identify the transcript and the encoding gene(s).

The "ChainType" concept is defined by the "Molecule_EntityType" and the "DomainType" concepts of identification and by concepts of classification (**From IMGT-ONTOLOGY CLASSIFICATION Axiom to IMGT Standardized Gene and Allele Nomenclature: For Immunoglobulins (IG) and T Cell Receptors (TR)** [Lefranc 2011c]). The "ChainType" concept is a "highconcept" as it corresponds to a hierarchy of concepts which allow one to identify the chain type at different levels of granularity (up to five for "ChainType").

The finest level of granularity, the "**GeneLevelChainType**" concept, identifies the type of chain by reference to the gene(s) which code(s) the chain. It represents the main concept for a very precise identification because it establishes a relationship with "**Gene**" (concept of classification generated from the CLASSIFICATION axiom) (the reciprocal relations are: "is_coded_by" and "codes").

Thus:

- A *Homo sapiens* IG-Heavy-Gamma-1-Chain is defined by a leafconcept of the "Molecule_EntityType" (V-D-J-C-chain), "DomainType" and numbers (1 V domain, 3 C domains), and the gene encoding the constant region (*Homo sapiens* IGHG1) (Fig. 3).
- Similarly, a *Homo sapiens* IG-Kappa chain is defined by a leafconcept of the "Molecule_EntityType" (V-J-C-chain), "DomainType" and numbers (1 V domain, 1 C domain), and the gene encoding the constant region (*Homo sapiens* IGKC).

The number of instances of the leafconcepts of the "GeneLevelChainType" concept (and therefore of the keywords in the IMGT databases and tools) depends on the number of functional genes and ORF ("FunctionalityType"), per haploid genome, in a given species (in the case of the IG and TR genes, it is the number of functional and ORF C genes which is taken into account). If only the functional genes are considered, the instances of this leafconcept correspond to the isotypes.

As the "ChainType" concept can be used by extrapolation for any of the four "MoleculeUnit" (gene, transcript, cDNA sequence, chain), in order to facilitate query and/or data extraction from the IMGT/LIGM-DB database (Giudicelli et al. 2006), the suffix -Chain (used in "ChainType" leafconcepts) is not specified in the "ChainType" keywords.

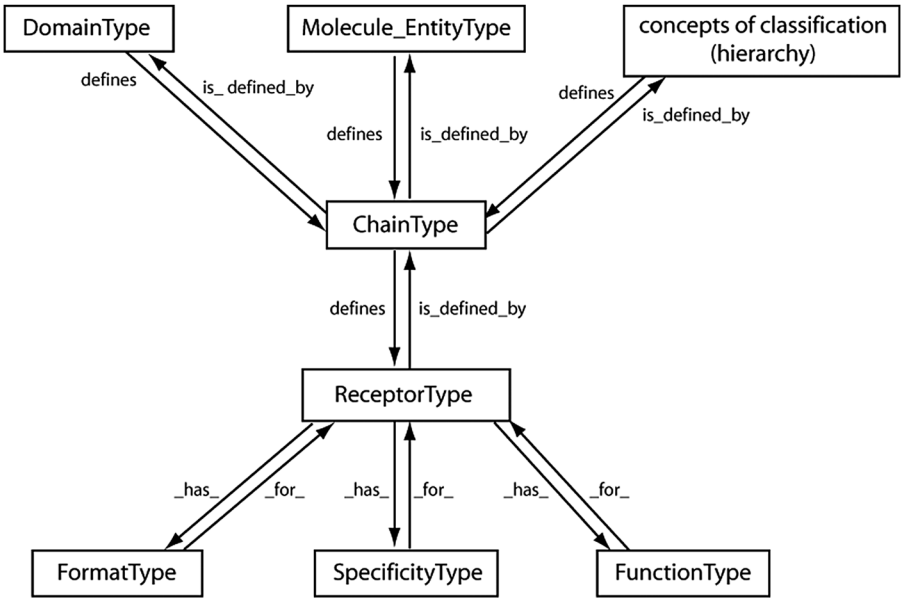


FIGURE 3. The "ReceptorType" concept. The "ReceptorType" concept is defined by the "ChainType" concept of identification and has properties identified in the "FormatType," "SpecificityType," and "FunctionType" concepts (IDENTIFICATION axiom). The "ChainType" concept is itself defined by the "Molecule_EntityType" and "DomainType" concepts and by concepts of classification organized in a hierarchy (see CLASSIFICATION axiom). Arrows indicate reciprocal relations "is_defined_by" and "defines," "_has_" and "_for_." The "ChainType" and "ReceptorType" concepts have different levels of granularity (up to five) and are highconcepts.

"DomainType"

"DomainType" is a concept of identification that allows one to identify the type of domain.

A domain is a chain subunit characterized by its three-dimensional (3D) structure, and by extension its amino acid sequence and the nucleotide sequence which encodes it.

In IMGT-ONTOLOGY, the "DomainType" concept has three major leafconcepts that were first described in the proteins of the adaptive immune response, IG or antibodies, TR and MH, and then extended to the proteins of the IgSF and MhSF (Lefranc 1997, 1999; Lefranc et al. 2003, 2005b,c). These leafconcepts (keywords in the IMGT databases and tools) are:

the **"V domain"** (Variable [V] domain) that includes the variable domains of the IG and TR and the V-like domains of other IgSF proteins; (Lefranc 1999; Lefranc et al. 2003)

the **"C domain"** (Constant [C] domain) that includes the constant domains of the IG and TR and the C-like domains of other IgSF proteins; (Lefranc et al. 2005c)

the **"G domain"** (Groove [G] domain) that includes the groove domains of the MH and the G-like domains of other MhSF proteins (Lafrnac et al. 2005b).

Standardized description of the V, C, and G domains is based on the IMGT unique numbering (Lafranc 1997, 1999; Lefranc et al. 2003, 2005b,c). Graphical two-dimensional (2D) representations or IMGT Colliers de Perles (Ruiz and Lefranc 2002; Kaas and Lefranc 2007; Kaas et al. 2007) can be obtained using the IMGT/DomainGapAlign and IMGT/Collier de Perles tools (Ehrenmann et al. 2010a,b; http://www.imgt.org).

"ReceptorType"

"ReceptorType" is a concept that allows one to identify the type of receptor, based on the ChainType leafconcept(s) that identify the associated chains. "ReceptorType" is a "highconcept" as it corresponds to a hierarchy of concepts (depending itself on the "ChainType" hierarchy) which allows one to identify the receptor type at different levels of granularity (up to five for "ReceptorType"). The

"ReceptorType" concept is defined by the "ChainType" concept of identification and has properties identified in the "FormatType," "SpecificityType," and "FunctionType" concepts (Fig. 3; Duroux et al. 2008; Lefranc et al. 2008).

Thus, "IG" identifies an IG-Heavy-Chain_IG-Light-Chain dimer, whereas at a different granularity ("GeneLevelChainType"), the same receptor will be identified as an IG-Heavy-Gamma-1-Chain_Light-Kappa-Chain dimer.

IG and TR are the specific antigen receptors of the adaptive immune response (and therefore major leafconcepts of the "ReceptorType" concept) characterized by their function of binding and recognition of the antigen ("**FunctionType**"). The "**SpecificityType**" concept allows an amino acid characterization of this binding.

The "**SpecificityType**" concept has two major leafconcepts (keywords in the IMGT databases and tools):

"**paratope**" identifies the amino acids of the antigen receptor V domains that recognize and bind the antigen (Ag).

"**epitope**" identifies the antigen part recognized and bound by the antigen receptor (IG or TR).

The "**FormatType**" concept has many leafconcepts (keywords in the IMGT databases and tools) that identify the different types of format of the receptors, and particularly those of the antibodies (Fab, F(ab′)2, Fc, scFv, Fv…) and of the fusion proteins for immune applications (FPIA) (Lefranc 2011f).

The corresponding keywords are extensively used in the identification of the IMGT/2Dstructure-DB and IMGT/3Dstructure-DB entries (Lefranc 2008, 2009; Lefranc et al. 2008; Ehrenmann et al. 2010a,b) and of the IMGT/mAb-DB therapeutic mAbs and FPIA (Poiron et al. 2010). Paratope/epitope are used in IMGT/3Dstructure-DB (http://www.imgt.org) for analysis of interactions in IG/Ag and TR/peptide/major histocompatibility (TR/pMH) complexes (Kaas and Lefranc 2005; Kaas et al. 2008).

ACKNOWLEDGMENTS

We thank Véronique Giudicelli, Fatena Bellahcene, Géraldine Folch, Joumana Jabado-Michaloud, Nelly Jouffre, Claire Poiron, Laetitia Regnier, and Gérard Lefranc, and the IMGT team for their motivation and expertise.

REFERENCES

Duroux P, Kaas Q, Brochet X, Lane J, Ginestoux C, Lefranc M-P, Giudicelli V. 2008. IMGT-Kaleidoscope, the Formal IMGT-ONTOLOGY paradigm. *Biochimie* **90:** 570–583.

Ehrenmann F, Lefranc M-P. 2011a. IMGT/DomainGapAlign: IMGT standardized analysis of amino acid sequences of variable, constant, and groove domains (IG, TR, MH, IgSF, MhSF). *Cold Spring Harb Protoc* doi: 10.1101/pdb.prot5636.

Ehrenmann F, Lefranc M-P. 2011b. IMGT/3Dstructure-DB: Querying the IMGT database for 3D structures in immunology and immunoinformatics (IG or antibodies, TR, MH, RPI, and FPIA). *Cold Spring Harb Protoc* doi: 10.1101/pdb.prot5637.

Ehrenmann F, Duroux P, Giudicelli V, Lefranc M-P. 2010a. Standardized sequence and structure analysis of antibody using IMGT. In *Antibody engineering*, 2nd ed. (ed. R Kontermann, S Dübel), Vol 2, pp. 11–31. Springer-Verlag, Berlin/Heidelberg.

Ehrenmann F, Kaas Q, Lefranc M-P. 2010b. IMGT/3Dstructure-DB and IMGT/DomainGapAlign: A database and a tool for immunoglobulins or antibodies, T cell receptors, MHC, IgSF and MhcSF. *Nucl Acids Res* **38:** D301–D307.

Ehrenmann F, Giudicelli V, Brochet X, Lefranc M-P. 2011. IMGT/Collier de Perles: IMGT standardized representation of domains (IG, TR, and IgSF variable and constant domains, MH and MhSF groove domains). *Cold Spring Harb Protoc* doi: 10.1101/pdb.prot5635.

Giudicelli V, Lefranc M-P. 1999. Ontology for Immunogenetics: The IMGT-ONTOLOGY. *Bioinformatics* **12:** 1047–1054.

Giudicelli V, Lefranc M-P. 2011. IMGT/JunctionAnalysis: IMGT standardized analysis of the V-J and V-D-J junctions of the rearranged immunoglobulins (IG) and T cell receptors (TR). *Cold Spring Harb Protoc* doi: 10.1101/pdb.prot5634.

Giudicelli V, Duroux P, Ginestoux C, Folch G, Jabado-Michaloud J, Chaume D, Lefranc M-P. 2006. IMGT/LIGM-DB, the IMGT comprehensive database of immunoglobulin and T cell receptor nucleotide sequences. *Nucl Acids Res* **34:** D781–D784.

Giudicelli V, Brochet X, Lefranc M-P. 2011. IMGT/V-QUEST: IMGT standardized analysis of the immunoglobulin (IG) and T cell receptor (TR) nucleotide sequences. *Cold Spring Harb Protoc* doi: 10.1101/pdb.prot5633.

Kaas Q, Lefranc M-P. 2005. T cell receptor/peptide/MHC molecular characterization and standardized pMHC contact sites in IMGT/3Dstructure-DB. *In Silico Biol* **5:** 505–528.

Kaas Q, Lefranc M-P. 2007. IMGT Colliers de Perles: Standardized sequence-structure representations of the IgSF and MhcSF superfamily domains. *Curr Bioinform* **2:** 21–30.

Kaas Q, Ehrenmann F, Lefranc M-P. 2007. IG, TR, MHC, IgSf and MhcSF: What do we learn from the IMGT Colliers de Perles?. *Brief Funct Genomic Proteomic* **6:** 253–264.

Kaas Q, Duprat E, Tourneur G, Lefranc M-P. 2008. IMGT standardization for molecular characterization of the T cell receptor/peptide/MHC complexes. In *Immunoinformatics* (ed. C Schoenbach et al.), pp. 19–49. Immunomics Reviews, Series of Springer Science and Business Media LLC, Springer, New York.

Lefranc M-P. 1997. Unique database numbering system for immunogenetic analysis. *Immunol Today* **18:** 509.

Lefranc M-P. 1999. The IMGT unique numbering for immunoglobulins, T cell receptors and Ig-like domains. *The Immunologist* **7:** 132–136.

Lefranc M-P. 2008. IMGT-ONTOLOGY, IMGT databases, tools and Web resources for Immunoinformatics. In *Immunoinformatics* (ed. C Schoenbach et al.), pp. 1–18. Immunomics Reviews,

Series of Springer Science and Business Media LLC, Springer, New York.

Lefranc M-P. 2009. Antibody database and tools: The IMGT experience. In *Therapeutic monoclonal antibodies: From bench to clinic* (ed. A Zhiqiang), pp. 91–114. John Wiley and Sons Inc, Hoboken, NJ.

Lefranc M-P. 2011a. IMGT, the international ImMunoGeneTics information system. *Cold Spring Harb Protoc* doi: 10.1101/pdb.top115.

Lefranc M-P. 2011b. From IMGT-ONTOLOGY DESCRIPTION axiom to IMGT standardized labels: For immunoglobulin (IG) and T cell receptor (TR) sequences and structures. *Cold Spring Harb Protoc* doi: 10.1101/pdb.ip83.

Lefranc M-P. 2011c. From IMGT-ONTOLOGY CLASSIFICATION axiom to IMGT standardized gene and allele nomenclature: For immunoglobulins (IG) and T cell receptors (TR). *Cold Spring Harb Protoc* doi: 10.1101/pdb.ip84.

Lefranc M-P. 2011d. IMGT unique numbering for the variable (V), constant (C), and groove (G) domains of IG, TR, MH, IgSF, and MhSF. *Cold Spring Harb Protoc* doi: 10.1101/pdb.ip85.

Lefranc M-P. 2011e. IMGT Collier de Perles for the variable (V), constant (C), and groove (G) domains of IG, TR, MH, IgSF, and MhSF. *Cold Spring Harb Protoc* doi: 10.1101/pdb.ip86.

Lefranc M-P. 2011f. Antibody nomenclature: From IMGT-ONTOLOGY to INN definitions. *MAbs* **3:** 1–2.

Lefranc M-P, Lefranc G. 2001a. *The immunoglobulin FactsBook*. Academic Press, London.

Lefranc M-P, Lefranc G. 2001b. *The T cell receptor FactsBook*. Academic Press, London.

Lefranc M-P, Pommié C, Ruiz M, Giudicelli V, Foulquier E, Truong L, Thouvenin-Contet V, Lefranc G. 2003. IMGT unique numbering for immunoglobulin and T cell receptor variable domains and Ig superfamily V-like domains. *Dev Comp Immunol* **27:** 55–77.

Lefranc M-P, Giudicelli V, Ginestoux C, Bosc N, Folch G, Guiraudou D, Jabado-Michaloud J, Magris S, Scaviner D, Thouvenin V, et al. 2004. IMGT-ONTOLOGY for Immunogenetics and Immunoinformatics. *In Silico Biol* **4:** 17–29.

Lefranc M-P, Clément O, Kaas Q, Duprat E, Chastellan P, Coelho I, Combres K, Ginestoux C, Giudicelli V, Chaume D, et al. 2005a. IMGT-Choreography for immunogenetics and immunoinformatics. *In Silico Biol* **5:** 45–60.

Lefranc M-P, Duprat E, Kaas Q, Tranne M, Thiriot A, Lefranc G. 2005b. IMGT unique numbering for MHC groove G-DOMAIN and MHC superfamily (MhcSF) G-LIKE-DOMAIN. *Dev Comp Immunol* **29:** 917–938.

Lefranc M-P, Pommié C, Kaas Q, Duprat E, Bosc N, Guiraudou D, Jean C, Ruiz M, Da Piedade I, Rouard M, et al. 2005c. IMGT unique numbering for immunoglobulin and T cell receptor constant domains and Ig superfamily C-like domains. *Dev Comp Immunol* **29:** 185–203.

Lefranc M-P, Giudicelli V, Regnier L, Duroux P. 2008. IMGT, a system and an ontology that bridge biological and computational spheres in bioinformatics. *Brief Bioinform* **9:** 263–275.

Lefranc M-P, Giudicelli V, Ginestoux C, Jabado-Michaloud J, Folch G, Bellahcene F, Wu Y, Gemrot E, Brochet X, Lane J, et al. 2009. IMGT, the international ImMunoGeneTics information system. *Nucl Acids Res* **37:** D1006–D1012.

Poiron C, Wu Y, Ginestoux C, Ehrenmann F, Duroux P, Lefranc M-P. 2010. IMGT/mAb-DB: The IMGT database for therapeutic monoclonal antibodies. JOBIM 2010 Poster 13. http://www.jobim2010.fr/?q=fr/node/55.

Ruiz M, Lefranc M-P. 2002. IMGT gene identification and Colliers de Perles of human immunoglobulin with known 3D structures. *Immunogenetics* **53:** 857–883.

Information Panel

From IMGT-ONTOLOGY DESCRIPTION Axiom to IMGT Standardized Labels: For Immunoglobulin (IG) and T Cell Receptor (TR) Sequences and Structures

Marie-Paule Lefranc[1]

IMGT, the international ImMunoGeneTics information system, Laboratoire d'ImmunoGénétique Moléculaire LIGM, Université Montpellier 2, Institut de Génétique Humaine IGH, UPR CNRS 1142, 34396 Montpellier cedex 5, France

INTRODUCTION

More than 560 IMGT standardized labels have been defined for the description of immunoglobulin (IG) and T cell receptor (TR) sequences and structures. These labels, as detailed here, are based on the concepts of description (generated from the DESCRIPTION axiom) of IMGT-ONTOLOGY, the global reference in immunogenetics and immunoinformatics, built by IMGT, the international ImMunoGeneTics information system. IMGT labels for nucleotide sequences correspond to the "Molecule_Entity-Prototype" concept with its leafconcepts (a leafconcept is the finest level of granularity), its associated concepts of description, and its prototypes or graphical representations. These labels have been essential for the IG and TR description in the IMGT sequence and gene databases (IMGT/LIGM-DB, IMGT/GENE-DB) and tools (IMGT/V-QUEST, IMGT/JunctionAnalysis, IMGT/HighV-QUEST, IMGT/LIGMotif). IMGT labels for amino acid sequences and structures correspond to the "RECEPTOR," "CHAIN," and "DOMAIN" concepts. These labels have been crucial for the IG and TR description in the IMGT two-dimensional (2D) and three-dimensional (3D) databases (IMGT/2Dstructure-DB, IMGT/3Dstructure-DB) and tools (IMGT/DomainDisplay, IMGT/DomainGapAlign). As all leafconcepts of description are directly translated into IMGT standardized labels for use in the IMGT databases and tools, IMGT-ONTOLOGY concepts of description contribute to data coherence and consistency and facilitate the interoperability between the different components of IMGT, bridging the description of sequences and structures.

RELATED INFORMATION

IMGT standardized labels for nucleotide sequences of IG and TR are available at the IMGT/LIGM-DB (Giudicelli et al. 2006) database query page (IMGT Home page http://www.imgt.org). IMGT standardized labels for amino acid sequences and structures (IMGT/2Dstructure-DB, IMGT/3Dstructure-DB [Kaas et al. 2004; Ehrenmann et al. 2010b]) are available in the IMGT Scientific chart at: http://www.imgt.org/textes/IMGTScientificChart/SequenceDescription/IMGT3Dkeywords.html. These labels describe the amino acid sequences and the 2D and 3D structure organization (in receptors, chains, and domains) of IG, TR, major histocompatibility (MH), related proteins of the immune system (RPI), and fusion proteins for immune applications (FPIA) (Lefranc 2011f). The definition of these labels is available in the IMGT Scientific chart at: http://www.imgt.org/textes/IMGTScientificChart/SequenceDescription/IMGT3Dlabeldef.html.

A detailed description of IMGT is provided in **IMGT, The International ImMunoGeneTics Information System** (Lefranc 2011a). Information is also available on **From IMGT-ONTOLOGY IDENTIFICATION Axiom to IMGT Standardized Keywords: For Immunoglobulins (IG), T Cell Receptors (TR), and Conventional Genes** (Lefranc 2011b), **From IMGT-ONTOLOGY CLASSIFICATION Axiom to IMGT Standardized Gene and Allele Nomenclature: For Immunoglobulins (IG) and T Cell Receptors (TR)** (Lefranc 2011c), **IMGT Unique Numbering for the Variable (V), Constant (C), and Groove**

[1]Corresponding author (Marie-Paule.Lefranc@igh.cnrs.fr).
Cite as: Cold Spring Harb Protoc; 2011; doi:10.1101/pdb.ip83

(G) Domains of IG, TR, MH, IgSF, and MhSF (Lefranc 2011d), and **IMGT Collier de Perles for the Variable (V), Constant (C), and Groove (G) Domains of IG, TR, MH, IgSF, and MhSF** (Lefranc 2011e).

In addition, protocols are available for **IMGT/V-QUEST: IMGT Standardized Analysis of the Immunoglobulin (IG) and T Cell Receptor (TR) Nucleotide Sequences** (Giudicelli et al. 2011), **IMGT/JunctionAnalysis: IMGT Standardized Analysis of the V-J and V-D-J Junctions of the Rearranged Immunoglobulins (IG) and T Cell Receptors (TR)** (Giudicelli and Lefranc 2011), **IMGT/Collier de Perles: IMGT Standardized Representation of Domains (IG, TR, and IgSF Variable and Constant Domains, MH and MhSF Groove Domains)** (Ehrenmann et al. 2011), **IMGT/DomainGapAlign: IMGT Standardized Analysis of Amino Acid Sequences of Variable, Constant, and Groove Domains (IG, TR, MH, IgSF, MhSF)** (Ehrenmann and Lefranc 2011a), and **IMGT/3Dstructure-DB: Querying the IMGT Database for 3D Structures in Immunology and Immunoinformatics (IG or Antibodies, TR, MH, RPI, and FPIA)** (Ehrenmann and Lefranc 2011b).

BACKGROUND INFORMATION

The DESCRIPTION axiom is an axiom of the Formal IMGT-ONTOLOGY or IMGT-Kaleidoscope (Giudicelli and Lefranc 1999; Lefranc et al. 2004, 2005a, 2008; Duroux et al. 2008). The DESCRIPTION axiom postulates that any molecule, cell, tissue, organ, organism, or population, any process and any relation, has to be described. The DESCRIPTION axiom has generated the concepts of description of IMGT-ONTOLOGY, the global reference in immunogenetics and immunoinformatics, built by IMGT (http://www.imgt.org; Lefranc et al. 2009). The DESCRIPTION axiom is one of seven axioms of the Formal IMGT-ONTOLOGY, the others being "IDENTIFICATION axiom," "CLASSIFICATION axiom," "NUMEROTATION axiom," "LOCALIZATION axiom," "ORIENTATION axiom," and "OBTENTION axiom." The leafconcepts (finest level of granularity of the concepts) of description are directly translated into IMGT standardized labels and used in the IMGT databases and tools. IMGT standardized labels have been essential for the sequence description in IMGT/LIGM-DB (Giudicelli et al. 2006), the first and the largest IMGT database that contains IG and TR nucleotide sequences and the translation of annotated sequences, IMGT/GENE-DB (Giudicelli et al. 2005), the IMGT gene database, IMGT/2Dstructure-DB, the IMGT amino acid sequence and 2D structure database (Ehrenmann et al. 2010b), and IMGT/3Dstructure-DB, the IMGT 3D structure database (Kaas et al. 2004; Ehrenmann et al. 2010b). IMGT standardized labels have been crucial in the IMGT tools for the analysis of nucleotide sequences (IMGT/V-QUEST [Giudicelli et al. 2004; Brochet et al. 2008; Giudicelli and Lefranc 2009], IMGT/JunctionAnalysis [Yousfi Monod et al. 2004; Bleakley et al. 2006], IMGT/HighV-QUEST [Alamyar et al. 2010]), of amino acid sequences (IMGT/DomainGapAlign [Ehrenmann et al. 2010b]), or of large genomic sequences (IMGT/LIGMotif [Lane et al. 2010]).

MOLECULE ENTITY PROTOTYPE

"Molecule_EntityPrototype" Concept

The "Molecule_EntityPrototype" is a concept, generated from the DESCRIPTION axiom, that provides the description for the "Molecule_EntityType" concept (IDENTIFICATION axiom). There are as many leafconcepts in the "Molecule_EntityPrototype" as there are leafconcepts in the "Molecule_EntityType" (**From IMGT-ONTOLOGY IDENTIFICATION Axiom to IMGT Standardized Keywords: For Immunoglobulins (IG), T Cell Receptors (TR), and Conventional Genes** [Lefranc 2011b]). Thus the "Molecule_EntityPrototype" comprises 38 leafconcepts that describe the organization of each entity with its constitutive motifs and relations. Each "Molecule_EntityPrototype" leafconcept is linked to a "Molecule_EntityType" leafconcept by the reciprocal relations "describes" and "is_described_by." For example, a "V-gene" is described by "V-GENE," a "V-D-J-gene" by "V-D-J-GENE," and a "L-V-D-J-sequence" by "L-V-D-J-SEQUENCE." Leafconcepts of description (labels in the IMGT databases and tools) are written in capital letters.

The 38 "Molecule_EntityPrototype" leafconcepts, as indicated by the suffix, describe the four major "MoleculeUnit" leafconcepts that are genes (10), transcripts (11), cDNA sequences (11), and chains (6). The 10 leafconcepts that are the most classical representatives of IG and TR data description in the IMGT information system are shown as examples (Table 1).

Table 1. IMGT-ONTOLOGY "Molecule_EntityPrototype" concept

	Molecule_EntityProtoType leafconcepts (labels in the IMGT databases and tools)									
Labels	V-gene	D-gene	J-gene	C-gene	V-J-gene	V-D-J-gene	L-V-J-C-sequence	L-V-D-J-C-sequence	V-J-C-chain	V-D-J-C-chain
1st-CYS	x			x	x	x	x	x	x	x
2nd-CYS	x			x	x	x	x	x	x	x
3'D-HEPTAMER		x								
3'D-NONAMER		x								
3'D-RS		x								
3'D-SPACER		x								
3'UTR	x	x	x	x	x	x	x	x		
5'D-HEPTAMER		x								
5'D-NONAMER		x								
5'D-RS		x								
5'D-SPACER		x								
5'UTR	x	x	x	x	x	x	x	x		
ACCEPTOR-SPLICE	x			x	x	x				
C-GENE-UNIT				x						
C-REGION				x			x	x	x	x
CDR1-IMGT	x				x	x	x	x	x	x
CDR2-IMGT	x				x	x	x	x	x	x
CDR3-IMGT	x				x	x	x	x	x	x
CONSERVED-TRP	x			x	x	x	x	x	x	x
D-GENE-UNIT		x								
D-REGION		x				x		x		
DONOR-SPLICE	x		x	x	x	x				
FR1-IMGT	x				x	x	x	x	x	x
FR2-IMGT	x				x	x	x	x	x	x
FR3-IMGT	x				x	x	x	x	x	x
FR4-IMGT					x	x	x	x	x	x
INIT-CODON	x				x	x	x	x		
J-GENE-UNIT			x							
J-HEPTAMER			x							
J-NONAMER			x							
J-PHE			x		x	x	x	x	x	x
J-REGION			x		x	x	x	x	x	x
J-RS			x							
J-SPACER			x							
J-TRP			x		x	x	x	x	x	x
JUNCTION					x	x	x	x	x	x
L-INTRON-L	x				x	x				

(*continued*)

Table 1. *Continued*

	Molecule_EntityProtoType leafconcepts (labels in the IMGT databases and tools)									
Labels	V-gene	D-gene	J-gene	C-gene	V-J-gene	V-D-J-gene	L-V-J-C-sequence	L-V-D-J-C-sequence	V-J-C-chain	V-D-J-C-chain
L-PART1	x				x	x				
L-PART2	x				x	x				
L-REGION							x	x		
L-V-D-J-C-REGION								x		
L-V-GENE-UNIT	x									
L-V-J-C-REGION							x			
(N-D)-REGION						x		x		x
N-REGION					x	x	x	x	x	
V-D-J-C-REGION								x		x
V-D-J-EXON						x				
V-D-J-REGION						x		x		x
V-EXON	x									
V-HEPTAMER	x									
V-INTRON	x				x	x				
V-J-C-REGION							x		x	
V-J-EXON					x					
V-J-REGION					x		x		x	
V-NONAMER	x									
V-REGION	x				x	x	x	x	x	x
V-RS	x									
V-SPACER	x									

The 10 most representative leafconcepts for IG and TR are shown with the most important labels (indicated by an "x") that describe them in the IMGT databases and tools. The labels for a detailed description of the C-GENE exons or of the C-REGION in these leafconcepts are not shown.

Associated Concepts of Description

The labels used for the description of the "Molecule_EntityPrototype" (Table 1) correspond to the leafconcepts of the following concepts:

"GeneUnit"

The "GeneUnit" concept has four leafconcepts (labels in the IMGT databases and tools): "L-V-GENE-UNIT," "D-GENE-UNIT," "J-GENE-UNIT," and "C-GENE-UNIT" (Lane et al. 2010). In contrast to "V-GENE," "D-GENE," "J-GENE," and "C-GENE," the "GeneUnit" labels are precisely delimited in 5′ and 3′, respectively, by the 5′ end and 3′ end of their constitutive labels: "L-PART1" and "V-RS" for V, "5′D-RS" and "3′D-RS" for D, "J-RS" and "J-REGION" for J, "ACCEPTOR-SPLICE" and "STOP-CODON" for C.

"RecombinationSignal"

The "RecombinationSignal" (RS) concept is specific to the IG and TR V, D, and J genes which rearrange at the DNA level in B cells (for the IG) and T cells (for the TR) (Lefranc and Lefranc 2001a,b) and represents the major difference from "conventional" genes. The "RecombinationSignal" labels describe the RS that are localized in 3′ of the V genes ("V-RS"), in 5′ of the J genes ("J-RS"), and on both sides of the D genes ("5′D-RS" and "3′D-RS") (Lefranc and Lefranc 2001a,b). They consist of conserved heptamers (-HEPTAMER) and nonamers (-NONAMER) separated by less conserved spacers (-SPACER) of 12 ± 1 or 23 ± 1 base pairs (bp) which vary between loci and species (IMGT Repertoire, http://www.imgt.org/).

"Core"

The "Core" concept allows one to describe the coding region of genes and contains five leafconcepts which are "REGION" (for "conventional" gene), "V-REGION," "D-REGION," "J-REGION," and "C-REGION" (for "variable" [V], "diversity" [D], "joining" [J], "constant" [C] genes). These important labels have permitted the definition and standardized description of the IG and TR alleles (concepts of classification), now approved at the international level (Lefranc and Lefranc 2001a,b) (**From IMGT-ONTOLOGY CLASSIFICATION Axiom to IMGT Standardized Gene and Allele Nomenclature: For Immunoglobulins (IG) and T Cell Receptors (TR)** [Lefranc 2011c]). Moreover, they have allowed one to link sequences and structures of the IG and TR by their composite labels that describe rearranged sequences. Thus, a V domain is encoded by a "V-J-REGION" or a "V-D-J-REGION," and IG and TR chains are encoded by a "V-J-C-REGION" or a "V-D-J-C-REGION."

"FR-IMGT and CDR-IMGT"

The "FR-IMGT" and "CDR-IMGT" concepts allow one to describe the framework regions and complementarity determining regions, respectively, based on the IMGT unique numbering (Lefranc 1997, 1999; Lefranc et al. 2003) (**IMGT Unique Numbering for the Variable (V), Constant (C), and Groove (G) Domains of IG, TR, MH, IgSF, and MhSF** [Lefranc 2011d]). "FR-IMGT" has four leafconcepts (labels in the IMGT databases and tools): "FR1-IMGT," "FR2-IMGT," "FR3-IMGT," and, for rearranged sequences, "FR4-IMGT." "CDR-IMGT" has three leafconcepts (labels in the IMGT databases and tools): "CDR1-IMGT," "CDR2-IMGT," and "CDR3-IMGT" (the composition of "CDR3-IMGT" is different depending on whether the sequence is germline or rearranged) (Lefranc and Lefranc 2001a,b).

"Junction"

The "JUNCTION" concept allows one to describe the junction resulting from the V-J or V-D-J rearrangement. "JUNCTION" leafconcepts (labels in the IMGT databases and tools) comprise: "3′V-REGION," "5′J-REGION," "N-REGION," "(N-D)-REGION," and "P-REGION" (Yousfi Monod et al. 2004; Bleakley et al. 2006).

"Leader"

The "leader" concept allows one to describe the leader (L) or signal peptide. "Leader" has three leafconcepts (labels in the IMGT databases and tools): "L-PART1," "L-PART2," and "L-REGION."

In the IG and TR V genes, "L-PART1" (first exon of "L-V-GENE-UNIT") and L-PART2 (first codons of V-EXON, the second exon of "L-V-GENE-UNIT") are separated by the "V-INTRON" and their splicing frame is of type 1 (sf1) (IMGT Aide-mémoire, http://www.imgt.org). In cDNA and chains (spliced L-PART1 and L-PART2), the leader is described as "L-REGION." The L-REGION is absent from mature chains as a result of proteolytic cleavage in the endoplamic reticulum (ER). The "L-PART1," "L-PART2," and "L-REGION" labels are also used for the description of "conventional-with-leader" genes (**From IMGT-ONTOLOGY IDENTIFICATION Axiom to IMGT Standardized Keywords: For Immunoglobulins (IG), T Cell Receptors (TR), and Conventional Genes** [Lefranc 2011b]).

"Conserved_AA_codons"

The "Conserved_AA_codon" concept allows one to describe conserved amino acids (AA) and/or codons. It includes for example, for the V-DOMAIN, five leafconcepts (labels in the IMGT databases and tools): "1st-CYS" (position 23), "CONSERVED-TRP" (position 41), "2nd-CYS" (position 104), "J-TRP," and "J-PHE" (position 118) (positions in parentheses are according to the IMGT unique numbering [Lefranc 1997, 1999; Lefranc et al. 2003]) (**IMGT Unique Numbering for the Variable (V), Constant (C), and Groove (G) Domains of IG, TR, MH, IgSF, and MhSF** [Lefranc 2011d]). The "2nd-CYS" 104 and "J-TRP" or "J-PHE" 118 (that belong to the FR3-IMGT and FR4-IMGT, respectively) are the anchors of the rearranged CDR3-IMGT. They delimit the "JUNCTION"that spans 104 to 118 and includes them at its ends (Yousfi Monod et al. 2004; Bleakley et al. 2006).

Conserved amino acids "J-TRP" and "J-PHE" are part of a conserved motif "[W,F]-G-X-G" in the "FR4-IMGT" of the "V-DOMAIN" of IG and TR (where W, tryptophan "J-TRP"; F, phenylalanine "J-PHE"; G, glycine; and X, any amino acid). Codons of conserved amino acids are "tgg" for "CONSERVED-TRP" and "J-TRP, "tgc," and "tgt" for "1st-CYS" and "2nd-CYS," and "ttt" and "ttc" for "J-PHE." These conserved codons and amino acids and their mutations are identified and described by IMGT/V-QUEST (Giudicelli et al. 2004; Brochet et al. 2008; Giudicelli and Lefranc 2009) and IMGT/HighV-QUEST (Alamyar et al. 2010) for nucleotide sequences.

The "1st-CYS" (position 23), "CONSERVED-TRP" (position 41), and "2nd-CYS" (position 104) are also conserved in the "C-DOMAIN" of IG and TR (Lefranc et al. 2005c). Conserved amino acids for V, C, or G domains are highlighted in IMGT/DomainDisplay (Ehrenmann et al. 2010b; http://www.imgt.org).

Prototype or Graphical Representation

A prototype is a graphical representation of a "Molecule_EntityPrototype" leafconcept. Two prototypes, those of "V-GENE" and "V-D-J-GENE," are shown in Figure 1 as examples of a germline entity and of a rearranged entity, respectively. Twenty-six labels for "V-GENE" and 29 labels for "V-D-J-GENE" (19 of them being shared by the two prototypes), on a total of 277 different labels for sequences in IMGT/LIGM-DB, are necessary and sufficient for a complete description of these prototypes. Definitions of the IMGT/LIGM-DB labels used for the two "V-GENE" and "V-D-J-GENE" prototypes are provided, as examples, in Table 2.

The organization of a prototype is based on the relations that order two labels. A set of 12 relations is necessary and sufficient to describe the relations between labels in a prototype (Table 3; Giudicelli and Lefranc 1999; Lefranc et al; 2004; Duroux et al. 2008; Lane et al. 2010). The reciprocal relations "is_in_5_prime_of" and "is_in_3_prime_of" describe the relative position of labels on a 5′–3′ DNA strand when there is no intersection between labels.

GENE CLUSTER

"GeneCluster" Concept

The "GeneCluster" concept allows one to describe gene clusters, which are genomic sequences containing several genes. The genes in a cluster can be described either by the same prototype (e.g., a "V-CLUSTER" only contains genes described by the "V-GENE" prototype) or by different prototypes (e.g., a "V-D-J-CLUSTER" contains at least one "V-GENE," one "D-GENE," and one "J-GENE"). Examples of IMGT-ONTOLOGY "GeneCluster" leafconcepts (labels in the IMGT databases and tools) are shown in Table 4. They are particularly useful for the annotation of large scale genomic IG and TR loci (Lane et al. 2010) and several of them are also used by Sequence Ontology (SO) (Table 4; Eilbeck et al. 2005).

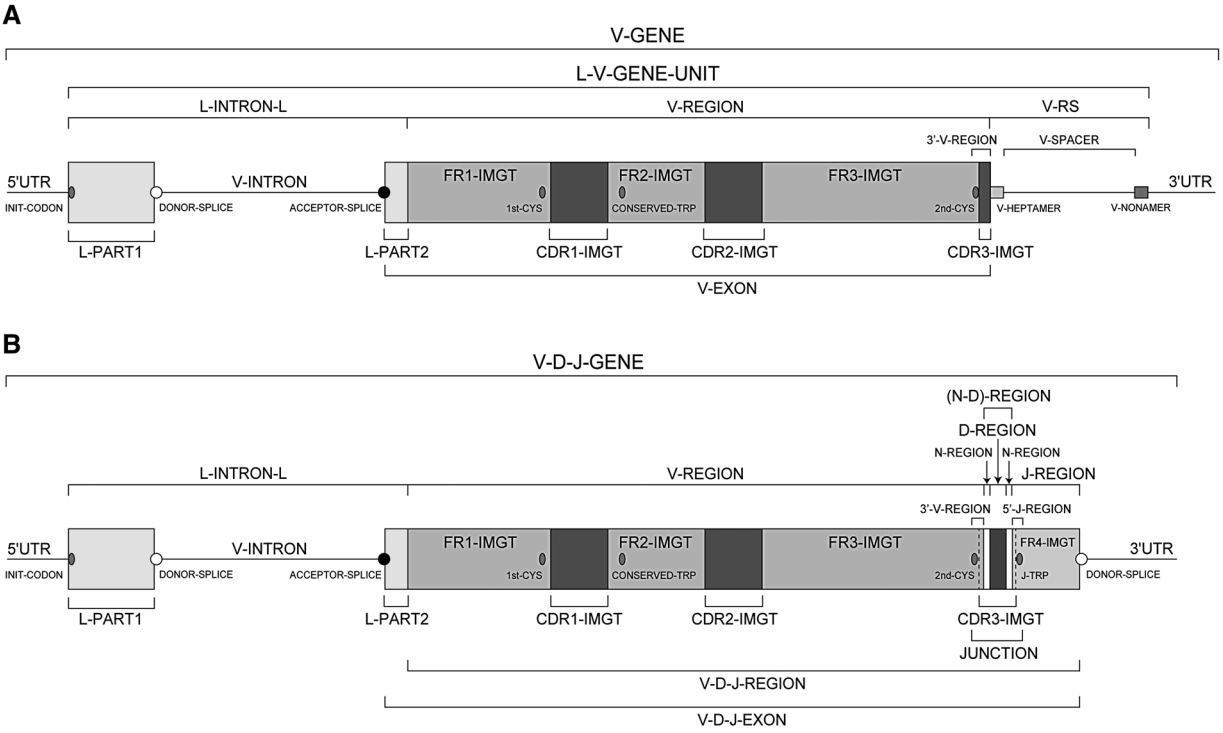


FIGURE 1. Prototype or graphical representation of two "Molecule_EntityPrototype" leafconcepts. (*A*) "V-GENE." (*B*) "V-D-J-GENE." Thirty-nine labels (27 for "V-GENE" and 32 for "V-D-J-GENE" of which 20 are shared) (Table 2) and 12 relations (Table 3) are necessary and sufficient for a complete description of these prototypes.

"RECEPTOR," "CHAIN," and "DOMAIN"

The "RECEPTOR," "CHAIN," and "DOMAIN" concepts allow one to describe amino acid sequences and structures. Two-hundred eighty-five leafconcepts (labels in IMGT/3Dstructure-DB) have been defined for a standardized description of the amino acid sequences and structures.

"RECEPTOR" Concept

The "RECEPTOR" concept allows one to describe a receptor made of one or several amino acid chains. In IMGT (IMGT/3Dstructure-DB, IMGT/DomainGapAlign), the receptor is described, whenever possible, at the finest level of granularity with each chain described at the gene level for a given species. For example, an IG (Fig. 2) is described by the label "*Homo sapiens* IG-GAMMA-1_KAPPA," and a TR (Fig. 3) by "*Homo sapiens* TR-ALPHA_BETA-1" or TR-GAMMA-1_DELTA. The same rules are used for the description of monoclonal antibody fragments whatever their format (e.g., *Mus musculus* FAB-GAMMA-1_KAPPA, *Homo sapiens* SCFV-HEAVY-KAPPA) and similar standards have been developed for the description of FPIA (Ehrenmann et al. 2010b).

"CHAIN" Concept

The "CHAIN" concept allows one to describe an amino acid chain. In IMGT (IMGT/3Dstructure-DB, IMGT/DomainGapAlign), the chain is described at the finest level of granularity, that is at the gene level for a given species: The number of functional IG and TR "constant" C genes defines the number of "CHAIN" leafconcepts (labels in the IMGT databases and tools). For *Homo sapiens*, the number of "CHAIN" labels (IMGT/3Dstructure-DB) is 6 for L-CHAIN (as there are six functional C genes: one IGKC, five IGLC), 9 for H-CHAIN (nine functional IGHC), 6 for TR (one TRAC, two TRBC, two TRGC, one TRDC) (Table 5). Relations with the sequence labels "V-J-C-REGION" and "V-D-J-C-REGION" (IMGT/LIGM-DB) and with the Molecule_EntityPrototype "V-J-C-CHAIN" and "V-D-J-C-CHAIN"

Table 2. List of the 39 labels used for the description of the "Molecule_EntityPrototype" "V-GENE" and "V-D-J-GENE" and their definitions (from IMGT/LIGM-DB)

Label name	Definition
1st-CYS	codon (3 nt) for conserved cysteine at position 23 in B-STRAND (FR1-IMGT)
2nd-CYS	codon (3 nt) for conserved cysteine at position 104 in F-STRAND (FR3-IMGT)
3'UTR	3' untranslated sequence, EMBL feature Key signification
3'V-REGION	region from 2nd-CYS to the 3' end of the V-REGION (for germline and rearranged)
5'UTR	5' untranslated sequence, EMBL feature Key signification
5'J-REGION	region from the 5' end of the J-REGION to the J-PHE or J-TRP (for germline and rearranged)
ACCEPTOR-SPLICE	splicing site in 5' of coding region (nagnn), with splicing occurring after g
CDR1-IMGT	first complementarity determining region according to the IMGT unique numbering
CDR2-IMGT	second complementarity determining region according to the IMGT unique numbering
CDR3-IMGT	third complementarity determining region according to the IMGT unique numbering
CONSERVED-TRP	codon (3 nt) for conserved tryptophan at position 41 in C-STRAND (FR2-IMGT)
D-REGION	coding region of D-GENE (plus 1 or 2 nt after the 5'D-HEPTAMER and/or before the 3'D-HEPTAMER, if present), or corresponding region in cDNA
DONOR-SPLICE	splicing site in 3' of coding region (ngt), with splicing occurring before g
FR1-IMGT	first framework according to the IMGT unique numbering
FR2-IMGT	second framework according to the IMGT unique numbering
FR3-IMGT	third framework according to the IMGT unique numbering
FR4-IMGT	fourth framework according to the IMGT unique numbering
INIT-CODON	initiation codon ATG
J-REGION	coding region of J-GENE (plus 1 or 2 nt after J-HEPTAMER, if present) or corresponding region in cDNA
J-TRP	codon (3 nt) for conserved tryptophan at position 118 in G-STRAND (FR4-IMGT) (in J-REGION of IG heavy chain)
JUNCTION	coding region encompassing the V-J or V-D-J junction from 2nd CYS to the J-PHE or J-TRP of the J-REGION
L-INTRON-L	sequence including L-PART1, V-INTRON, and L-PART2, in genomic DNA, or corresponding sequence in unspliced cDNA
L-PART1	exon encoding the first part of the leader peptide of a V-, V-D-, V-D-J-, or V-J-GENE or corresponding region in unspliced cDNA
L-PART2	5' region of V-EXON encoding the second part of leader peptide of a V-, V-D-, V-D-J-, or V-J-GENE or corresponding region in unspliced cDNA
L-V-GENE-UNIT	germline genomic DNA including L-PART1, V-INTRON, V-EXON, and V-RS
N-REGION	coding region encompassing an N diversity sequence, in a JUNCTION
(N-D)-REGION	coding region encompassing N-REGION and D-REGION, and if present P-REGION, in a JUNCTION
P-REGION	region encompassing a palindromic (P) sequence at the untrimmed end of a V-REGION, D-REGION, or J-REGION, in a JUNCTION
V-D-J-EXON	rearranged genomic DNA including L-PART2, V-, any D- and N-REGION, and J-REGION
V-D-J-GENE	rearranged genomic DNA including L-PART1, V-INTRON, and V-D-J-EXON, with the 5'UTR and 3'UTR
V-D-J-REGION	coding region including V-, any D- and N-REGION, and J-REGION, in rearranged genomic DNA, or corresponding region in cDNA
V-EXON	germline genomic DNA including L-PART2 and V-REGION
V-GENE	germline genomic DNA including L-PART1, V-INTRON, and V-EXON, with the 5'UTR and 3'UTR
V-HEPTAMER	7 nt recombination site, like CACAGTG, part of V-RS
V-INTRON	noncoding sequence between L-PART1 and V-EXON, in genomic DNA, or corresponding sequence in unspliced cDNA
V-NONAMER	9 nt recombination site, like ACAAAAACC, part of V-RS
V-REGION	coding region of V-GENE without the leader peptide (plus 1 or 2 nt before the V-HEPTAMER, if present), or corresponding region in cDNA
V-RS	recombination signal including V-HEPTAMER, V-SPACER, and V-NONAMER in 3' of V-REGION of a V-GENE or V-SEQUENCE
V-SPACER	12 or 23 nt spacer between the V-HEPTAMER and the V-NONAMER of a V-RS

A description of IMGT/LIGM-DB can be found in Giudicelli et al. (2006). In the list, 20 labels are shared by the two prototypes, seven are specific for "V-J-GENE," and 12 are specific for "V-D-J-GENE."

Table 3. IMGT-ONTOLOGY relations between labels used for the description of prototypes

Relation	Reciprocal relation
"adjacent_at_its_5_prime_to"	"adjacent_at_its_3_prime_to"
"included_with_same_5_prime_in"	"includes_with_same_5_prime"
"included_with_same_3_prime_in"	"includes_with_same_3_prime"
"overlaps_at_its_5_prime_with"	"overlaps_at_its_3_prime_with"
"included_in"	"includes"
"is_in_5_prime_of"	"is_in_3_prime_of"

Twelve relations are necessary and sufficient for the description of prototypes (Giudicelli and Lefranc 1999; Lefranc et al. 2004; Duroux et al. 2008; Lane et al. 2010).

leafconcepts (labels in the IMGT databases and tools) have allowed us to bridge the gap between sequence and structure description.

"DOMAIN" Concept

The "DOMAIN" concept allows one to describe a subunit of a chain characterized by its 3D structure (Kaas et al. 2007). The IMGT-ONTOLOGY "DOMAIN" concept includes six main leafconcepts (labels in IMGT/3Dstructure-DB and in IMGT/DomainGapAlign [Ehrenmann et al. 2010b]):

- the "V-DOMAIN" (V for variable)
- the "C-DOMAIN" (C for constant)
- the "G-DOMAIN" (G for groove)
- the "V-LIKE-DOMAIN"
- the "C-LIKE-DOMAIN"
- the "G-LIKE-DOMAIN"

"V-DOMAIN" and "C-DOMAIN" are found in IG and TR chains and "V-LIKE-DOMAIN" and "C-LIKE-DOMAIN" in chains other than IG or TR (Lefranc 1997, 1999; Lefranc et al. 2003, 2005c). "G-DOMAIN" is found in MH chains and "G-LIKE-DOMAIN" in chains other than MH (Lefranc et al.

Table 4. Examples of IMGT-ONTOLOGY "GeneCluster" leafconcepts

IMGT-ONTOLOGY "GeneCluster" leafconcepts (labels in the IMGT databases and tools)	Sequence ontology	"Molecule_EntityPrototype" leafconcepts (labels and prototypes in the IMGT databases and tools)	
		Minimum number of different prototypes	Name of the different prototypes
V-CLUSTER	SO:0000526	1	V-GENE
J-CLUSTER	SO:0000513	1	J-GENE
D-CLUSTER	SO:0000559	1	D-GENE
D-J-CLUSTER	SO:0000560	2	D-GENE J-GENE
V-D-CLUSTER		2	V-GENE D-GENE
V-J-CLUSTER	SO:0000534	2	V-GENE J-GENE
V-D-J-CLUSTER	SO:0000532	3	V-GENE D-GENE J-GENE

These examples are those used in IMGT/LIGMotif (Lane et al. 2010). Six of them are also used by SO (Eilbeck et al. 2005). Relations with the "Molecule_EntityPrototype" leafconcepts comprise the minimum number of different prototypes and the name of these different prototypes.

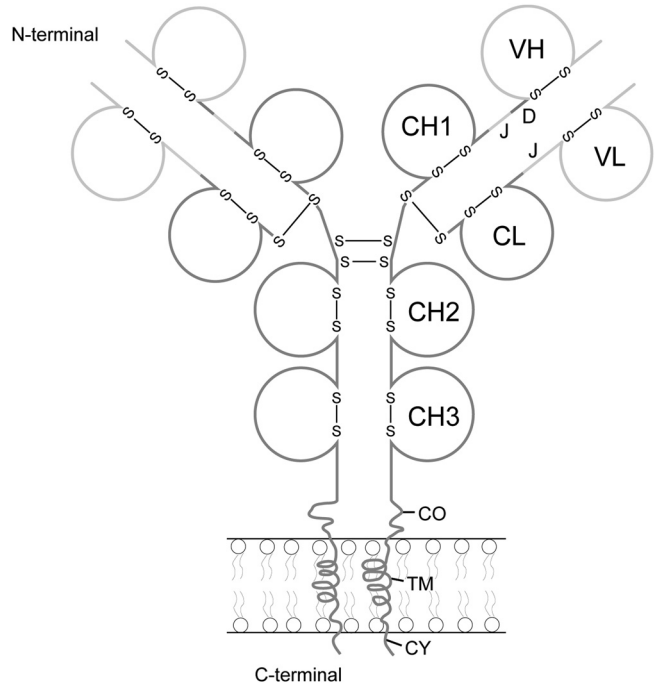


FIGURE 2. Schematic representation of an IG. The IG chain and domain labels (IMGT/3Dstructure-DB) and the corresponding sequence labels (IMGT/LIGM-DB) are shown in Tables 5 and 6. A membrane IG (mIG) is shown. CO, CONNECTING-REGION; TM, TRANSMEMBRANE-REGION; CY, CYTOPLASMIC-REGION. The example is *Homo sapiens* IG-GAMMA1_KAPPA (IgG1κ) (Lefranc and Lefranc 2001a).

2005b) (**IMGT Collier de Perles for the Variable (V), Constant (C), and Groove (G) Domains of IG, TR, MH, IgSF, and MhSF** [Lefranc 2011e]).

Correspondence between domain labels used for the description of amino acid sequences and structures (e.g., in IMGT/3Dstructure-DB) and sequence labels used for the description of nucleotide

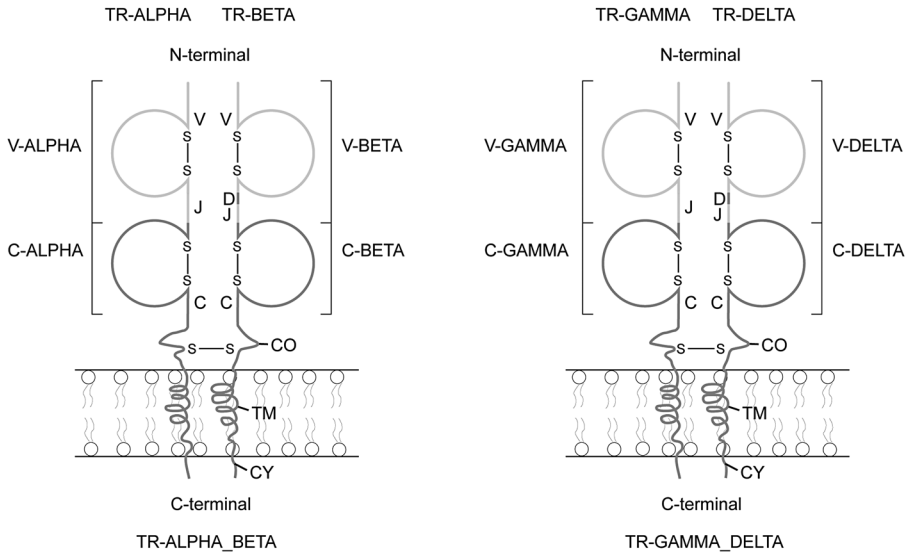


FIGURE 3. Schematic representation of TR. The TR chain and domain labels (IMGT/3Dstructure-DB) and the corresponding sequence labels (IMGT/LIGM-DB) are shown in Tables 5 and 7. CO, CONNECTING-REGION; TM, TRANSMEMBRANE-REGION; CY, CYTOPLASMIC-REGION. The examples are *Homo sapiens* TR-ALPHA_BETA and TR-GAMMA1_DELTA (TR-αβ and TR-γ1δ) (Lefranc and Lefranc 2001b).

Table 5. *Homo sapiens* "CHAIN" labels and relations with the sequence labels and the Molecule_EntityPrototype leafconcepts

CHAIN labels (IMGT/3Dstructure-DB)			Sequence labels (IMGT/LIGM-DB)	Molecule_EntityPrototype leafconcepts (labels in the IMGT databases and tools)
IG-CHAIN	L-CHAIN	L-KAPPA	V-J-C-REGION	V-J-C-CHAIN
		L-LAMBDA-1		
		L-LAMBDA-2		
		L-LAMBDA-3		
		L-LAMBDA-6		
		L-LAMBDA-7		
	H-CHAIN	H-MU	V-D-J-C-REGION	V-D-J-C-CHAIN
		H-DELTA		
		H-GAMMA-1		
		H-GAMMA-2		
		H-GAMMA-3		
		H-GAMMA-4		
		H-ALPHA-1		
		H-ALPHA-2		
		H-EPSILON		
TR-CHAIN		TR-ALPHA	V-J-C-REGION	V-J-C-CHAIN
		TR-BETA-1	V-D-J-C-REGION	V-D-J-C-CHAIN
		TR-BETA-2		
		TR-GAMMA-1	V-J-C-REGION	V-J-C-CHAIN
		TR-GAMMA-2		
		TR-DELTA	V-D-J-C-REGION	V-D-J-C-CHAIN

sequences (e.g., in IMGT/LIGM-DB) are shown for the IG domains (Table 6; Fig. 2) and the TR domains (Table 7; Fig. 3) from *Homo sapiens*.

IMGT LABELS: A PARADIGM FOR IMMUNOGENETICS AND ANTIBODY ENGINEERING

With >560 IMGT standardized labels for the nucleotide and amino acid sequences, 2D and 3D structures, IMGT databases and tools provide a detailed and accurate description of IG and TR sequences and structures that is crucial in basic and clinical research (Lefranc 2005, 2008a,b; Ghia et al. 2007; Giudicelli and Lefranc 2009) and in biotechnology related to antibody engineering and humanization (Lefranc 2009; Ehrenmann et al. 2010a). IMGT/mAb-DB (Poiron et al. 2010) allows one to query for therapeutic antibodies for which sequences and 3D structures are available. Novel IMGT chain labels have been defined for the description of new antibody formats and FPIA (Ehrenmann et al. 2010b). Antibodies

Table 6. Correspondence between domain and sequence labels for *Homo sampiens* IG domains

Domain labels (IMGT/3Dstructure-DB)			Sequence labels (IMGT/LIGM-DB)	
V-DOMAIN	VL	V-KAPPA	V-J-REGION	
		V-LAMBDA	V-J-REGION	
	VH		V-D-J-REGION	
C-DOMAIN	CL	C-KAPPA	C-REGION	
		C-LAMBDA		
	CH	CH1	CH1	C-REGION[a]
		CH2	CH2	
		CH3	CH3	
		CH4[b]	CH4[b]	

[a]The H-CHAIN "C-REGION" also includes the "HINGE-REGION" for the "H-ALPHA," "H-DELTA," and "H-GAMMA" chains and, for membrane IG (mIG), the "CONNECTING-REGION" (CO), "TRANSMEMBRANE-REGION" (TM), and "CYTOPLASMIC-REGION" (CY) (Fig. 2); for secreted IG (sIG), the "C-REGION" includes "CH-S" instead of CO, TM, and CY.
[b]For "H-MU" and "H-EPSILON."

Table 7. Correspondence between domain and sequence labels for *Homo sampiens* TR domains

Domain labels (IMGT/3Dstructure-DB)		Sequence labels (IMGT/LIGM-DB)
V-DOMAIN	V-ALPHA	V-J-REGION
	V-BETA	V-D-J-REGION
	V-GAMMA	V-J-REGION
	V-DELTA	V-D-J-REGION
C-DOMAIN	C-ALPHA	C-REGION[a]
	C-BETA	C-REGION[a]
	C-GAMMA	C-REGION[a]
	C-DELTA	C-REGION[a]

[a]The TR-CHAIN "C-REGION" also includes the CONNECTING-REGION (CO), TRANSMEMBRANE-REGION (TM), and CYTOPLASMIC-REGION (CY) (Fig. 3).

and FPIA represent a large number of the pharmaceutical substances submitted to the World Health Organization/International Nonproprietary Names (WHO/INN) Program (General policies for monoclonal antibodies World Health Organization. INN Working Document 09.251. Update 18/12/2009. http://www.who.int/medicines/services/inn/en/). The INN definition of antibodies is based on the IMGT-ONTOLOGY concepts of classification (nomenclature), description (labels), and numerotation (IMGT unique numbering) and this information is obtained with the IMGT/DomainGapAlign tool (Lefranc 2011f). Since 2008, amino acid sequences of monoclonal antibodies (mAb, suffix -mab) and of fusion proteins for immune applications (FPIA, suffix -cept) from WHO/INN have been entered in IMGT. These therapeutic applications emphasize the importance of the IMGT-ONTOLOGY concepts in bridging the gap between antibody sequences, 2D and 3D structures.

ACKNOWLEDGMENTS

We thank Véronique Giudicelli, Chantal Ginestoux, Amandine Lacan, Eltaf Alamyar, François Ehrenmann, Jérôme Lane, Mansour Saljoqi, Patrice Duroux, and Gérard Lefranc and the IMGT team for their motivation and expertise.

REFERENCES

Alamyar E, Giudicelli V, Duroux P, Lefranc M-P. 2010. IMGT/HighV-QUEST: A high-throughput system and Web portal for the analysis of rearranged nucleotide sequences of antigen receptors—High-throughput version of IMGT/V-QUEST. JOBIM 2010 Poster 60. http://www.jobim2010.fr/?q=fr/node/55.

Bleakley K, Giudicelli V, Wu Y, Lefranc M-P, Biau G. 2006. IMGT standardization for statistical analyses of T cell receptor junctions: The TRAV-TRAJ example. *In Silico Biol* **6:** 573–588.

Brochet X, Lefranc M-P, Giudicelli V. 2008. IMGT/V-QUEST: The highly customized and integrated system for IG and TR standardized V-J and V-D-J sequence analysis. *Nucl Acids Res* **36:** W503–W508.

Duroux P, Kaas Q, Brochet X, Lane J, Ginestoux C, Lefranc M-P, Giudicelli V. 2008. IMGT-kaleidoscope, the formal IMGT-ONTOLOGY paradigm. *Biochimie* **90:** 570–583.

Ehrenmann F, Lefranc M-P. 2011a. IMGT/DomainGapAlign: IMGT standardized analysis of amino acid sequences of variable, constant, and groove domains (IG, TR, MH, IgSF, MhSF). *Cold Spring Harb Protoc* doi: 10.1101/pdb.prot5636.

Ehrenmann F, Lefranc M-P. 2011b. IMGT/3Dstructure-DB: Querying the IMGT database for 3D structures in immunology and immunoinformatics (IG or antibodies, TR, MH, RPI, and FPIA). *Cold Spring Harb Protoc* doi: 10.1101/pdb.prot5637.

Ehrenmann F, Duroux P, Giudicelli V, Lefranc M-P. 2010a. Standardized sequence and structure analysis of antibody using IMGT. In *Antibody engineering*, 2nd ed. (ed. R Kontermann, S Dübel), Vol 2, pp. 11–31. Springer-Verlag, Berlin/Heidelberg.

Ehrenmann F, Kaas Q, Lefranc M-P. 2010b. IMGT/3Dstructure-DB and IMGT/DomainGapAlign: A database and a tool for immunoglobulins or antibodies, T cell receptors, MHC, IgSF and MhcSF. *Nucl Acids Res* **38:** D301–D307.

Ehrenmann F, Giudicelli V, Brochet X, Lefranc M-P. 2011. IMGT/Collier de Perles: IMGT standardized representation of domains (IG, TR, and IgSF variable and constant domains, MH and MhSF groove domains). *Cold Spring Harb Protoc* doi: 10.1101/pdb.prot5635.

Eilbeck K, Lewis SE, Mungall CJ, Yandell M, Stein L, Durbin R, Ashburner M. 2005. The Sequence Ontology: A tool for the unification of genome annotations. *Genome Biol* **6:** pR44. doi: 10.1186/gb-2005-6-5-r44.

Ghia P, Stamatopoulos K, Belessi C, Moreno C, Stilgenbauer S, Stevenson F, Davi F, Rosenquist R, European research Initiative on CLL (ERIC). 2007. ERIC recommendations on IGHV gene mutational status analysis in chronic lymphocytic leukemia. *Leukemia* **21:** 1–3.

Giudicelli V, Lefranc M-P. 1999. Ontology for Immunogenetics: The IMGT-ONTOLOGY. *Bioinformatics* **12:** 1047–1054.

Giudicelli V, Lefranc M-P. 2009. IMGT standardized analysis of immunoglobulin rearranged sequences. In *Immunoglobulin gene analysis in chronic lymphocytic leukemia* (ed. P Ghia et al.), pp. 33–52. Wolters Kluwer Health Italy Ltd, Milan.

Giudicelli V, Lefranc M-P. 2011. IMGT/JunctionAnalysis: IMGT standardized analysis of the V-J and V-D-J junctions of the rearranged immunoglobulins (IG) and T cell receptors (TR). *Cold Spring Harb Protoc* doi: 10.1101/pdb.prot5634.

Giudicelli V, Chaume D, Lefranc M-P. 2004. IMGT/V-QUEST, an integrated software for immunoglobulin and T cell receptor V-J and V-D-J rearrangement analysis. *Nucl Acids Res* **32:** W435–W440.

Giudicelli V, Chaume D, Lefranc M-P. 2005. IMGT/GENE-DB: A comprehensive database for human and mouse immunoglobulin and T cell receptor genes. *Nucl Acids Res* **33:** D256–D261.

Giudicelli V, Duroux P, Ginestoux C, Folch G, Jabado-Michaloud J, Chaume D, Lefranc M-P. 2006. IMGT/LIGM-DB, the IMGT comprehensive database of immunoglobulin and T cell receptor nucleotide sequences. *Nucl Acids Res* **34:** D781–D784.

Giudicelli V, Brochet X, Lefranc M-P. 2011. IMGT/V-QUEST: IMGT standardized analysis of the immunoglobulin (IG) and T cell receptor (TR) nucleotide sequences. *Cold Spring Harb Protoc* doi: 10.1101/pdb.prot5633.

Kaas Q, Ruiz M, Lefranc M-P. 2004. IMGT/3Dstructure-DB and IMGT/StructuralQuery, a database and a tool for immunoglobulin, T cell receptor and MHC structural data. *Nucl Acids Res* **32:** D208–D210.

Kaas Q, Ehrenmann F, Lefranc M-P. 2007. IG, TR, MHC, IgSf and MhcSF: What do we learn from the IMGT Colliers de Perles? *Brief Funct Genomic Proteomic* **6:** 253–264.

Lane L, Duroux P, Lefranc M-P. 2010. From IMGT-ONTOLOGY to IMGT/LIGMotif: The IMGT standardized approach for immunoglobulin and T cell receptor gene identification and description in large genomic sequences. *BMC Bioinformatics* **11:** 223. doi: 10.1186/1471-2105-11-223.

Lefranc M-P. 1997. Unique database numbering system for immunogenetic analysis. *Immunol Today* **18:** 509.

Lefranc M-P. 1999. The IMGT unique numbering for Immunoglobulins, T cell receptors and Ig-like domains. *The Immunologist* **7:** 132–136.

Lefranc M-P. 2005. IMGT, the international ImMunoGeneTics information system: A standardized approach for immunogenetics and immunoinformatics. *Immunome Res* **1:** 3.

Lefranc M-P. 2008a. IMGT, the international ImMunoGeneTics information system for immunoinformatics. Methods for querying IMGT databases, tools and Web resources in the context of immunoinformatics. *Mol Biotechnol* **40:** 101–111.

Lefranc M-P. 2008b. IMGT-ONTOLOGY, IMGT databases, tools and Web resources for Immunoinformatics. In *Immunoinformatics* (ed. C Schoenbach et al.), pp. 1–18. Springer, New York.

Lefranc M-P. 2009. Antibody database and tools: The IMGT experience. In *Therapeutic monoclonal antibodies: From bench to clinic* (ed. A Zhiqiang), pp. 91–114. John Wiley and Sons, Inc, Hoboken, NJ.

Lefranc M-P. 2011a. IMGT, the international ImMunoGeneTics information system. *Cold Spring Harb Protoc* doi: 10.1101/pdb.top115.

Lefranc M-P. 2011b. From IMGT-ONTOLOGY IDENTIFICATION axiom to IMGT standardized keywords: For immunoglobulins (IG), T cell receptors (TR), and conventional genes. *Cold Spring Harb Protoc* doi: 10.1101/pdb.ip82.

Lefranc M-P. 2011c. From IMGT-ONTOLOGY CLASSIFICATION axiom to IMGT standardized gene and allele nomenclature: For immunoglobulins (IG) and T cell receptors (TR). *Cold Spring Harb Protoc* doi: 10.1101/pdb.ip84.

Lefranc M-P. 2011d. IMGT unique numbering for the variable (V), constant (C), and groove (G) domains of IG, TR, MH, IgSF, and MhSF. *Cold Spring Harb Protoc* doi: 10.1101/pdb.ip85.

Lefranc M-P. 2011e. IMGT Collier de Perles for the Variable (V), Constant (C), and Groove (G) Domains of IG, TR, MH, IgSF, and MhSF. *Cold Spring Harb Protoc* doi: 10.1101/pdb.ip86.

Lefranc M-P. 2011f. Antibody nomenclature: From IMGT-ONTOLOGY to INN definition. *MAbs* **3:** 1–2.

Lefranc M-P, Lefranc G. 2001a. *The immunoglobulin FactsBook*. Academic Press, London.

Lefranc M-P, Lefranc G. 2001b. *The T cell receptor FactsBook*. Academic Press, London.

Lefranc M-P, Pommié C, Ruiz M, Giudicelli V, Foulquier E, Truong L, Thouvenin-Contet V, Lefranc G. 2003. IMGT unique numbering for immunoglobulin and T cell receptor variable domains and Ig superfamily V-like domains. *Dev Comp Immunol* **27:** 55–77.

Lefranc M-P, Giudicelli V, Ginestoux C, Bosc N, Folch G, Guiraudou D, Jabado-Michaloud J, Magris S, Scaviner D, Thouvenin V, et al. 2004. IMGT-ONTOLOGY for immunogenetics and immunoinformatics. *In Silico Biol* **4:** 17–29.

Lefranc M-P, Clément O, Kaas Q, Duprat E, Chastellan P, Coelho I, Combres K, Ginestoux C, Giudicelli V, Chaume D, et al. 2005a. IMGT-Choreography for immunogenetics and immunoinformatics. *In Silico Biol* **5:** 45–60.

Lefranc M-P, Duprat E, Kaas Q, Tranne M, Thiriot A, Lefranc G. 2005b. IMGT unique numbering for MHC groove G-DOMAIN and MHC superfamily (MhcSF) G-LIKE-DOMAIN. *Dev Comp Immunol* **29:** 917–938.

Lefranc M-P, Pommié C, Kaas Q, Duprat E, Bosc N, Guiraudou D, Jean C, Ruiz M, Da Piedade I, Rouard M, et al. 2005c. IMGT unique numbering for immunoglobulin and T cell receptor constant domains and Ig superfamily C-like domains. *Dev Comp Immunol* **29:** 185–203.

Lefranc M-P, Giudicelli V, Regnier L, Duroux P. 2008. IMGT, a system and an ontology that bridge biological and computational spheres in bioinformatics. *Brief Bioinform* **9:** 263–275.

Lefranc M-P, Giudicelli V, Ginestoux C, Jabado-Michaloud J, Folch G, Bellahcene F, Wu Y, Gemrot E, Brochet X, Lane J, et al. 2009. IMGT, the international ImMunoGeneTics information system. *Nucl Acids Res* **37:** D1006–D1012.

Poiron C, Wu Y, Ginestoux C, Ehrenmann F, Duroux P, Lefranc M-P. 2010. IMGT/mAb-DB: The IMGT database for therapeutic monoclonal antibodies. JOBIM 2010 Poster 13. http://www.jobim2010.fr/?q=fr/node/55.

Yousfi Monod M, Giudicelli V, Chaume D, Lefranc M-P. 2004. IMGT/JunctionAnalysis: The first tool for the analysis of the immunoglobulin and T cell receptor complex V-J and V-D-J JUNCTIONs. *Bioinformatics* **20:** i379–i385.

Information Panel

From IMGT-ONTOLOGY CLASSIFICATION Axiom to IMGT Standardized Gene and Allele Nomenclature: For Immunoglobulins (IG) and T Cell Receptors (TR)

Marie-Paule Lefranc[1]

IMGT, the international ImMunoGeneTics information system, Laboratoire d'ImmunoGénétique Moléculaire LIGM, Université Montpellier 2, Institut de Génétique Humaine IGH, UPR CNRS 1142, 34396 Montpellier cedex 5, France

INTRODUCTION

Since the creation of IMGT, the international ImMunoGeneTics information system in 1989, at New Haven during the 10th Human Genome Mapping Workshop (HGM10), the standardized classification and nomenclature of the immunoglobulins (IG) and T cell receptors (TR) of human and other vertebrate species have been under the responsibility of the IMGT Nomenclature Committee (IMGT-NC). In 1995, following the first demonstration online of the nucleotide database IMGT/LIGM-DB at the 9th International Congress of Immunology in San Francisco, IMGT-NC has become the World Health Organization-International Union of Immunological Societies (WHO-IUIS)/IMGT Nomenclature Subcommittee for IG and TR. As described here, IMGT gene and allele names are based on the concepts of classification of "Group," "Subgroup," "Gene," and "Allele," generated from the IMGT-ONTOLOGY CLASSIFICATION axiom. The IMGT gene nomenclature for IG and TR genes was approved at the international level by the Human Genome Organisation (HUGO) Nomenclature Committee (HGNC) in 1999 and by the WHO-IUIS. The IMGT IG and TR gene names are the official reference for the vertebrate genome projects and, as such, have been entered in IMGT/GENE-DB, the IMGT gene database, in Entrez Gene (National Center for Biotechnology Information [NCBI]), in Ensembl (European Bioinformatics Institute [EBI]), and in the Vega Genome Browser (Wellcome Trust Sanger Institute).

RELATED INFORMATION

IMGT/GENE-DB, the IMGT gene database, can be found at http://www.imgt.org. Following approval by the WHO-IUIS/IMGT Nomenclature Subcommittee for IG and TR, "IMGT Locus in Focus" publications have provided overviews on the standardized IMGT nomenclature of the IG and TR genes and alleles, with functionality and polymorphism, according to the IMGT Scientific chart rules. "IMGT Locus in Focus" publications (available as pdf from IMGT Index>IMGT Locus in Focus http://www.imgt.org/textes/IMGTindex/imgtfocus.html) comprise:

- *Homo sapiens* IGHV, IGHD, and IGHJ (Pallarès et al. 1999; Ruiz et al. 1999), IGKV and IGKJ (Barbié and Lefranc 1998), IGLV and IGLJ (Pallarès et al. 1998), TRAV and TRAJ (Scaviner and Lefranc 2000a,b), TRBV, TRBD, and TRBJ (Folch and Lefranc 2000a,b); nomenclature of the *Homo sapiens* IGH (Lefranc 2001a), IGK (Lefranc 2001b), IGL (Lefranc 2001c); protein displays of the human IG (Scaviner et al. 1999) and TR (Folch et al. 2000) variable and joining regions.
- *Mus musculus* IGKV, IGKJ, and IGKC (Martinez and Lefranc 1998; Martinez-Jean et al. 2001), TRAV and TRDV (Bosc and Lefranc 2003), TRBV, TRBD, and TRBJ (Bosc and Lefranc 2000), TRDV, TRDD, and TRDJ (Bosc et al. 2001).
- Teleostei IGH and IGI (Artero and Lefranc 2000a,b).

Abbreviations are explained below in IMGT-ONTOLOGY CONCEPTS OF CLASSIFICATION.

A detailed description of IMGT is provided in **IMGT, The International ImMunoGeneTics Information System** (Lefranc 2011a). Information is also available on **From IMGT-ONTOLOGY IDENTIFICATION Axiom to IMGT Standardized Keywords: For Immunoglobulins (IG), T Cell Receptors**

[1]Corresponding author (Marie-Paule.Lefranc@igh.cnrs.fr).
Cite as: Cold Spring Harb Protoc; 2011; doi:10.1101/pdb.ip84

(TR), and Conventional Genes (Lefranc 2011b), **From IMGT-ONTOLOGY DESCRIPTION Axiom to IMGT Standardized Labels: For Immunoglobulin (IG) and T Cell Receptor (TR) Sequences and Structures** (Lefranc 2011c), **IMGT Unique Numbering for the Variable (V), Constant (C), and Groove (G) Domains of IG, TR, MH, IgSF, and MhSF** (Lefranc 2011d), and **IMGT Collier de Perles for the Variable (V), Constant (C), and Groove (G) Domains of IG, TR, MH, IgSF, and MhSF** (Lefranc 2011e).

In addition, protocols are available for **IMGT/V-QUEST: IMGT Standardized Analysis of the Immunoglobulin (IG) and T Cell Receptor (TR) Nucleotide Sequences** (Giudicelli et al. 2011), **IMGT/JunctionAnalysis: IMGT Standardized Analysis of the V-J and V-D-J Junctions of the Rearranged Immunoglobulins (IG) and T Cell Receptors (TR)** (Giudicelli and Lefranc 2011), **IMGT/Collier de Perles: IMGT Standardized Representation of Domains (IG, TR, and IgSF Variable and Constant Domains, MH and MhSF Groove Domains)** (Ehrenmann et al. 2011), **IMGT/DomainGapAlign: IMGT Standardized Analysis of Amino Acid Sequences of Variable, Constant, and Groove Domains (IG, TR, MH, IgSF, MhSF)** (Ehrenmann and Lefranc 2011a), and **IMGT/3Dstructure-DB: Querying the IMGT Database for 3D Structures in Immunology and Immunoinformatics (IG or Antibodies, TR, MH, RPI, and FPIA)** (Ehrenmann and Lefranc 2011b).

BACKGROUND INFORMATION

The CLASSIFICATION axiom is an axiom of the Formal IMGT-ONTOLOGY or IMGT-Kaleidoscope (Giudicelli and Lefranc 1999; Lefranc et al. 2004, 2005, 2008; Duroux et al. 2008). The CLASSIFICATION axiom postulates that any molecule, cell, tissue, organ, organism, or population, any process and any relation, has to be **classified**. The CLASSIFICATION axiom has generated the concepts of classification of IMGT-ONTOLOGY, the global reference in immunogenetics and immunoinformatics, built by IMGT (http://www.imgt.org) (Lefranc et al. 2009). The CLASSIFICATION axiom is one of seven axioms of the Formal IMGT-ONTOLOGY, the others being "IDENTIFICATION axiom," "DESCRIPTION axiom," "NUMEROTATION axiom," "LOCALIZATION axiom," ORIENTATION axiom," and "OBTENTION axiom." The IMGT gene and allele nomenclature for IG (Lefranc 2000a; Lefranc and Lefranc 2001a) and TR (Lefranc 2000b; Lefranc and Lefranc 2001b) derives from the concepts of classification generated from the CLASSIFICATION axiom, and among them, the concepts of "Group," "Subgroup," "Gene," and "Allele."

IMGT-ONTOLOGY CONCEPTS OF CLASSIFICATION

The IMGT-ONTOLOGY CLASSIFICATION axiom has generated the concepts of classification which have been necessary to propose a standardized nomenclature for the IG and TR genes. These concepts take into account the highly polymorphic multigenic loci and families to which the IG and TR belong, their rearrangements, and their allelic polymorphisms. These concepts are used whatever the antigen receptor (IG or TR), whatever the locus (for mammals, e.g., immunoglobulin heavy IGH, immunoglobulin kappa IGK, immunoglobulin lambda IGL, T cell receptor alpha TRA, T cell receptor beta TRB, T cell receptor gamma TRG, and T cell receptor delta TRD), whatever the gene configuration (germline, undefined, or rearranged), and whatever the species, from fish to human. Among the concepts of classification, the "Group," "Subgroup," "Gene," and "Allele" concepts are essential (Fig. 1):

"Group" is a concept of classification that allows one to classify a set of genes which belong to the same multigene family, within the same species or between different species. For the IG and TR, the set of genes is identified by the "GeneType" concept: V (for "variable"), D (for "diversity"), J (for "joining"), or C (for "constant"), e.g., **"IGHV"** (Fig. 1). In *Homo sapiens*, the "Group" concept has 24 leafconcepts (a leafconcept is a concept that corresponds to the finest level of granularity), of which 10 are for IG ("IGHV," "IGHD," "IGHJ," "IGHC," "IGKV," "IGKJ," "IGKC," "IGLV," "IGLJ," "IGLC") and 14 are for TR ("TRAV," "TRAJ," "TRAC," "TRBV," "TRBD," "TRBJ," "TRBC," "TRGV," "TRGJ," "TRGC," "TRDV," "TRDD," "TRDJ," "TRDC").

"**Subgroup**" is a concept of classification that allows one to classify a subset of genes which belong to the same group, and which, in a given species, share at least 75% of identity at the nucleotide sequence level (and in the germline configuration for the V, D, and J genes), for instance "***Homo sapiens* IGHV1**" (Fig. 1).

"**Gene**" is a concept of classification that allows one to classify a unit of DNA sequence that can be potentially transcribed and/or translated (this definition includes the regulatory elements in 5′ and 3′,

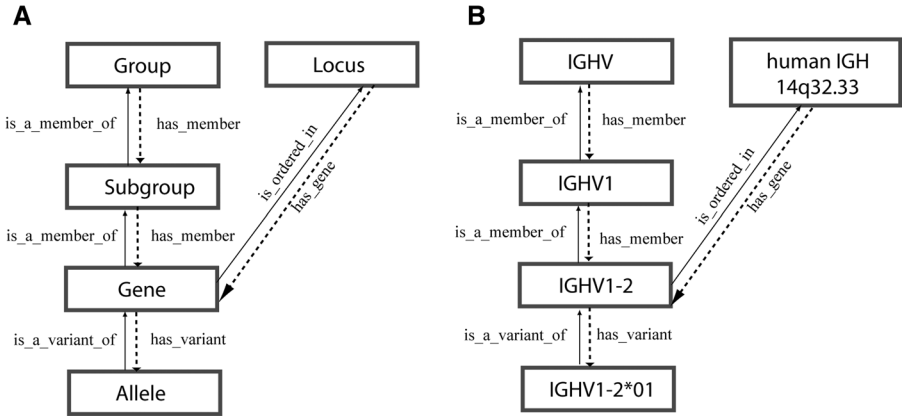


FIGURE 1. Concepts of classification for gene and allele nomenclature (generated from the IMGT-ONTOLOGY CLASSIFICATION axiom). (*A*) Hierarchy of the concepts of classification and their relations. (*B*) Examples of leafconcepts for each concept of classification. They are associated with a "TaxonRank" level, and more precisely for the "Gene" and "Allele" concepts with a leafconcept of "Species" (here, *Homo sapiens*). The "Locus" concept is a concept of localization (LOCALIZATION axiom).

and the introns, if present). The leafconcepts of the "Gene" concept are gene names. In IMGT-ONTOLOGY, a gene name is composed of the name of the species (leafconcept of the "Species" concept) and of the international HGNC/IMGT gene symbol, e.g., "***Homo sapiens* IGHV1-2**" (human immunoglobulin heavy variable 1–2) (Fig. 1). By extension, orphons and pseudogenes are also instances of the "Gene" concept. For example, in December 2010, the "Gene" concept had 685 *Homo sapiens* IG and TR gene names (leafconcepts) (of which 442 were for IG and 243 were for TR).

"**Allele**" is a concept of classification that allows one to classify a polymorphic variant of a gene. The leafconcepts of the "Allele" concept are allele names. In IMGT-ONTOLOGY, an allele name identified by mutations at the nucleotide level is composed of the IMGT gene name with an asterisk (*) and a two-figure number (for instance, "***Homo sapiens* IGHV1-2*01**") (Fig. 1). The alleles are classified by reference to allele *01. Full description of mutations and allele name designations are currently recorded for the core sequences (V-REGION, D-REGION, J-REGION, C-REGION). For example, in December 2010, the "Allele" concept had 1254 *Homo sapiens* IG and TR allele names (leafconcepts) (of which 832 were for IG and 422 were for TR).

IMGT SCIENTIFIC CHART RULES FOR NOMENCLATURE

IMGT nomenclature for IG and TR gene names of human and other vertebrates, based on the IMGT-ONTOLOGY concepts of classification, follows as closely as possible the rules of the HUGO Gene Nomenclature Committee (http://www.genenames.org/). This was applied as early as 1988 for the human IGL and IGH loci (Bensmana et al. 1988; Ghanem et al. 1988) and 1989 for all the genes of the human TRG locus (Lefranc 1989a,b; Lefranc and Rabbitts 1990): **Gene names in capital letters, no greek letters, no commas, no dots** (hyphens are accepted, as for the human leucocyte antigens [HLA] and major histocompatibility [MH] genes). The three first letters indicate the locus: IGH, IGK, IGL, TRA, TRB, TRG, TRD. The fourth letter is V, D, J, or C (except for IGH, see below). The following number(s) and/or letter(s) allow, if necessary, unambiguous identification of the gene, a single number or letter being used whenever it is possible (IGKC, TRGV9, TRGVA, TRAJ4…).

- For the IGH locus, the constant genes are designated by the letter (and eventually number) corresponding to the encoded isotype (IGHM, IGHD, IGHG3,…), instead of using the letter C.
- For the more complex situation where genes have not yet been localized and have to be named accordingly to their subgroup (for V) or cluster (for D and J), a temporary designation is used in which the Arabic number (for the subgroup or cluster) is followed by the letter S (for subgroup or sequential, respectively), in turn followed by the number of the gene in the subgroup or cluster.

- Orphon genes are designated by the IMGT four letter gene designation, followed by a number for the subgroup (if known), followed by a slash, OR (for orphon), the chromosome number (if known), a dash, and a specific gene number and/or letter.
- Examples include IGKV2/OR2-1, IGHD1/OR15-1a, IGKV1/OR-1. In the IG and TR HGNC symbols, slashes and parentheses are omitted, and capital letters replace the lower-case letters found in some provisional IMGT gene names. Otherwise, the gene symbols and all the full names (including slashes and parentheses) are identical in IMGT and HGNC nomenclatures.

Genes and alleles are recorded in IMGT/GENE-DB (Giudicelli et al. 2005) and displayed in Alignments of alleles (IMGT Repertoire, http://www.imgt.org).

INTEROPERABILITY BETWEEN IMGT, HGNC, AND NCBI GENE

Since the creation of IMGT, in 1989, at New Haven during HGM10, the standardized classification and nomenclature of the IG and TR of human and other vertebrate species have been under the responsibility of the IMGT-NC. The WHO-IUIS/IMGT Nomenclature Subcommittee for IG and TR has received official delegation from HGNC for the IG and TR genes and alleles (Wain et al. 2002). The IMGT gene nomenclature for IG and TR genes (Lefranc 2000a,b; Lefranc and Lefranc 2001a,b) was approved by the HGNC in 1999 and by the WHO-IUIS (Lefranc 2007, 2008). IMGT IG and TR gene names have been entered in IMGT/GENE-DB (Giudicelli et al. 2005), the IMGT gene database, in the Human Genome Database (GDB), in LocusLink at the NCBI in 1999–2000, in Entrez Gene (NCBI) (Maglott et al. 2007) when this gene database superseded LocusLink, in MapViewer (NCBI), in Ensembl at the EBI in 2006, and in the Vega Genome Browser at the Wellcome Trust Sanger Institute.

The IMGT nomenclature is acknowledged as the official standard for IG and TR by the scientific community (Glusman et al. 2001; Lefranc 2005, 2007, 2008, 2011f) and the IMGT IG and TR gene names are the official reference for the vertebrate genome projects. Indeed the antigen receptors (IG and TR) that characterize the adaptive immune response of the vertebrates with jaws (gnathostomes) have specific characteristics compared to the conventional genes. It is through the IMGT-ONTOLOGY concepts that the IG and TR genes are now managed in the genome databases. The first unified presentation of the locus "maps" of the seven human loci (IGH, IGK, IGL, TRA, TRB, TRG, and TRD) (Lefranc 2000c,d) has become the standard for the IG and TR locus representation in other vertebrate species (Bosc and Lefranc 2003; Lefranc and Lefranc 2004) (IMGT Repertoire, http://www.imgt.org).

To facilitate interoperability between the databases, IMGT works in close collaboration with HGNC (Wain et al. 2004; Bruford et al. 2008), NCBI Gene (Maglott et al. 2007), the Mouse Genomic Nomenclature Committee (MGNC) (to which IMGT provided all the mouse IG and TR genes in 2002), the Nomenclature Committees of newly sequenced genomes, the national and international societies (Immunology, Immunogenetics, Genetics, Antibody), and editors and publishers for recommendations to authors.

ACKNOWLEDGMENTS

We thank Véronique Giudicelli and Gérard Lefranc and the IMGT team for their motivation and expertise.

REFERENCES

Artero S, Lefranc M-P. 2000a. The teleostei immunoglobulin heavy IGH genes. *Exp Clin Immunogene* **17:** 148–161.

Artero S, Lefranc M-P. 2000b. The teleostei immunoglobulin light IGL1 and IGL2 V, J and C genes. *Exp Clin Immunogenet* **17:** 162–172.

Barbié V, Lefranc M-P. 1998. The human immunoglobulin kappa variable (IGKV) genes and joining (IGKJ) segments. *Exp Clin Immunogenet* **15:** 171–183.

Bensmana M, Huck S, Lefranc G, Lefranc M-P. 1988. The human immunoglobulin pseudo-gamma IGHGP gene shows no major structural defect. *Nucl Acids Res* **16:** 3108.

Bosc N, Lefranc M-P. 2000. The mouse (*Mus musculus*) T-cell receptor beta variable (TRBV), diversity (TRBD) and joining (TRBJ) genes. *Exp Clin Immunogenet* **17:** 216–228.

Bosc N, Lefranc M-P. 2003. IMGT locus in focus: The mouse *(Mus musculus)* T cell receptor alpha (TRA) and delta (TRD) variable genes. *Dev Comp Immunol* **27:** 465–497.

Bosc N, Contet V, Lefranc M-P. 2001. The mouse (*Mus musculus*) T cell receptor delta variable (TRDV), diversity (TRDD) and joining (TRDJ) genes. *Exp Clin Immunogenet* **18:** 51–58.

Bruford EA, Lush MJ, Wright MW, Sneddon TP, Povey S, Birney E. 2008. The HGNC database in 2008: A resource for the human genome. *Nucl Acids Res* **36:** D445–D448.

Duroux P, Kaas Q, Brochet X, Lane J, Ginestoux C, Lefranc M-P, Giudicelli V. 2008. IMGT-kaleidoscope, the formal IMGT-ONTOLOGY paradigm. *Biochimie* **90:** 570–583.

Ehrenmann F, Lefranc M-P. 2011a. IMGT/DomainGapAlign: IMGT standardized analysis of amino acid sequences of variable,

constant, and groove domains (IG, TR, MH, IgSF, MhSF). *Cold Spring Harb Protoc* doi: 10.1101/pdb.prot5636.

Ehrenmann F, Lefranc M-P. 2011b. IMGT/3Dstructure-DB: Querying the IMGT database for 3D structures in immunology and immunoinformatics (IG or antibodies, TR, MH, RPI, and FPIA). *Cold Spring Harb Protoc* doi: 10.1101/pdb.prot5637.

Ehrenmann F, Giudicelli V, Brochet X, Lefranc M-P. 2011. IMGT/Collier de Perles: IMGT standardized representation of domains (IG, TR, and IgSF variable and constant domains, MH and MhSF groove domains). *Cold Spring Harb Protoc* doi: 10.1101/pdb.prot5635.

Folch G, Lefranc M-P. 2000a. The human T cell Receptor Beta Variable (TRBV) genes. *Exp Clin Immunogenet* **17:** 42–54.

Folch G, Lefranc M-P. 2000b. The human T cell Receptor Beta Diversity (TRBD) and Beta Joining (TRBJ) genes. *Exp Clin Immunogenet* **17:** 107–114.

Folch G, Scaviner D, Contet V, Lefranc M-P. 2000. Protein displays of the Human T cell Receptor Alpha, Beta, Gamma and Delta variable and joining regions. *Exp Clin Immunogenet* **17:** 205–215.

Ghanem N, Dariavach P, Bensmana M, Chibani J, Lefranc M-P, Lefranc G. 1988. Polymorphism of immunoglobulin lambda constant region genes in populations from France, Lebanon and Tunisia. *Exp Clin Immunogenet* **5:** 186–195.

Giudicelli V, Lefranc M-P. 1999. Ontology for immunogenetics: The IMGT-ONTOLOGY. *Bioinformatics* **12:** 1047–1054.

Giudicelli V, Lefranc M-P. 2011. IMGT/JunctionAnalysis: IMGT standardized analysis of the V-J and V-D-J junctions of the rearranged immunoglobulins (IG) and T cell receptors (TR). *Cold Spring Harb Protoc* doi: 10.1101/pdb.prot5634.

Giudicelli V, Chaume D, Lefranc M-P. 2005. IMGT/GENE-DB: A comprehensive database for human and mouse immunoglobulin and T cell receptor genes. *Nucl Acids Res* **33:** D256–D261.

Giudicelli V, Brochet X, Lefranc M-P. 2011. IMGT/V-QUEST: IMGT standardized analysis of the immunoglobulin (IG) and T cell receptor (TR) nucleotide sequences. *Cold Spring Harb Protoc* doi: 10.1101/pdb.prot5633.

Glusman G, Rowen L, Lee I, Boysen C, Roach JC, Smit AF, Wang K, Koop BF, Hood L. 2001. Comparative genomics of the human and mouse T cell receptor loci. *Immunity* **15:** 337–349.

Lefranc M-P. 1989a. The joining segments and constant region genes of the human T-cell receptor gamma chain and linkage of the variable and constant regions. Human Gene Mapping 10. Tenth International Workshop on Human Gene Mapping. New Haven, USA, 11-17 June 1989. *Cytogenet Cell Genet* **51:** 1031 (A2337).

Lefranc M-P. 1989b. The variable region genes of the human T-cell receptor gamma chain Human Gene Mapping 10. Tenth International Workshop on Human Gene Mapping. New Haven, USA, 11-17 June 1989. *Cytogenet Cell Genet* **51:** 1031 (A2338).

Lefranc M-P. 2000a. Nomenclature of the human immunoglobulin genes. In *Current protocols in immunology* (ed. JE Coligan et al.), pp. A.1P.1–A.1P.37. John Wiley and Sons, Inc, Hoboken, NJ.

Lefranc M-P. 2000b. Nomenclature of the human T cell receptor genes. In *Current protocols in immunology* (ed. JE Coligan et al.), pp. A.1O.1–A.1O.23. John Wiley and Sons, Inc, Hoboken, NJ.

Lefranc M-P. 2000c. Locus maps and genomic repertoire of the human T cell receptor genes. *The Immunologist* **8:** 72–79.

Lefranc M-P. 2000d. Locus maps and genomic repertoire of the human immunoglobulin genes. *The Immunologist* **8:** 80–87.

Lefranc M-P. 2001a. Nomenclature of the human immunoglobulin heavy (IGH) genes. *Exp Clin Immunogenet* **18:** 100–116.

Lefranc M-P. 2001b. Nomenclature of the human immunoglobulin kappa (IGK) genes. *Exp Clin Immunogenet* **18:** 161–174.

Lefranc M-P. 2001c. Nomenclature of the human immunoglobulin lambda (IGL) genes. *Exp Clin Immunogenet* **18:** 242–254.

Lefranc M-P. 2005. IMGT, the international ImMunoGeneTics information system: A standardized approach for immunogenetics and immunoinformatics. *Immunome Res* **1:** 3. doi: 10.1186/1745-7580-1-3.

Lefranc M-P. 2007. WHO-IUIS Nomenclature Subcommittee for immunoglobulins and T cell receptors report. *Immunogenetics* **59:** 899–902.

Lefranc M-P. 2008. WHO-IUIS Nomenclature Subcommittee for immunoglobulins and T cell receptors report August 2007, 13th International Congress of Immunology, Rio de Janeiro, Brazil. *Dev Comp Immunol* **32:** 461–463.

Lefranc M-P. 2011a. IMGT, the international ImMunoGeneTics information system. *Cold Spring Harb Protoc* doi: 10.1101/pdb.top115.

Lefranc M-P. 2011b. From IMGT-ONTOLOGY IDENTIFICATION axiom to IMGT standardized keywords: For immunoglobulins (IG), T cell receptors (TR), and conventional genes. *Cold Spring Harb Protoc* doi: 10.1101/pdb.ip82.

Lefranc M-P. 2011c. From IMGT-ONTOLOGY DESCRIPTION axiom to IMGT standardized labels: For immunoglobulin (IG) and T cell receptor (TR) sequences and structures. *Cold Spring Harb Protoc* doi: 10.1101/pdb.ip83.

Lefranc M-P. 2011d. IMGT unique numbering for the variable (V), constant (C), and groove (G) domains of IG, TR, MH, IgSF, and MhSF. *Cold Spring Harb Protoc* doi: 10.1101/pdb.ip85.

Lefranc M-P. 2011e. IMGT Collier de Perles for the variable (V), constant (C), and groove (G) domains of IG, TR, MH, IgSF, and MhSF. *Cold Spring Harb Protoc* doi: 10.1101/pdb.ip86.

Lefranc M-P. 2011f. Antibody nomenclature: From IMGT-ONTOLOGY to INN definition. *MAbs* **3:** 1–2.

Lefranc M-P, Lefranc G. 2001a. *The immunoglobulin FactsBook*. Academic Press, London.

Lefranc M-P, Lefranc G. 2001b. *The T cell receptor FactsBook*. Academic Press, London.

Lefranc M-P, Lefranc G. 2004. Immunoglobulin lambda (IGL) genes of human and mouse. In *Molecular biology of B cells* (ed. T Honjo et al.), pp. 37–59. Elsevier Academic Press, London.

Lefranc M-P, Rabbitts TH. 1990. A nomenclature to fit the organization of the human T cell receptor gamma and delta genes. *Res Immunol* **141:** 615–618.

Lefranc M-P, Giudicelli V, Ginestoux C, Bosc N, Folch G, Guiraudou D, Jabado-Michaloud J, Magris S, Scaviner D, Thouvenin V, et al. 2004. IMGT-ONTOLOGY for Immunogenetics and Immunoinformatics. *In Silico Biol* **4:** 17–29.

Lefranc M-P, Clément O, Kaas Q, Duprat E, Chastellan P, Coelho I, Combres K, Ginestoux C, Giudicelli V, Chaume D, et al. 2005. IMGT-Choreography for immunogenetics and immunoinformatics. *In Silico Biol* **5:** 45–60.

Lefranc M-P, Giudicelli V, Regnier L, Duroux P. 2008. IMGT, a system and an ontology that bridge biological and computational spheres in bioinformatics. *Brief Bioinform* **9:** 263–275.

Lefranc M-P, Giudicelli V, Ginestoux C, Jabado-Michaloud J, Folch G, Bellahcene F, Wu Y, Gemrot E, Brochet X, Lane J, et al. 2009. IMGT, the international ImMunoGeneTics information system. *Nucl Acids Res* **37:** D1006–D1012.

Maglott D, Ostell J, Pruitt KD, Tatusova T. 2007. Entrez Gene: Gene-centered information at NCBI. *Nucl Acids Res* **35:** D26–D31.

Martinez C, Lefranc M-P. 1998. The mouse (*Mus musculus*) immunoglobulin kappa variable (IGKV) genes and joining (IGKJ) segments. *Exp Clin Immunogenet* **15:** 184–193.

Martinez-Jean C, Folch G, Lefranc M-P. 2001. Nomenclature and overview of the mouse (*Mus musculus* and *Mus* sp.) immunoglobulin kappa (IGK) genes. *Exp Clin Immunogenet* **18:** 255–279.

Pallarès N, Frippiat J-P, Giudicelli V, Lefranc M-P. 1998. The human immunoglobulin lambda variable (IGLV) genes and joining (IGLJ) segments. *Exp Clin Immunogenet* **15:** 8–18.

Pallarès N, Lefebvre S, Contet V, Mastuda F, Lefranc M-P. 1999. The human immunoglobulin heavy variable (IGHV) genes. *Exp Clin Immunogenet* **16:** 36–60.

Ruiz M, Pallarès N, Contet V, Barbié V, Lefranc M-P. 1999. The human immunoglobulin heavy diversity (IGHD) and joining (IGHJ) segments. *Exp Clin Immunogenet* **16:** 173–184.

Scaviner D, Lefranc M-P. 2000a. The human T cell Receptor Alpha Variable (TRAV) genes. *Exp Clin Immunogenet* **17:** 83–96.

Scaviner D, Lefranc M-P. 2000b. The human T cell Receptor Alpha Joining (TRAJ) genes. *Exp Clin Immunogenet* **17:** 97–106.

Scaviner D, Barbié V, Ruiz M, Lefranc M-P. 1999. Protein displays of the human immunoglobulin heavy, kappa and lambda variable and joining regions. *Exp Clin Immunogenet* **16:** 234–240.

Wain HM, Bruford EA, Lovering RC, Lush MJ, Wright MW, Povey S. 2002. Guidelines for human gene nomenclature. *Genomics* **79:** 464–470.

Wain HM, Lush MJ, Ducluzeau F, Khodiyar VK, Povey S. 2004. Genew: The Human Gene Nomenclature Database, 2004 updates. *Nucl Acids Res* **32:** D255–D257.

Information Panel

IMGT Unique Numbering for the Variable (V), Constant (C), and Groove (G) Domains of IG, TR, MH, IgSF, and MhSF

Marie-Paule Lefranc[1]

IMGT, the international ImMunoGeneTics information system, Laboratoire d'ImmunoGénétique Moléculaire LIGM, Université Montpellier 2, Institut de Génétique Humaine IGH, UPR CNRS 1142, 34396 Montpellier cedex 5, France

INTRODUCTION

The "IMGT unique numbering" (or "IMGT_unique_numbering") concept is a major concept of numerotation (generated from the NUMEROTATION axiom) of IMGT-ONTOLOGY, the global reference in immunogenetics and immunoinformatics, built by IMGT, the international ImMunoGeneTics information system. The "IMGT unique numbering" concept, described here, allows one to number domain types that are characteristic of protein superfamilies, whatever the species, the molecule type or the chain type. Three leafconcepts (a leafconcept is a concept that corresponds to the finest level of granularity) have been defined, respectively, for the variable (V) domain and constant (C) domain of the immunoglobulin superfamily (IgSF) and for the groove (G) domain of the major histocompatibility (MH) superfamily (MhSF). The IMGT unique numbering concept has been a great breakthrough in immunogenetics and systems biology in allowing, for the first time, the bridging of the gap between amino acid (and nucleotide) sequences of any V, C, or G domain and their two-dimensional (2D) and three-dimensional (3D) structures. The IMGT unique numbering concept has been fundamental in the creation of the IMGT Collier de Perles and in the standardization of the description of mutations, amino acid changes, polymorphisms, and contact analysis in the IMGT databases, tools, and web resources.

RELATED INFORMATION

A detailed description of IMGT is provided in **IMGT, the International ImMunoGeneTics Information System** (Lefranc 2011a). The NUMEROTATION axiom is an axiom of the Formal IMGT-ONTOLOGY or IMGT-Kaleidoscope (Giudicelli and Lefranc 1999; Lefranc et al. 2004, 2005a, 2008; Duroux et al. 2008). The NUMEROTATION axiom postulates that any molecule, cell, tissue, organ, organism, or population, any process and any relation, has to be numbered. The NUMEROTATION axiom has generated the concepts of numerotation of IMGT-ONTOLOGY, the global reference in immunogenetics and immunoinformatics, built by IMGT (http://www.imgt.org; Lefranc et al. 2009). The NUMEROTATION axiom is one of seven axioms of the Formal IMGT-ONTOLOGY, the others being "IDENTIFICATION axiom" (From IMGT-ONTOLOGY IDENTIFICATION Axiom to IMGT Standardized Keywords: **For Immunoglobulins (IG), T Cell Receptors (TR), and Conventional Genes** [Lefranc 2011b]), "DESCRIPTION axiom" (**From IMGT-ONTOLOGY DESCRIPTION Axiom to IMGT Standardized Labels: For Immunoglobulin (IG) and T Cell Receptor (TR) Sequences and Structures** [Lefranc 2011c]), "CLASSIFICATION axiom" (**From IMGT-ONTOLOGY CLASSIFICATION Axiom to IMGT Standardized Gene and Allele Nomenclature: For Immunoglobulins (IG) and T Cell Receptors (TR)** [Lefranc 2011d]), "LOCALIZATION axiom," "ORIENTATION axiom," and "OBTENTION axiom." The IMGT unique numbering (Lefranc 1997, 1999; Lefranc et al. 2003, 2005b,c) is one of the major concepts of numerotation, generated from the NUMEROTATION axiom, the other being the IMGT Collier de Perles (**IMGT Collier de Perles for the Variable (V), Constant (C), and Groove (G) Domains of IG, TR, MH, IgSF, and MhSF** [Lefranc 2011e]).

Protocols are available for **IMGT/V-QUEST: IMGT Standardized Analysis of the Immunoglobulin (IG) and T Cell Receptor (TR) Nucleotide Sequences** (Giudicelli et al. 2011), **IMGT/JunctionAnalysis: IMGT Standardized Analysis of the V-J and V-D-J Junctions of the Rearranged Immunoglobulins**

[1]Corresponding author (Marie-Paule.Lefranc@igh.cnrs.fr).
Cite as: Cold Spring Harb Protoc; 2011; doi:10.1101/pdb.ip85

(IG) and T Cell Receptors (TR) (Giudicelli and Lefranc 2011), **IMGT/Collier de Perles: IMGT Standardized Representation of Domains (IG, TR, and IgSF Variable and Constant Domains, MH and MhSF Groove Domains)** (Ehrenmann et al. 2011), **IMGT/DomainGapAlign: IMGT Standardized Analysis of Amino Acid Sequences of Variable, Constant, and Groove Domains (IG, TR, MH, IgSF, MhSF)** (Ehrenmann and Lefranc 2011a), and **IMGT/3Dstructure-DB: Querying the IMGT Database for 3D Structures in Immunology and Immunoinformatics (IG or Antibodies, TR, MH, RPI, and FPIA)** (Ehrenmann and Lefranc 2011b).

BACKGROUND INFORMATION

The immunoglobulin (IG) and T cell receptor (TR) variable domains form a huge repertoire of 2×10^{12} antigen receptors per individual and, owing to the particularities of their synthesis that involve DNA rearrangements (Fig. 1), there was a need for a systematic and coherent numbering of the amino acids and codons, whatever the molecule, configuration, or chain type. It was therefore a great breakthrough when the "IMGT unique numbering" and its representation, the IMGT Collier de Perles (**IMGT Collier de Perles for the Variable (V), Constant (C), and Groove (G) Domains of IG, TR, MH, IgSF, and MhSF** [Lefranc 2011e]), were defined for the first time in 1997 for the V domain (Lefranc 1997, 1999). This IMGT unique numbering was created by taking into account the high conservation of the structure of the V domain and by integrating knowledge acquired by the analysis of multiple sources: alignment of >5000 sequences, literature data on the framework (FR) and complementarity determining regions (CDR), structural data from X-ray diffraction studies, and characterization of the CDR hypervariable loops (Lefranc 1997, 1999). The standardized delimitation of the FR-IMGT and CDR-IMGT, a new and crucial feature of this numbering, was defined based on the longest CDR1-IMGT and CDR2-IMGT found in the IMGT multiple alignments of the germline IG and TR genes and, for the rearrranged CDR3-IMGT, on statistical analysis of the IG and TR rearrangements (Lefranc 1997, 1999; Lefranc et al. 2003). The IMGT unique numbering, originally defined for the numerotation of the IG and TR V-DOMAIN (Lefranc 1997), was rapidly extended to the V-LIKE-DOMAIN of the IgSF other than IG and TR (Lefranc 1999; Lefranc et al. 2003), then to the C domain (C-DOMAIN of IG and TR and C-LIKE-DOMAIN of IgSF other than IG and TR) (Lefranc et al. 2005a). Based on the same concepts, and despite a very different structure of the G domain, the IMGT unique numbering for G domain

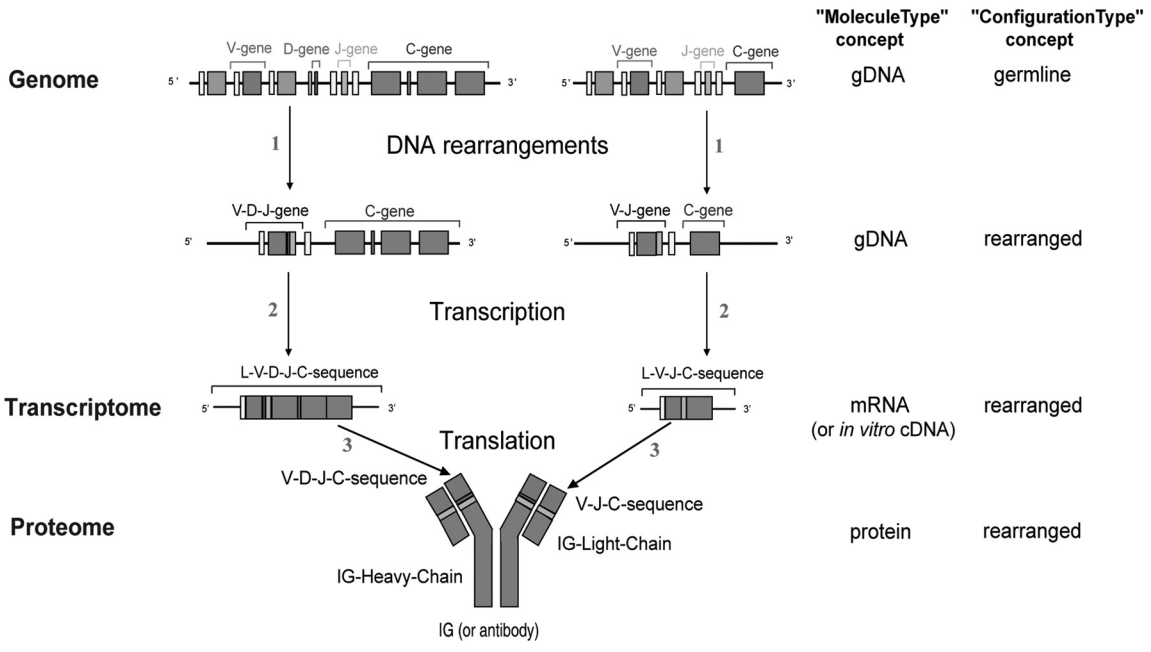


FIGURE 1. Synthesis of an IG or antibody in humans. A human being may potentially synthesize 10^{12} different antibodies (Lefranc and Lefranc 2001a). 1, DNA rearrangements (is_rearranged_into); 2, Transcription (is_transcribed_into); 3, Translation (is_translated_into). The configuration of C-GENE is "undefined" (ConfigurationType) (IMGT Repertoire, http://www.imgt.org).

was successfully set up for the G-DOMAIN of MH and G-LIKE-DOMAIN of MhSF other than MH (Lefranc et al. 2005b).

IMGT UNIQUE NUMBERING FOR V DOMAIN

V Domain Strands and Loops

The IMGT unique numbering for V domain numbers the V-DOMAIN of IG or antibodies and TR and the V-LIKE-DOMAIN of the IgSF other than IG and TR. The V domain strands and loops and their IMGT positions and lengths, based on the IMGT unique numbering for V domain (V-DOMAIN and V-LIKE-DOMAIN) (Lefranc et al. 2003), are shown in Table 1.

The V domain (V-DOMAIN of IG and TR and V-LIKE-DOMAIN of IgSF other than IG and TR) is composed of the A-STRAND of 15 (or 14 if gap at position 10) amino acids (positions 1 to 15), the B-STRAND of 11 amino acids (positions 16 to 26) with the first conserved cysteine (1st-CYS) at position 23, the BC-LOOP (positions 27 to 38; the longest BC loops have 12 amino acids), the C-STRAND of eight amino acids (positions 39 to 46) with the tryptophan (CONSERVED-TRP) at position 41, the C′-STRAND of nine amino acids (positions 47 to 55), the C′C″-LOOP (positions 56 to 65; the longest C′C″ loops have 10 amino acids), the C″-STRAND of nine (or eight if gap at position 73) amino acids (positions 66 to 74), the D-STRAND of 10 (or eight if gaps at positions 81 and 82) amino acids (positions 75 to 84), the E-STRAND of 12 amino acids (positions 85 to 96) with a conserved hydrophobic amino acid at position 89, the F-STRAND of eight amino acids (positions 97 to 104) with the second conserved cystein (2nd-CYS) at position 104, the FG-LOOP (positions 105 to 117; these positions correspond to a FG loop of 13 amino acids), and the G-STRAND of 11 (or 10) amino acids (positions 118 to 128) (Table 1; Fig. 2). In the IG and TR V-DOMAIN, the G-STRAND is the C-terminal part of the J-REGION, with J-PHE or J-TRP 118 and the canonical motif F/W-G-X-G at positions 118–121 (Table 1; Lefranc et al. 2003).

In the IG and TR V-DOMAIN, the structurally conserved antiparallel beta strands are also designated as framework regions (FR-IMGT), whereas the loops are designated as complementarity determining

Table 1. V domain strands and loops, IMGT positions and lengths, based on the IMGT unique numbering for V domain (V-DOMAIN and V-LIKE-DOMAIN)

V domain strands and loops[a]	IMGT positions	Lengths[b]	Characteristic Residue@Position[c]	V-DOMAIN FR-IMGT and CDR-IMGT
A-STRAND	1–15	15 (14 if gap at 10)		FR1-IMGT
B-STRAND	16–26	11	1st-CYS 23	
BC-LOOP	27–38	12 (or less)		CDR1-IMGT
C-STRAND	39–46	8	CONSERVED-TRP 41	FR2-IMGT
C′-STRAND	47–55	9		
C′C″-LOOP	56–65	10 (or less)		CDR2-IMGT
C″-STRAND	66–74	9 (or 8 if gap at 73)		FR3-IMGT
D-STRAND	75–84	10 (or 8 if gaps at 81, 82)		
E-STRAND	85–96	12	hydrophobic 89	
F-STRAND	97–104	8	2nd-CYS 104	
FG-LOOP	105–117	13 (or less, or more)		CDR3-IMGT
G-STRAND	118–128	11 (or 10)	V-DOMAIN J-PHE 118 or J-TRP 118[d]	FR4-IMGT

From Lefranc et al. (2003).

[a]IMGT labels (concepts of description) are written in capital letters (**From IMGT-ONTOLOGY DESCRIPTION Axiom to IMGT Standardized Labels: For Immunoglobulin (IG) and T Cell Receptor (TR) Sequences and Structures** [Lefranc 2011c]).

[b]In number of amino acids (or codons).

[c]Residue@Position is an IMGT concept of numerotation that numbers the position of a given residue (or that of a conserved property amino acid class), based on the IMGT unique numbering.

[d]In the IG and TR V-DOMAIN, the G-STRAND (or FR4-IMGT) is the C-terminal part of the J-REGION, with J-PHE or J-TRP 118 and the canonical motif F/W-G-X-G at positions 118–121.

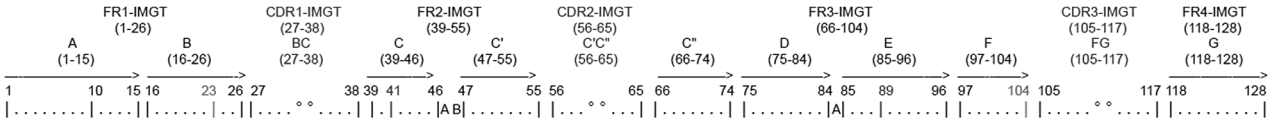


FIGURE 2. "IMGT Protein display for V domain" header. The header comprises seven lines. Line 1: Framework regions FR-IMGT (e.g. FR1-IMGT) and complementarity determining regions CDR-IMGT (e.g. CDR1-IMGT). Line 2: start and end positions of FR-IMGT and CDR-IMGT, e.g. (1–26). Line 3: name of strands (A, B…) and loops (AB, BC…). Line 4: start and end positions of strands, e.g. (1–15). Line 5: arrows (for the strands). Line 6: positions according to the IMGT unique numbering for V-DOMAIN and V-LIKE-DOMAIN (Lefranc 1999; Lefranc et al. 2003). Line 7: Pipes for the start and end positions of strands and loops and highlighted positions, dots for the other positions, plus if needed, letters or numbers for additional positions (for example AB or 123…). Optionally small circles above the line (instead of dots) may indicate the positions at the top of the BC, C′C″, and FG loops considering loop lengths of [12.10.13] (these positions are 32, 33 for BC, 60, 61 for C′C″, and 111, 112 for FG).

regions (CDR-IMGT) (Lefranc et al. 2003). Strands A and B correspond to the FR1-IMGT (positions 1 to 26), strands C and C′ to the FR2-IMGT (positions 39 to 55), strands C″, D, E, and F to the FR3-IMGT (positions 66 to 104), and strand G to the FR4-IMGT (positions 118 to 128). The BC, C′C″, and FG loops correspond to the CDR1-IMGT, CDR2-IMGT, and CDR3-IMGT, respectively (Table 1; Fig. 2; Lefranc et al. 2003).

The loop length (number of amino acids (or codons), that is the number of occupied positions) is a crucial and original concept of IMGT-ONTOLOGY. The lengths of the BC (CDR1-IMGT), C′C″ (CDR2-IMGT), and FG (CDR3-IMGT) loops characterize the V-DOMAIN and V-LIKE-DOMAIN. Thus, the length of the three loops BC, C′C″, and FG is shown, in number of amino acids (or codons), in brackets and separated by dots. For example [9.6.9] means that the BC, C′C″, and FG loops (or CDR1-IMGT, CDR2-IMGT, and CDR3-IMGT for V-DOMAIN) have a length of 9, 6, and 9 amino acids (or codons), respectively.

IMGT Gaps and Additional Positions

Dots in IMGT Protein displays (Fig. 2; Scaviner et al. 1999; Folch et al. 2000; Lefranc and Lefranc 2001a,b) and hatched circles or squares in IMGT Colliers de Perles for V domain (**IMGT Collier de Perles for the Variable (V), Constant (C), and Groove (G) Domains of IG, TR, MH, IgSF, and MhSF** [Lefranc 2011e]) correspond to missing positions according to the IMGT unique numbering for V domain (Lefranc et al. 2003).

For BC, C′C″, or FG loops shorter than 12, 10, 13 amino acids, respectively, gaps are created at the apex (missing positions, hatched in IMGT Collier de Perles, or not shown in structural data representations). The gaps are placed at the apex of the loop with an equal number of amino acids (or codons) on both sides if the loop length is an even number, or with one more amino acid (or codon) in the left part if it is an odd number. For example, for FG loops <13 amino acids, gaps are created from the apex of the loop, in the following order: 111, 112, 110, 113, 109, etc. For FG loops >13 amino acids, additional positions are created, between positions 111 and 112 at the top of the FG loop, in the following order: 112.1, 111.1, 112.2, 111.2, 112.3, etc. (Lefranc et al. 2003).

IMGT UNIQUE NUMBERING FOR C DOMAIN

C Domain Strands, Loops, and Turns

The IMGT unique numbering for C domain numbers the C-DOMAIN of IG or antibodies and TR and the C-LIKE-DOMAIN of the IgSF other than IG and TR. The C domain strands, turns, and loops and their IMGT positions and lengths, based on the IMGT unique numbering for C domain (C-DOMAIN and C-LIKE-DOMAIN) (Lefranc et al. 2005a), are shown in Table 2.

The C domain (C-DOMAIN of IG and TR and C-LIKE-DOMAIN of IgSF other than IG and TR) is composed of the A-STRAND of 15 (or 14 if gap at 10) amino acids (positions 1 to 15), the AB-TURN (additional positions 15.1, 15.2, and 15.3; the longest AB-TURN have three amino acids), the B-STRAND of 11 amino acids (positions 16 to 26) with the 1st-CYS at position 23, the BC-LOOP (positions 27 to 31, 34 to 38), the C-STRAND of seven amino acids (positions 39 to 45) with the CONSERVED-TRP at position 41, the CD-STRAND of one to nine amino acids (additional positions 45.1 to 45.9), the D-STRAND of eight (or seven if gap at 82) amino acids (positions 77 to 84), the DE-TURN (additional

Table 2. C domain strands, turns, and loops, IMGT positions and lengths, based on the IMGT unique numbering for C domain (C-DOMAIN and C-LIKE-DOMAIN)

C domain strands, turns and loops[a]	IMGT positions	Lengths[b]	Characteristic Residue@Position[c]
A-STRAND	1–15	15 (14 if gap at 10)	
AB-TURN	15.1–15.3	0–3	
B-STRAND	16–26	11	1st-CYS 23
BC-LOOP	27–31 34–38	10 (or less)	
C-STRAND	39–45	7	CONSERVED-TRP 41
CD-STRAND	45.1–45.9	0–9	
D-STRAND	77–84	8 (or 7 if gap at 82)	
DE-TURN	84.1–84.7 85.1–85.7	0–14	
E-STRAND	85–96	12	hydrophobic 89
EF-TURN	96.1–96.2	0–2	
F-STRAND	97–104	8	2nd-CYS 104
FG-LOOP	105–117	13 (or less, or more)	
G-STRAND	118–128	11 (or less)	

From Lefranc et al. (2005a).

[a]IMGT labels (concepts of description) are written in capital letters (**From IMGT-ONTOLOGY DESCRIPTION Axiom to IMGT Standardized Labels: For Immunoglobulin (IG) and T Cell Receptor (TR) Sequences and Structures** [Lefranc 2011c]).

[b]In number of amino acids (or codons).

[c]Residue@Position is a IMGT concept of numerotation that numbers the position of a given residue (or that of a conserved property amino acid class), based on the IMGT unique numbering.

positions 84.1 to 84.7 and from the apex 85.7 to 85.1, corresponding to 14 amino acids), the E-STRAND of 12 amino acids (positions 85 to 96) with a conserved hydrophobic amino acid at position 89, the EF-TURN (additional positions 96.1 and 96.2, corresponding to two amino acids), the F-STRAND of eight amino acids (positions 97 to 104) with the 2nd-CYS at position 104, the FG-LOOP (positions 105 to 117, these positions corresponding to an FG loop of 13 amino acids), and the G-STRAND of 11 (or less) amino acids (positions 118 to 128) (Table 2; Fig. 3; Lefranc et al. 2005a).

C Domain and V Domain Comparison

The A-STRAND and B-STRAND of the C domain are similar to those of the V domain (Lefranc et al. 2005a). The longest BC-LOOP of the C domain have 10 amino acids (missing positions 32 and 33), instead of 12 amino acids in the V domain. The C-STRAND and the D-STRAND of the C domain are shorter by one position (46) and two positions (75, 76), respectively, compared to those of the V domain. The transversal CD-STRAND is a characteristic of the C domain (a V domain has instead two antiparallel beta strands C′-STRAND and C″-STRAND linked by the C′C″-LOOP). The E-STRAND, F-STRAND, and G-STRAND of the C domain are similar to those of the V domain (Lefranc et al. 2005a) (IMGT Scientific chart rules, http://www.imgt.org).

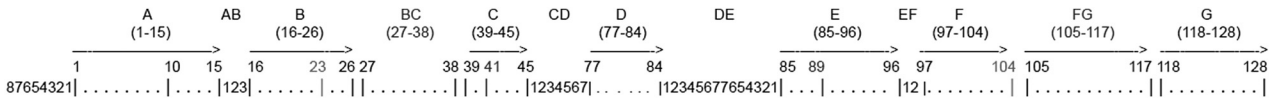


FIGURE 3. "IMGT Protein display for C domain" header. The header comprises five lines. Line 1: name of strands (A, B…), turns, and loops (AB, BC…). Line 2: start and end positions of strands, e.g. (1-15). Line 3: arrows (for the strands). Line 4: positions according to the IMGT unique numbering for C-DOMAIN and C-LIKE-DOMAIN (Lefranc et al. 2005b). Missing positions (32 and 33 in the BC loop and 46 to 76) in the C domain by comparison with the V domain are not shown. Line 5: Pipes for the start and end positions of strands and loops and highlighted positions, dots for the other positions, plus if needed, numbers for additional positions (for example: 123…) (Lefranc et al. 2005a).

IMGT Gaps and Additional Positions

Dots in IMGT Protein displays (Fig. 3) and hatched circles or squares in IMGT Colliers de Perles for C domain (**IMGT Collier de Perles for the Variable (V), Constant (C), and Groove (G) Domains of IG, TR, MH, IgSF, and MhSF** [Lefranc 2011e]) correspond to missing positions according to the IMGT unique numbering for C domain (Lefranc et al. 2005a).

The longest BC-LOOP of the C domain have 10 amino acids (missing positions 32 and 33, that are a feature of the C domain are not shown in the IMGT Colliers de Perles and IMGT Protein displays for C domain). For BC loops <10 amino acids, gaps are created from the apex in the following order: 34, 31, 35, 30, 36, etc. The FG-LOOP of the C domain is similar to that of the V domain. Gaps for FG loops <13 amino acids and additional positions for FG loops >13 amino acids are created following the same rules as those of the V domain.

Additional positions in the C domain define the AB-TURN, DE-TURN, and EF-TURN (Table 2). For AB-TURN shorter than three amino acids, gaps are created (hatched in IMGT Colliers de Perles, or not shown in structural data representations) in a decreasing ordinal manner. For DE-TURN <14 amino acids, gaps are created in the following order: 85.7, 84.7, 85.6, 84.6, 85.5, etc. For EF-TURN shorter than two amino acids, gaps are created in the following order: 96.2, 96.1 (Lefranc et al. 2005a).

IMGT UNIQUE NUMBERING FOR G DOMAIN

G Domain Strands, Turns, and Helix

The IMGT unique numbering for G domain numbers the G-DOMAIN of MH and the G-LIKE-DOMAIN of MhSF other than MH. The G domain strands, turns, and helix and their IMGT positions and lengths, based on the IMGT unique numbering for G domain (G-DOMAIN and G-LIKE-DOMAIN) (Lefranc et al. 2005b), are shown in Table 3.

The G domain (G-DOMAIN of MH and G-LIKE-DOMAIN of MhSF other than MH) comprises a sheet of four antiparallel beta strands linked by turns and a helix that sits on the beta strands, its axis forming an angle of ~40 degrees with the strands. Two G domains are needed to form the MhSF groove made of a "floor" and two "walls." Each G domain contributes by its strands and turns to half of the groove floor and by its helix to one wall of the groove (Lefranc et al. 2005b).

Table 3. G domain strands, turns, and helix, IMGT positions and lengths, based on the IMGT unique numbering for G domain (G-DOMAIN and G-LIKE-DOMAIN)

G domain strands, turns, and helix[a]	IMGT positions	Lengths[b]	Characteristic Residue@Position[c] and additional positions[d]
A-STRAND	1–14	14	7A, CYS-11
AB-TURN	15–17	3 (or 2 or 0)	
B-STRAND	18–28	11 (or 10[e])	
BC-TURN	29–30	2	
C-STRAND	31–38	8	
CD-TURN	39–41	3 (or 1[f])	
D-STRAND	42–49	8	49.1 to 49.5
HELIX	50–92	43 (or less or more)	54A, 61A, 61B, 72A, CYS-74, 92A

From Lefranc et al. (2005b).

[a]IMGT labels (concepts of description) are written in capital letters (**From IMGT-ONTOLOGY DESCRIPTION Axiom to IMGT Standardized Labels: For Immunoglobulin (IG) and T Cell Receptor (TR) Sequences and Structures** [Lefranc 2011c]).

[b]In number of amino acids (or codons).

[c]Residue@Position is a IMGT concept of numerotation that numbers the position of a given residue (or that of a conserved property amino acid class), based on the IMGT unique numbering.

[d]For details on the characteristic Residue@Position and additional positions, see text (Lefranc et al. 2005b).

[e]Or 9 in some G-BETA (Lefranc et al. 2005b).

[f]Or 0 in some G-ALPHA2-LIKE (Lefranc et al. 2005b).

The G domain is composed of the A-STRAND of 14 amino acids (positions 1 to 14), the AB-TURN of three or less amino acids (positions 15 to 17), the B-STRAND of 11 or less amino acids (positions 18 to 28), the BC-TURN of two amino acids (positions 29 and 30), the C-STRAND of eight amino acids (positions 31 to 38), the CD-TURN of three or less amino acids (positions 39 to 41), the D-STRAND of eight amino acids (positions 42 to 49), and a helix of 43 (or less or more) amino acids (positions 50 to 92) (Table 3; Fig. 4; Lefranc et al. 2005b).

The helix is split into two parts separated by a kink, positions 58 of G-ALPHA1, 61 of G-ALPHA2, 63 of G-ALPHA, and 62 of G-BETA being the "highest" points on the groove floor (Lefranc et al. 2005b).

IMGT Gaps and Additional Positions

Dots in IMGT Protein displays (Fig. 4) and hatched circles or squares in IMGT Collier de Perles for G domain (**IMGT Collier de Perles for the Variable (V), Constant (C), and Groove (G) Domains of IG, TR, MH, IgSF, and MhSF** [Lefranc 2011e]) correspond to missing positions according to the IMGT unique numbering for G domain (Lefranc et al. 2005b).

The gaps are localized in the turns. The AB-TURN (positions 15 to 17) comprises three amino acids in the G-ALPHA1, G-ALPHA2, and G-BETA domains but these positions are unoccupied in the G-ALPHA domains (as well as position 18 of B-STRAND). The BC-TURN (positions 29 and 30) comprises two positions that are occupied in all G-DOMAIN. The CD-TURN (positions 39 to 41) is occupied by three amino acids in the G-ALPHA1 domains and by one amino acid in most of the other G domains (Lefranc et al. 2005b).

The additional position 7A represents a bulge in 3D structures and is present in some G-ALPHA domains, for instance those of the human HLA-DQA1 and HLA-DOA, and mouse H2-AA and H2-DOA chains (Lefranc et al. 2005b). Additional positions at the N terminus of A-STRAND or at the C terminus of D-STRAND can be added if necessary. Thus, two additional positions (1.2 and 1.1) are added at the N terminus of the A-STRAND of the G-ALPHA domains as the presence of these two amino acids was demonstrated by protein sequencing of the HLA-DRA. However the proteolytic cleavage site of the leader peptide (L-REGION) needs to be confirmed experimentally for the other G-ALPHA (Lefranc et al. 2005b). In each [D1] G-DOMAIN (except the G-ALPHA domain of human HLA-DMA and mouse H2-DMA discussed below), the amino acid at position 1 (shown within parentheses in IMGT Protein displays, http://www.imgt.org) is encoded by the codon that results from the splicing between the first exon (EX1) that encodes the L-REGION and the second exon (EX2) that encodes [D1] (Lefranc et al. 2005b). Ten amino acids have been added at positions 1.10 to 1.1 of HLA-DMA and H2-DMA but it is necessary to confirm experimentally if they belong, or not, to the mature protein. Four additional positions (49.1 to 49.4) are also observed at the C terminus of D-STRAND of the G-BETA domain of HLA-DMB and H2-DMB1 (Lefranc et al. 2005b).

Interestingly, despite the high sequence divergence, only five additional positions in the helix (54A, 61A, 61B, 72A, and 92A) are necessary to align **any** G domain (Lefranc et al. 2005b). Three of them (61A, 61B, 72A) characterize the G-ALPHA2 and/or G-BETA domains. Indeed, positions 61A and 72A are occupied in the G-ALPHA2 domain, whereas positions 61A, 61B, and 72A are occupied in G-BETA domains (except for the mouse H2-AB chain). The position 92A is only occupied in the HLA-DMA and H2-DMA G-ALPHA domains. It is worthwhile to note that position 54A in G-ALPHA1-LIKE is the only additional position needed to extend the IMGT numbering for G-DOMAIN to the G-LIKE-DOMAIN of the MhSF proteins other than MH, or related proteins of the immune system (RPI)-MH1Like (Lefranc et al. 2005b).

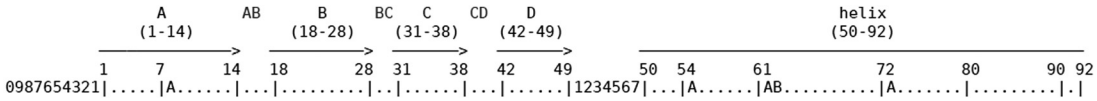


FIGURE 4. "IMGT Protein display for G domain" header. The header comprises five lines. Line 1: name of strands (A, B, C, D), turns (AB, BC, CD), and helix. Line 2: start and end positions of strands (1–14), (18–28), (31–38), (42–49), and helix (50–92). Line 3: arrows (for the strands) and line (for the helix). Line 4: positions according to the IMGT unique numbering for G-DOMAIN and G-LIKE-DOMAIN (Lefranc et al. 2005c). Line 5: Pipes for the start and end positions of strands and helix and highlighted positions, dots for the other positions, plus if needed, letters or numbers for additional positions that characterize some G domains (for example: A, AB or 123…) (Lefranc et al. 2005b).

G Domain Characteristics Comparison

Two cysteine, CYS-11 (in strand A) and CYS-74 (in the helix), are well conserved in the G-ALPHA2 and G-BETA domains where they participate in a disulfide bridge that fastens the helix on the groove floor. The G-ALPHA1 and G-ALPHA domains have a conserved N-glycosylation site at position 86 (N-X-S/T, where N is asparagine, X is any amino acid except proline, S is serine, and T is threonine). An N-glycosylation site is also found at position 86 in the G-ALPHA2 domain of the mouse classical MH1 chains (H2-D1, H2-K1, and H2-L). The G-BETA domains (except for the human HLA-DMB and mouse H2-DMB1 chains) have a conserved N-glycosylation site at position 15 (AB-TURN) (Lefranc et al. 2005b).

Interestingly, the G-ALPHA domains of the HLA-DMA and H2-DMA chains have specific features compared to the other G-ALPHA domains and share common characteristics with the G-ALPHA2 and G-BETA domains: there is a conserved CYS-11_CYS-74 disulfide bridge, positions 61A and 61B are occupied (as in the G-BETA domains), and there is no N-glycosylation site at position 86 (Lefranc et al. 2005b) (IMGT Protein display in IMGT Repertoire, http://www.imgt.org).

Practically, the IMGT unique numbering for positions 1–39 and 73–92 of the G-ALPHA2 domains can be obtained very easily by substracting 90 from the mature protein numbering (91–129 and 163–182). Between these positions, the two gaps (at positions 40 and 41) and the two insertions (at positions 61A and 72A) are necessary, in the IMGT unique numbering, to allow meaningful sequence and structure alignment and comparison between the G-ALPHA1 and G-ALPHA2 sequences.

BRIDGING THE GAP BETWEEN IMMUNOGENETICS AND SYSTEMS BIOLOGY

By allowing one to number the V and C domains of the IG, TR, and IgSF other than IG and TR, and the G domain of the MH and MhSF other than MH (Lefranc 1997, 1999; Lefranc et al. 2003, 2005a,b), the IMGT unique numbering has been fundamental in:

- the creation of the IMGT Colliers de Perles for V, C, and G domains (**IMGT Collier de Perles for the Variable (V), Constant (C), and Groove (G) Domains of IG, TR, MH, IgSF, and MhSF** [Lefranc 2011e])
- the definition of the Residue@Position concept (standardized numerotation of amino acid changes and polymorphisms)
- the standardized numerotation of mutations (at the codon and nucleotide level) and definition of alleles
- the delimitation of the FR-IMGT and CDR-IMGT for antibody engineering and antibody humanization
- the comparison of interactions and contact analysis between domains in 3D structures

The IMGT unique numbering bridges the gap between amino acid (and nucleotide) sequences and 3D structures and is essential for domain comparison in molecular immunogenetics (Bertrand et al. 2004; Duprat et al. 2004; Frigoul and Lefranc 2005; Garapati and Lefranc 2007), in structural biology and immunology for receptor/ligand interaction analysis (Kaas and Lefranc 2005; Kaas et al. 2008), and in biotechnology related to antibody engineering and humanization (Lefranc 2009; Ehrenmann et al. 2010a). The IMGT unique numbering is used (i) for the annotation of sequences, genes, and structures in the IMGT databases: IMGT/LIGM-DB (Giudicelli et al. 2006), IMGT/CLL-DB (Brochet 2008), IMGT/GENE-DB (Giudicelli et al. 2005), IMGT/2Dstructure-DB and IMGT/3Dstructure-DB (Kaas et al. 2004; Ehrenmann et al. 2010b), (ii) for the analysis of sequences and structures in the IMGT online tools: IMGT/V-QUEST (Giudicelli et al. 2004; Brochet et al. 2008; Giudicelli and Lefranc 2009), IMGT/JunctionAnalysis (Yousfi Monod et al. 2004; Bleakley et al. 2006), IMGT/HighV-QUEST (Alamyar et al. 2010), IMGT/Phylogene (Elemento and Lefranc 2003), IMGT/GeneInfo (Baum et al. 2004, 2006), IMGT/LIGMotif (Lane et al. 2010), IMGT/DomainGapAlign, IMGT/DomainDisplay and IMGT/Collier de Perles (Ehrenmann et al. 2010b), and (iii) in the data management of the IMGT knowledge web resources: IMGT Protein displays, IMGT Alignments of alleles, IMGT Table of alleles, IMGT Colliers de Perles, etc. (http://www.imgt.org) (**IMGT, the International ImMunoGeneTics Information System** [Lefranc 2011a]).

As a major IMGT-ONTOLOGY concept of numerotation, the IMGT unique numbering is used, with the IMGT-ONTOLOGY concepts of classification (**From IMGT-ONTOLOGY CLASSIFICATION Axiom to IMGT Standardized Gene and Allele Nomenclature: For Immunoglobulins (IG) and T Cell Receptors (TR)** [Lefranc 2011d]) and concepts of description (**From IMGT-ONTOLOGY DESCRIPTION Axiom to**

IMGT Standardized Labels: For Immunoglobulin (IG) and T Cell Receptor (TR) Sequences and Structures [Lefranc 2011c]), in the definition of monoclonal antibodies (mAb, suffix -mab) and fusion proteins for immune applications (FPIA, suffix -cept) of the World Health Organization/International Nonproprietary Name (WHO/INN) programme (Lefranc 2011f).

REFERENCES

Alamyar E, Giudicelli V, Duroux P, Lefranc M-P. 2010. IMGT/HighV-QUEST: A high-throughput system and Web portal for the analysis of rearranged nucleotide sequences of antigen receptors - High-throughput version of IMGT/V-QUEST. JOBIM 2010 Poster 60. http://www.jobim2010.fr/?q=fr/node/55.

Baum TP, Pasqual N, Thuderoz F, Hierle V, Chaume D, Lefranc M-P, Jouvin-Marche E, Marche N, Demongeot J. 2004. IMGT/GeneInfo: Enhancing V(D)J recombination database accessibility. *Nucl Acids Res* **32:** D51–D54.

Baum TP, Hierle V, Pascal N, Bellahcene F, Chaume D, Lefranc M-P, Jouvin-Marche E, Marche PN, Demongeot J. 2006. IMGT/GeneInfo: T cell receptor gamma TRG and delta TRD genes in database give access to all TR potential V(D)J recombinations. *BMC Bioinformatics* **7:** 224. doi: 10.1186/1471-2105-7-224.

Bertrand G, Duprat E, Lefranc M-P, Marti J, Coste J. 2004. Characterization of human fcgr3b*02 (hna-1b, na2) cDNAs and imgt standardized description of fcgr3b alleles. *Tissue Antigens* **64:** 119–131.

Bleakley K, Giudicelli V, Wu Y, Lefranc M-P, Biau G. 2006. IMGT standardization for statistical analyses of T cell receptor junctions: The TRAV-TRAJ example. *In Silico Biol* **6:** 573–588.

Brochet X. 2008. "Conception et intégration d'un système d'information dédié à l'analyse et à la gestion des séquences réarrangées des récepteurs d'antigènes au sein d'IMGT: Application à la Leucémie Lymphoïde Chronique." PhD Thesis, Université Montpellier 1, Montpellier.

Brochet X, Lefranc M-P, Giudicelli V. 2008. IMGT/V-QUEST: The highly customized and integrated system for IG and TR standardized V-J and V-D-J sequence analysis. *Nucl Acids Res* **36:** W503–W508.

Duprat E, Kaas Q, Garelle V, Lefranc G, Lefranc M-P. 2004. IMGT standardization for alleles and mutations of the V-LIKE-DOMAINs and C-LIKE-DOMAINs of the immunoglobulin superfamily. In *Recent research developments in human genetics* (ed. SG Pandalai), Vol 2, pp. 111–136. Research Signpost, Trivandrum, Kerala, India.

Duroux P, Kaas Q, Brochet X, Lane J, Ginestoux C, Lefranc M-P, Giudicelli V. 2008. IMGT-Kaleidoscope, the Formal IMGT-ONTOLOGY paradigm. *Biochimie* **90:** 570–583.

Ehrenmann F, Lefranc M-P. 2011a. IMGT/DomainGapAlign: IMGT standardized analysis of amino acid sequences of variable, constant, and groove domains (IG, TR, MH, IgSF, MhSF). *Cold Spring Harb Protoc* doi: 10.1101/pdb.prot5636.

Ehrenmann F, Lefranc M-P. 2011b. IMGT/3Dstructure-DB: Querying the IMGT database for 3D structures in immunology and immunoinformatics (IG or antibodies, TR, MH, RPI, and FPIA). *Cold Spring Harb Protoc* doi: 10.1101/pdb.prot5637.

Ehrenmann F, Duroux P, Giudicelli V, Lefranc M-P. 2010a. Standardized sequence and structure analysis of antibody using IMGT. In *Antibody engineering*, 2nd ed. (ed. R Kontermann, S Dübel), Vol 2, pp. 11–31. Springer-Verlag, Berlin/Heidelberg.

Ehrenmann F, Kaas Q, Lefranc M-P. 2010b. IMGT/3Dstructure-DB and IMGT/DomainGapAlign: A database and a tool for immunoglobulins or antibodies, T cell receptors, MHC, IgSF and MhcSF. *Nucl Acids Res* **38:** D301–D307.

Ehrenmann F, Giudicelli V, Brochet X, Lefranc M-P. 2011. IMGT/Collier de Perles: IMGT standardized representation of domains (IG, TR, and IgSF variable and constant domains, MH and MhSF groove domains). *Cold Spring Harb Protoc* doi: 10.1101/pdb.prot5635.

Elemento O, Lefranc M-P. 2003. IMGT/PhyloGene, an online software package for phylogenetic analysis of immunoglobulin and T cell receptor genes. *Dev Comp Immunol* **27:** 763–779.

Folch G, Scaviner D, Contet V, Lefranc M-P. 2000. Protein displays of the human T cell Receptor alpha, beta, gamma and delta variable and joining regions. *Exp Clin Immunogenet* **17:** 205–215.

Frigoul A, Lefranc M-P. 2005. MICA: Standardized IMGT allele nomenclature, polymorphisms and diseases. In *Recent research developments in human genetics* (ed. SG Pandalai), Vol 3, pp. 95–145. Research Signpost, Trivandrum, Kerala, India.

Garapati VP, Lefranc M-P. 2007. IMGT Colliers de Perles and IgSF domain standardization for T cell costimulatory activatory (CD28, ICOS) and inhibitory (CTLA4, PDCD1 and BTLA) receptors. *Dev Comp Immunol* **31:** 1050–1072.

Giudicelli V, Lefranc M-P. 1999. Ontology for immunogenetics: The IMGT-ONTOLOGY. *Bioinformatics* **12:** 1047–1054.

Giudicelli V, Lefranc M-P. 2009. IMGT standardized analysis of immunoglobulin rearranged sequences. In *Immunoglobulin gene analysis in chronic lymphocytic leukemia* (ed. P Ghia et al.), pp. 33–52. Wolters Kluwer Health Italy Ltd, Milan.

Giudicelli V, Lefranc M-P. 2011. IMGT/JunctionAnalysis: IMGT standardized analysis of the V-J and V-D-J junctions of the rearranged immunoglobulins (IG) and T cell receptors (TR). *Cold Spring Harb Protoc* doi: 10.1101/pdb.prot5634.

Giudicelli V, Chaume D, Lefranc M-P. 2004. IMGT/V-QUEST, an integrated software for immunoglobulin and T cell receptor V-J and V-D-J rearrangement analysis. *Nucl Acids Res* **32:** W435–W440.

Giudicelli V, Chaume D, Lefranc M-P. 2005. IMGT/GENE-DB: A comprehensive database for human and mouse immunoglobulin and T cell receptor genes. *Nucl Acids Res* **33:** D256–D261.

Giudicelli V, Duroux P, Ginestoux C, Folch G, Jabado-Michaloud J, Chaume D, Lefranc M-P. 2006. IMGT/LIGM-DB, the IMGT comprehensive database of immunoglobulin and T cell receptor nucleotide sequences. *Nucl Acids Res* **34:** D781–D784.

Giudicelli V, Brochet X, Lefranc M-P. 2011. IMGT/V-QUEST: IMGT standardized analysis of the immunoglobulin (IG) and T cell receptor (TR) nucleotide sequences. *Cold Spring Harb Protoc* doi: 10.1101/pdb.prot5633.

Kaas Q, Lefranc M-P. 2005. T cell receptor/peptide/MHC molecular characterization and standardized pMHC contact sites in IMGT/3Dstructure-DB. *In Silico Biol* **5:** 505–528.

Kaas Q, Ruiz M, Lefranc M-P. 2004. IMGT/3Dstructure-DB and IMGT/StructuralQuery, a database and a tool for immunoglobulin, T cell receptor and MHC structural data. *Nucl Acids Res* **32:** D208–D210.

Kaas Q, Duprat E, Tourneur G, Lefranc M-P. 2008. IMGT standardization for molecular characterization of the T cell receptor/peptide/MHC complexes. In *Immunoinformatics* (ed. C Schoenbach et al.), pp. 19–49. Immunomics Reviews, Series of Springer Science and Business Media LLC, Springer, New York.

Lane L, Duroux P, Lefranc M-P. 2010. From IMGT-ONTOLOGY to IMGT/LIGMotif: The IMGT standardized approach for immunoglobulin and T cell receptor gene identification and description in large genomic sequences. *BMC Bioinformatics* **11:** 223. doi: 10.1186/1471-2105-11-223.

Lefranc M-P. 1997. Unique database numbering system for immunogenetic analysis. *Immunol Today* **18:** 509.

Lefranc M-P. 1999. The IMGT unique numbering for immunoglobulins, T cell receptors and Ig-like domains. *The Immunologist* **7:** 132–136.

Lefranc M-P. 2009. Antibody database and tools: The IMGT experience. In *Therapeutic monoclonal antibodies: From bench to clinic* (ed. A Zhiqiang), pp. 91–114. John Wiley Sons Inc, Hoboken, NJ.

Lefranc M-P. 2011a. IMGT, the international ImMunoGeneTics information system. *Cold Spring Harb Protoc* doi: 10.1101/pdb.top115.
Lefranc M-P. 2011b. From IMGT-ONTOLOGY IDENTIFICATION axiom to IMGT standardized keywords: For immunoglobulins (IG), T cell receptors (TR), and conventional genes. *Cold Spring Harb Protoc* doi: 10.1101/pdb.ip82.
Lefranc M-P. 2011c. From IMGT-ONTOLOGY DESCRIPTION axiom to IMGT standardized labels: For immunoglobulin (IG) and T cell receptor (TR) sequences and structures. *Cold Spring Harb Protoc* doi: 10.1101/pdb.ip83.
Lefranc M-P. 2011d. From IMGT-ONTOLOGY CLASSIFICATION axiom to IMGT standardized gene and allele nomenclature: For immunoglobulins (IG) and T cell receptors (TR). *Cold Spring Harb Protoc* doi: 10.1101/pdb.ip84.
Lefranc M-P. 2011e. IMGT Collier de Perles for the Variable (V), Constant (C), and Groove (G) Domains of IG, TR, MH, IgSF, and MhSF. *Cold Spring Harb Protoc* doi: 10.1101/pdb.ip86.
Lefranc M-P. 2011f. Antibody nomenclature: From IMGT-ONTOLOGY to INN definition. *MAbs* **3:** 1–2.
Lefranc M-P, Lefranc G. 2001a. *The immunoglobulin FactsBook*. Academic Press, London.
Lefranc M-P, Lefranc G. 2001b. *The T cell receptor FactsBook*. Academic Press, London.
Lefranc M-P, Pommié C, Ruiz M, Giudicelli V, Foulquier E, Truong L, Thouvenin-Contet V, Lefranc G. 2003. IMGT unique numbering for immunoglobulin and T cell receptor variable domains and Ig superfamily V-like domains. *Dev Comp Immunol* **27:** 55–77.
Lefranc M-P, Giudicelli V, Ginestoux C, Bosc N, Folch G, Guiraudou D, Jabado-Michaloud J, Magris S, Scaviner D, Thouvenin V, et al. 2004. IMGT-ONTOLOGY for immunogenetics and immunoinformatics. *In Silico Biol* **4:** 17–29.
Lefranc M-P, Clément O, Kaas Q, Duprat E, Chastellan P, Coelho I, Combres K, Ginestoux C, Giudicelli V, Chaume D, et al. 2005a. IMGT-Choreography for immunogenetics and immunoinformatics. *In Silico Biol* **5:** 45–60.
Lefranc M-P, Duprat E, Kaas Q, Tranne M, Thiriot A, Lefranc G. 2005b. IMGT unique numbering for MHC groove G-DOMAIN and MHC superfamily (MhcSF) G-LIKE-DOMAIN. *Dev Comp Immunol* **29:** 917–938.
Lefranc M-P, Pommié C, Kaas Q, Duprat E, Bosc N, Guiraudou D, Jean C, Ruiz M, Da Piedade I, Rouard M, et al. 2005c. IMGT unique numbering for immunoglobulin and T cell receptor constant domains and Ig superfamily C-like domains. *Dev Comp Immunol* **29:** 185–203.
Lefranc M-P, Giudicelli V, Regnier L, Duroux P. 2008. IMGT, a system and an ontology that bridge biological and computational spheres in bioinformatics. *Brief Bioinform* **9:** 263–275.
Lefranc M-P, Giudicelli V, Ginestoux C, Jabado-Michaloud J, Folch G, Bellahcene F, Wu Y, Gemrot E, Brochet X, Lane J, et al. 2009. IMGT, the international ImMunoGeneTics information system. *Nucl Acids Res* **37:** D1006–D1012.
Scaviner D, Barbié V, Ruiz M, Lefranc M-P. 1999. Protein displays of the human immunoglobulin heavy, kappa and lambda variable and joining regions. *Exp Clin Immunogenet* **16:** 234–240.
Yousfi Monod M, Giudicelli V, Chaume D, Lefranc M-P. 2004. IMGT/JunctionAnalysis: The first tool for the analysis of the immunoglobulin and T cell receptor complex V-J and V-D-J JUNCTIONs. *Bioinformatics* **20:** i379–i385.

Information Panel

IMGT Collier de Perles for the Variable (V), Constant (C), and Groove (G) Domains of IG, TR, MH, IgSF, and MhSF

Marie-Paule Lefranc

IMGT, the international ImMunoGeneTics information system, Laboratoire d'ImmunoGénétique Moléculaire LIGM, Université Montpellier 2, Institut de Génétique Humaine IGH, UPR CNRS 1142, 34396 Montpellier cedex 5, France

INTRODUCTION

The "IMGT Collier de Perles" (or "IMGT_Collier_de_Perles") concept is a major concept of numerotation (generated from the NUMEROTATION axiom) of IMGT-ONTOLOGY, the global reference in immunogenetics and immunoinformatics, built by IMGT, the international ImMunoGeneTics information system. The "IMGT Collier de Perles" concept, described here, allows standardized two-dimensional (2D) graphical representations of the domains, based on the IMGT unique numbering. Three leafconcepts (a leafconcept is a concept that corresponds to the finest level of granularity) have been defined: for the variable (V) domain and constant (C) domain of the immunoglobulin superfamily (IgSF) and for the groove (G) domain of the major histocompatibility (MH) superfamily (MhSF). IMGT Colliers de Perles are obtained, starting from V, C, or G domain amino acid sequences, using IMGT/DomainGapAlign and IMGT/Collier de Perles tools. In IMGT/3Dstructure-DB, IMGT Colliers de Perles of V and C domains are provided with hydrogen bonds and those of G domains with IMGT pMH contact analysis. IMGT Colliers de Perles allows one to bridge the gap between sequences and three-dimensional (3D) structures, whatever the species, the IgSF or MhSF protein, or the chain type. They are particularly useful for antibody engineering, sequence-structure analysis, visualization and comparison of positions for mutations, polymorphisms and contact analysis of immunoglobulins (IG), T cell receptors (TR), MH, and related proteins of the immune system (RPI) belonging to the IgSF and MhSF.

RELATED INFORMATION

A detailed description of IMGT is provided in **IMGT, the International ImMunoGeneTics Information System** (Lefranc 2011a). The NUMEROTATION axiom is an axiom of the Formal IMGT-ONTOLOGY or IMGT-Kaleidoscope (Giudicelli and Lefranc 1999; Lefranc et al. 2004, 2005a, 2008; Duroux et al. 2008). The NUMEROTATION axiom postulates that any molecule, cell, tissue, organ, organism or population, any process and any relation, has to be numbered. The NUMEROTATION axiom has generated the concepts of numerotation of IMGT-ONTOLOGY, the global reference in immunogenetics and immunoinformatics, built by IMGT (http://www.imgt.org) (Lefranc et al. 2009). The NUMEROTATION axiom is one of seven axioms of the Formal IMGT-ONTOLOGY, the others being "IDENTIFICATION axiom" **(From IMGT-ONTOLOGY IDENTIFICATION Axiom to IMGT Standardized Keywords: For Immunoglobulins (IG), T Cell Receptors (TR), and Conventional Genes**) (Lefranc 2011b), "DESCRIPTION axiom" **(From IMGT-ONTOLOGY DESCRIPTION Axiom to IMGT Standardized Labels: For Immunoglobulin (IG) and T Cell Receptor (TR) Sequences and Structures)** (Lefranc 2011c), "CLASSIFICATION axiom" **(From IMGT-ONTOLOGY CLASSIFICATION Axiom to IMGT Standardized Gene and Allele Nomenclature: For Immunoglobulins (IG) and T Cell Receptors (TR))** (Lefranc 2011d), "LOCALIZATION axiom", ORIENTATION axiom", and "OBTENTION axiom". The IMGT Collier de Perles (Lefranc 1999; Ruiz and Lefranc 2002; Kaas and Lefranc 2007; Kaas et al. 2007) and the IMGT unique numbering (Lefranc 1997, 1999; Lefranc et al. 2003, 2005b,c) **(IMGT Unique Numbering for the Variable (V), Constant (C), and Groove (G) Domains of IG, TR, MH, IgSF, and MhSF)** (Lefranc 2011e) are the two major concepts of numerotation, generated from the NUMEROTATION axiom.

Protocols are available for **IMGT/V-QUEST: IMGT Standardized Analysis of the Immunoglobulin (IG) and T Cell Receptor (TR) Nucleotide Sequences** (Giudicelli et al. 2011), **IMGT/JunctionAnalysis: IMGT**

Corresponding author (Marie-Paule.Lefranc@igh.cnrs.fr).
Cite as: Cold Spring Harb Protoc; 2011; doi:10.1101/pdb.ip86

Standardized Analysis of the V-J and V-D-J Junctions of the Rearranged Immunoglobulins (IG) and T Cell Receptors (TR) (Giudicelli and Lefranc 2011), **IMGT/Collier de Perles: IMGT Standardized Representation of Domains (IG, TR, and IgSF Variable and Constant Domains, MH and MhSF Groove Domains)** (Ehrenmann et al. 2011), **IMGT/DomainGapAlign: IMGT Standardized Analysis of Amino Acid Sequences of Variable, Constant, and Groove Domains (IG, TR, MH, IgSF, MhSF)** (Ehrenmann and Lefranc 2011a), and **IMGT/3Dstructure-DB: Querying the IMGT Database for 3D Structures in Immunology and Immunoinformatics (IG or Antibodies, TR, MH, RPI, and FPIA)** (Ehrenmann and Lefranc 2011b).

BACKGROUND INFORMATION

A standardized approach for immunogenetics and immunoinformatics was strongly needed in the 1990s. The first IMGT Collier de Perles to be put online was the VH domain from mouse E5.2 Fv in December 1997, still available in IMGT Repertoire (http://www.imgt.org) in its original drawing. By creating the IMGT Colliers de Perles (Lefranc 1999; Ruiz and Lefranc 2002; Kaas and Lefranc 2007; Kaas et al. 2007), based on the IMGT unique numbering (Lefranc 1997, 1999; Lefranc et al. 2003, 2005b,c), IMGT has bridged the gap between sequences and 3D structures (Lefranc 2005, 2008a,b; Lefranc et al. 2008). IMGT Collier de Perles can be obtained by the users from their own nucleotide sequences using IMGT/V-QUEST (Giudicelli et al. 2004; Brochet 2008; Brochet et al. 2008; Giudicelli and Lefranc 2009), from their amino acid sequences using IMGT/DomainGapAlign (Ehrenmann et al. 2010b), or whether their amino acid sequences are already gapped according to the IMGT unique numbering, using the IMGT/Collier de Perles tool (http://www.imgt.org). IMGT Colliers de Perles are available in IMGT/3Dstructure-DB, with hydrogen bonds for V and C domains, for proteins with known 3D structures, and in IMGT/2Dstructure-DB for therapeutic monoclonal antibodies and fusion proteins for immune applications (FPIA) (Edelmann et al. 2010b; Poiron et al. 2010). IMGT Colliers de Perles are widely used for structural biology and immunology (Kaas and Lefranc 2007; Kaas et al. 2007) and biotechnologies related to antibody engineering and humanization (Lefranc 2009; Ehrenmann et al. 2010a).

IMGT COLLIER DE PERLES FOR V DOMAIN

"IMGT Collier de Perles for V domain" (or "IMGT_Collier_de_Perles_for_V_domain") is a leafconcept of the "IMGT Collier de Perles" concept of numerotation. It represents the V domain that includes the V-DOMAIN of IG or antibodies and TR and the V-LIKE-DOMAIN of the IgSF proteins other than IG and TR. It is based on the IMGT unique numbering for V domain (Lefranc 1997, 1999; Lefranc et al. 2003) **(IMGT Unique Numbering for the Variable (V), Constant (C), and Groove (G) Domains of IG, TR, MH, IgSF, and MhSF)** (Lefranc 2011e).

The topology and 3D structure of a V-LIKE-DOMAIN are very similar to those of an IG or TR V-DOMAIN (Fig. 1A).

A V domain consists of about 100 amino acids in nine antiparallel β strands (A, B, C, C′, C″, D, E, F, and G), linked by β turns (AB, CC′, C″ D, DE, and EF) and loops (BC, C′C″, and FG), located on two layers, forming a sandwich of two sheets. The sheets are closely packed against each other through hydrophobic interactions, giving a hydrophobic core and joined together by a disulfide bridge between the B-STRAND in the first sheet and the F-STRAND in the second sheet (Lefranc et al. 2003). Four strands form one sheet and five strands form a second sheet.

It is remarkable that the 3D structure fold of the V domain has been conserved through evolution, despite the sequence divergence between IgSF domains and, even more surprisingly, despite the particularities of IG and TR synthesis compared with the other proteins (Lefranc and Lefranc 2001a,b). Indeed, in a conventional gene, the V-LIKE-DOMAIN is usually encoded by a unique exon, or more rarely by two spliced exons, whereas in IG and TR, the V-DOMAIN results from the rearrangement of two or three types of genes: variable (V) gene, joining (J) gene and, in some loci, diversity (D) gene (Lefranc and Lefranc 2001a,b) (Fig. 2).

The V-LIKE-DOMAIN is usually, as is the IG and TR V-DOMAIN, the most N-terminal (and extracellular) domain of the protein. However, in contrast to the IG and TR V-DOMAIN, which is always unique, the V-LIKE-DOMAIN may be present in several copies in the same protein and interspersed with the C-LIKE-DOMAIN or with domains of other superfamilies.

Based on the IMGT unique numbering, the conserved amino acids always have the same position. For the V domain, five positions are essential for the core structure: cysteine **23** (1st-CYS), tryptophan **41**

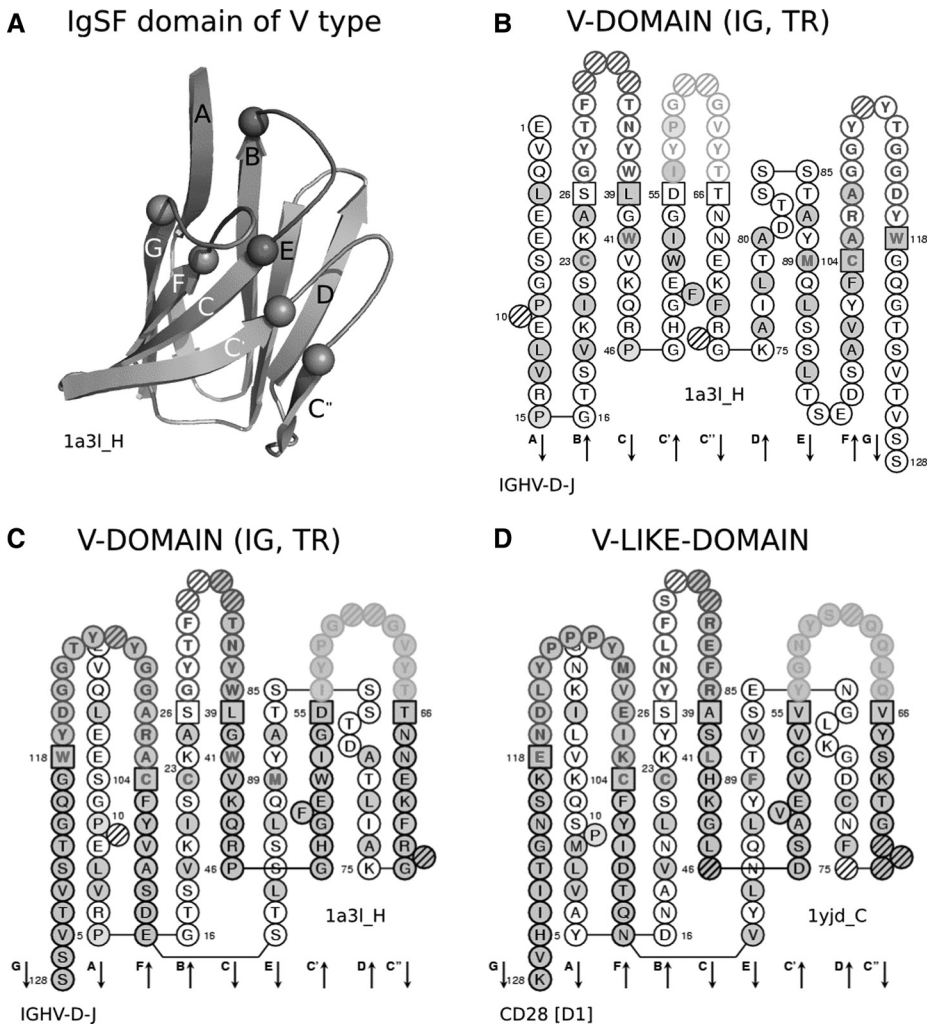


FIGURE 1. IMGT Collier de Perles for V domain. (*A*) Ribbon representation of a V-DOMAIN as an example. A similar topology and 3D structure characterize a V-LIKE-DOMAIN. (*B*,*C*) V-DOMAIN on one layer and on two layers, respectively (*Mus musculus* VH [8.8.12]). (*D*) V-LIKE-DOMAIN on two layers (*Homo sapiens* CD28 [9.9.13]). Amino acids are shown in the one-letter abbreviation. Position at which hydrophobic amino acids (hydropathy index with positive value: I, V, L, F, C, M, A) and tryptophan (W) are found in >50% of analyzed sequences are shown online in blue. All proline (P) are shown online in yellow. The loops BC, C′C″and FG (corresponding to the CDR-IMGT) are limited by amino acids shown in squares (anchor positions), which belong to the neighboring strands (FR-IMGT). BC loops are represented online in red, C′C″ loops in orange, and FG loops in purple. Hatched circles or squares correspond to missing positions according to the IMGT unique numbering for V domain (Lefranc et al. 2003). Arrows indicate the direction of the β strands and their designations in 3D structures. The IMGT Colliers de Perles on two layers show, in the forefront, the GFCC′C″ strands and, in the back, the ABED strands. IMGT Colliers de Perles on two layers with hydrogen bonds are available in IMGT/3Dstructure-DB (http://www.imgt.org): 1a3l_H, 1yjd_C.

(CONSERVED-TRP), conserved hydrophobic amino acid **89**, cysteine **104** (2nd-CYS), and "FG anchor" **118** (J-PHE or J-TRP in V-DOMAIN) (Lefranc et al. 2003). In the IG and TR V-DOMAIN, the structurally conserved antiparallel β strands are designated as framework regions (FR-IMGT), whereas the BC, C′C″, and FG loops are designated as complementarity determining regions (CDR-IMGT) (Lefranc et al. 2003). The length (number of amino acids or by extrapolation number of codons, that is the number of occupied positions) of the strands and loops is a crucial and original concept of IMGT-ONTOLOGY. The lengths of the BC (CDR1-IMGT), C′C″ (CDR2-IMGT) and FG (CDR3-IMGT) loops characterize the V domain. Thus, the length of the three loops BC, C′C″, and FG is shown, in number of amino acids (or codons), in brackets and separated by dots. In Figure 1, the lengths are [8.8.12] for the V-DOMAIN CDR-IMGT and [9.9.13] for the V-LIKE-DOMAIN. IMGT Colliers de Perles

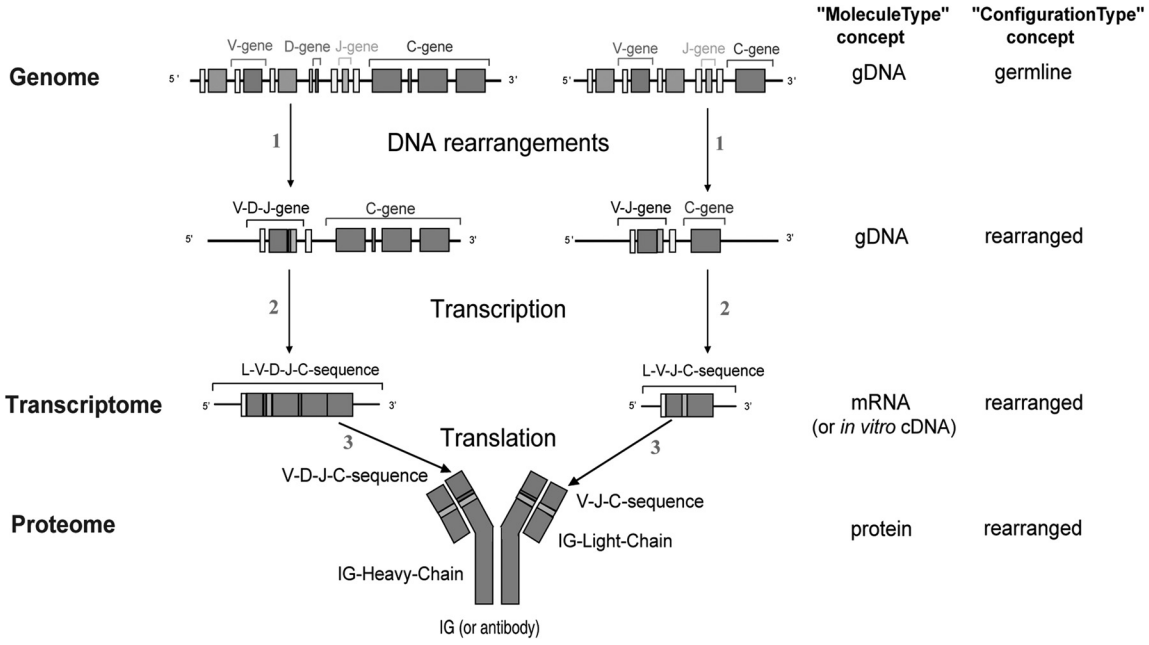


FIGURE 2. Synthesis of an IG or antibody in humans. A human being may potentially synthesize 10^{12} different antibodies (Lefranc and Lefranc 2001a). (1) DNA rearrangements (is_rearranged_into), (2) Transcription (is_transcribed_into), (3) Translation (is_translated_into). The configuration of C-GENE is "undefined" (ConfigurationType) (IMGT Repertoire, http://www/imgt/org).

for V domain can be represented on one layer (Fig. 1B) or two layers (Fig. 1C,D) (Lefranc 1999; Ruiz and Lefranc 2002; Lefranc et al. 2003; Kaas and Lefranc 2007; Kaas et al. 2007). If 3D structures are available, IMGT Colliers de Perles on two layers are provided with hydrogen bonds, in IMGT/3Dstructure-DB (http://www.imgt.org) (Ehrenmann et al. 2010b).

IMGT COLLIER DE PERLES FOR C DOMAIN

"IMGT Collier de Perles for C domain" (or "IMGT_Collier_de_Perles_for_C_domain") is a leafconcept of the "IMGT Collier de Perles" concept of numerotation. It represents the C domain, which includes the C-DOMAIN of IG or antibodies and TR and the C-LIKE-DOMAIN of IgSF proteins other than IG and TR. It is based on the IMGT unique numbering for C domain (Lefranc et al. 2005c) **(IMGT Unique Numbering for the Variable (V), Constant (C), and Groove (G) Domains of IG, TR, MH, IgSF, and MhSF)** (Lefranc 2011e). The topology and 3D structure of a C-LIKE-DOMAIN are very similar to those of an IG or TR C-DOMAIN (Fig. 3A).

A C domain consists of about 100 amino acids in seven antiparallel β strands (A, B, C, D, E, F, and G), linked by β turns (AB, DE, and EF), a transverse strand (CD), and loops (BC and FG), located in two layers, forming a sandwich of two sheets. The sheets are closely packed against each other through hydrophobic interactions, giving a hydrophobic core, and joined together by a disulfide bridge between the B-STRAND in the first sheet and the F-STRAND in the second sheet (Lefranc et al. 2005c). Four strands form one sheet and three strands form a second sheet.

The C domain has a topology and a 3D structure similar to those of the V domain, but differs by the number of antiparallel β strands (seven instead of nine). By comparison with a V domain, the C′-STRAND, C′C″-LOOP and C″-STRAND are missing in the C domain and are replaced by the characteristic transverse CD-STRAND that links the two sheets. Depending on the CD-STRAND length, the D-STRAND is in the first or in the second sheet (Lefranc et al. 2005c). Additional positions observed in the C domain define the AB-TURN, DE-TURN, and EF-TURN (Lefranc et al. 2005c).

The IMGT unique numbering for C domain (Lefranc et al. 2005c) is derived from the IMGT unique numbering first described for the V-REGION and for the V domain (Lefranc et al. 2003). The five positions that are essential for the core structure have the same number in the C domain: cysteine **23** (1st-CYS),

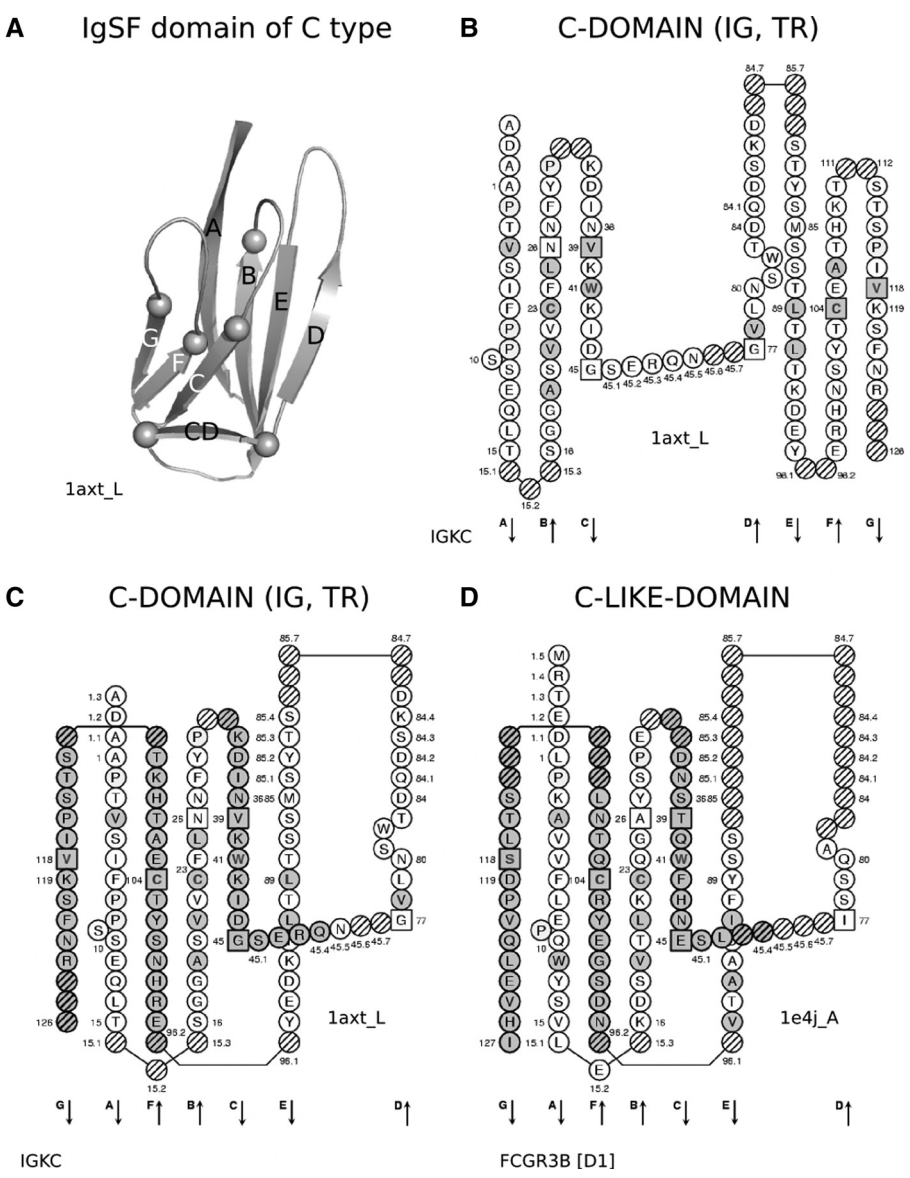


FIGURE 3. IMGT Collier de Perles for C domain. (*A*) Ribbon representation of a C-DOMAIN as an example. A similar topology and 3D structure characterize a C-LIKE-DOMAIN. (*B*,*C*) C-DOMAIN on one layer and two layers, respectively (*Mus musculus* C-KAPPA). (*D*) C-LIKE-DOMAIN on two layers (*Homo sapiens* FCGR3B [D1]). Amino acids are shown in the one-letter abbreviation. Position at which hydrophobic amino acids (hydropathy index with positive value: I, V, L, F, C, M, A) and tryptophan (W) are found in >50% of analyzed sequences are shown in gray. The positions 26, 39, and 104 are shown in squares by homology with the corresponding positions in the V domain. Positions 45 and 77, which delimit the characteristic CD-STRAND of the C domain, and position 118, which corresponds structurally to J-PHE or J-TRP of the IG and TR J REGION (Lefranc and Lefranc 2001a,b), are also shown in squares. Hatched circles correspond to missing positions according to the IMGT unique numbering for C domain (Lefranc et al. 2005c). Arrows indicate the direction of the β strands and their designations in 3D structures. The IMGT Colliers de Perles on two layers show, in the *forefront*, the GFC strands and, in the *back*, the ABE strands. IMGT Colliers de Perles on two layers with hydrogen bonds are available in IMGT/3Dstructure-DB (http://www.imgt.org): 1axt_L, 1e4j_A.

tryptophan **41** (CONSERVED-TRP), conserved hydrophobic amino acid **89**, cysteine **104** (2nd-CYS), and "FG anchor" **118** (Lefranc et al. 2005c).

IMGT Colliers de Perles for C domain can be represented on one layer (Fig. 3B) or two layers (Fig. 3C, D) (Lefranc et al. 2005c; Kaas and Lefranc 2007; Kaas et al. 2007). If 3D structures are available, IMGT Colliers de Perles on two layers are provided with hydrogen bonds in IMGT/3Dstructure-DB (http://www.imgt.org) (Ehrenmann et al. 2010b).

IMGT COLLIER DE PERLES FOR G DOMAIN

"IMGT Collier de Perles for G domain" (or "IMGT_Collier_de_Perles_for_G_domain") is a leafconcept of the "IMGT Collier de Perles" concept of numerotation. It represents the G domain that includes the G-DOMAIN of MH and the G-LIKE-DOMAIN of the MhSF proteins other than MH. It is based on the IMGT unique numbering for G domain (Lefranc et al. 2005b) **(IMGT Unique Numbering for the Variable (V), Constant (C), and Groove (G) Domains of IG, TR, MH, IgSF, and MhSF)** (Lefranc 2011e). The topology and 3D structure of a G-LIKE-DOMAIN are very similar to those of an MH G-DOMAIN (Fig. 4A). Two G domains are necessary to form the characteristic groove of the MhSF proteins that

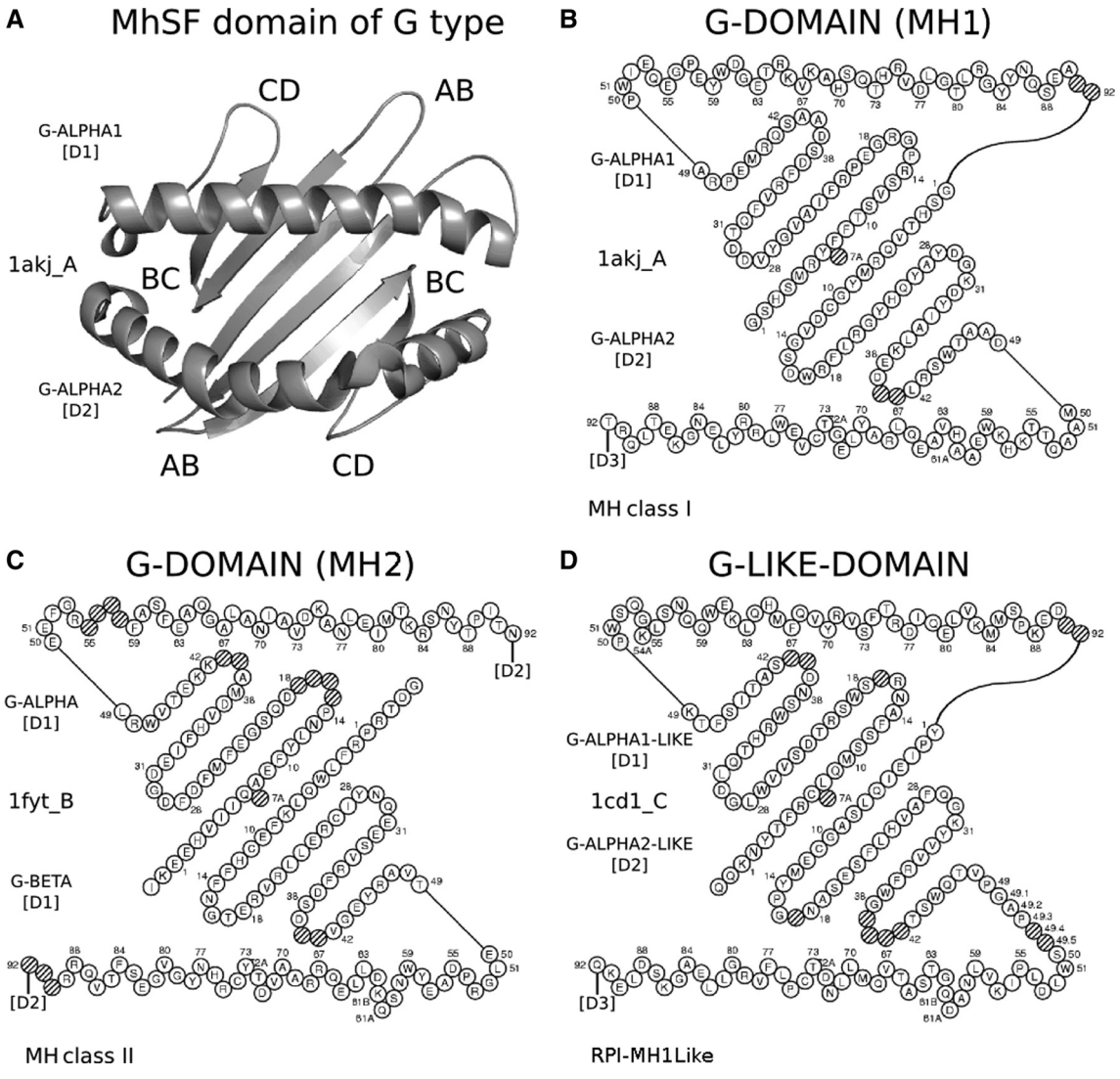


FIGURE 4. IMGT Collier de Perles for G domain. (*A*) Ribbon representation of two G-DOMAINs as an example. A similar topology and 3D structure characterize the G-LIKE-DOMAIN. (*B*) G-DOMAIN of MH class I (MH1). G-ALPHA1 [D1] and G-ALPHA2 [D2] (*Homo sapiens* HLA-A*0201). (*C*) G-DOMAIN of MH class II (MH2). G-ALPHA [D1] and G-BETA [D1] (*Homo sapiens* HLA-DRA*0101 and HLA-DRB1*0101). (*D*) G-LIKE-DOMAIN of RPI-MH1Like. G-ALPHA1-LIKE [D1] and G-ALPHA2-LIKE [D2] (*Mus musculus* CD1D1). Amino acids are shown in the one-letter abbreviation. Hatched circles correspond to missing positions according to the IMGT unique numbering for G domain (Lefranc et al. 2005b). In IMGT Colliers de Perles, position 7A is only displayed in the G-ALPHA, G-ALPHA1, and G-ALPHA1-LIKE domains, and positions 61A and 61B in the G-BETA, G-ALPHA2, and G-ALPHA2-LIKE domains. Position 54A is only occupied in G-ALPHA1-LIKE of RPI-MH1like proteins. Position 92A is only added for HLA-DMA and H2-DMA IMGT Colliers de Perles. Note that the N-terminal end of a peptide in the cleft would be on the *left* side. IMGT Colliers de Perles are available in IMGT/3Dstructure-DB (http://www.imgt.org): 1akj_A, 1fyt_B, 1cd1_C.

comprise the MH (MH class I or MH1, and MH class II or MH2) and the RPI-MH1Like proteins (Lefranc et al. 2005b).

A G domain consists of about 90 amino acids in four antiparallel β strands (A, B, C, and D), linked by turns (AB, BC, CD), forming half of the groove floor, and one α helix, forming one wall of the groove. For each G domain, the positions that contribute to the groove floor comprise positions 1–49, corresponding to the four strands linked by turns (Lefranc et al. 2005b). The numbering of the G domain helix starts at position 50 and ends at position 92 (Fig. 4).

Interestingly, despite the high sequence divergence, only five additional positions (54A, 61A, 61B, 72A, and 92A) are necessary to align any G-DOMAIN and G-LIKE-DOMAIN (Lefranc et al. 2005b). It is worthwhile to note that position 54A in G-ALPHA1-LIKE is the only additional position needed to extend the IMGT numbering for G-DOMAIN to the G-LIKE-DOMAIN.

The helix (positions 50–92) sits on the β sheet and its axis forms an angle of about 40 degrees with the β strands. Two cysteines, CYS-11 (in strand A) and CYS-74 (in the helix) are well conserved in the G-ALPHA2, G-BETA, and G-ALPHA2-LIKE domains, where they participate to a disulfide bridge that fastens the helix on the groove floor. The IMGT Collier de Perles allows one to describe specific features (Lefranc et al. 2005b; Kaas et al. 2007; Kaas and Lefranc 2007). As an example, the G-ALPHA1 and G-ALPHA domains have a conserved N-glycosylation site at position 86 (N-X-S/T, where N is asparagine, X is any amino acid except proline, S is serine, and T is threonine).

If 3D structures of peptide/MH are available, IMGT Colliers de Perles are provided with IMGT pMH contact analysis, in IMGT/3Dstructure-DB (http://www.imgt.org) (Kaas et al. 2004, 2008; Kaas and Lefranc 2005; Ehrenmann et al. 2010b).

WHAT DO IMGT COLLIERS DE PERLES TEACH US?

1. Any domain represented by an IMGT Collier de Perles is characterized by the **length** of its strands, loops, and turns and, for the G domain, by the length of its helix (Lefranc 1997, 1999; Lefranc et al. 2003, 2005b,c) **(IMGT Unique Numbering for the Variable (V), Constant (C), and Groove (G) Domains of IG, TR, MH, IgSF, and MhSF)** (Lefranc 2011e). The strand, loop, turn, or helix lengths (number of amino acids or by extrapolation number of codons, that is the number of occupied positions) become crucial information that characterizes the domains. This first feature of the IMGT standardization based on the IMGT unique numbering allowed us, for instance, to show that the distinction between the C1, C2, I1, and I2 domain types found in the literature and in the databases to describe the IgSF C domains is unnecessary and moreover inapplicable when dealing with sequences for which no structural data are known (discussed in Lefranc et al. 2005c).
2. A second feature of the IMGT standardization is the comparison of cDNA and/or amino acid sequences with **genomic** sequences, and the identification of the splicing sites, to delimit precisely the domains: a V-LIKE-DOMAIN, a C-DOMAIN, a C-LIKE-DOMAIN, a G-DOMAIN, or a G-LIKE-DOMAIN is usually encoded by a unique exon or more rarely by two spliced exons (Lefranc et al. 2003, 2005b,c). This IMGT standardization for the domain delimitation explains the discrepancies observed with the generalist UniProt/Swiss-Prot database, which identifies domains based on amino acid sequences and does not take into account the genomic information. The IMGT Collier de Perles also addresses the question of the leader region. Indeed, the N-terminal end of the first domain of an IgSF or MhSF chain depends on the proteolytic cleavage site of the leader region (peptide signal), which is rarely determined experimentally. When this site is not known, the IMGT Colliers de Perles start with the first amino acid resulting from the splicing (usually a splicing frame 1) ("Splicing sites" in IMGT Aide-mémoire, http://www.imgt.org). For an IG and TR V-DOMAIN, the leader proteolytic site is known (or is extrapolated) and the IMGT Colliers de Perles start with the first amino acid of the V-REGION (Lefranc and Lefranc 2001a,b).
3. The IMGT Colliers de Perles allow a precise visualization of the differences between proteins and/or between species for the IgSF V and C domain strands and loops, and MhSF G domain strands and helix, even in the absence of 3D structures. This has been applied, for example, to the CD28 family members and their B7 family ligands and to the BTLA proteins that belong to the IgSF by their V and/or C domains (Garapati and Lefranc 2007) and to other members of the IgSF (Bertrand et al. 2004; Duprat et al. 2004) and MhSF (Frigoul and Lefranc 2005). The IMGT Colliers de Perles are particularly useful in molecular engineering and antibody humanization design based on CDR grafting. Indeed, they allow one to precisely define the CDR-IMGT and to easily compare the amino acid

sequences of the four FR-IMGT (FR1-IMGT: positions 1–26, FR2-IMGT: 39–55, FR3-IMGT: 66–104, and FR4-IMGT: 118–128) between the mouse (or other species) and the closest human V-DOMAIN. Analyses performed on humanized therapeutic antibodies underline the importance of a correct delimitation of the CDR regions to be grafted (Lefranc 2009; Ehrenmann et al. 2010a).

4. The IMGT Colliers de Perles also allow a comparison to the IMGT Colliers de Perles **statistical profiles** for the human expressed IGHV, IGKV, and IGLV repertoires. These statistical profiles are based on the definition of 11 IMGT amino acid physicochemical characteristics classes that take into account the hydropathy, volume, and chemical characteristics of the 20 common amino acids (Pommié et al. 2004) ("Amino acids" in IMGT Aide-mémoire, http://www.imgt.org). The statistical profiles identified positions that are conserved for the physicochemical characteristics: 41 FR-IMGT positions for the human IGHV and 59 FR-IMGT positions for the human IGKV and IGLV at >80% threshold (see Plate 3 in Pommié et al. 2004). After assignment of the IMGT Collier de Perles amino acids to the IMGT amino acid physicochemical classes, comparison can be made with the statistical profiles of the human expressed repertoires. This comparison is useful to identify potential immunogenic residues at given positions in chimeric or humanized antibodies or to evaluate immunogenicity of therapeutic antibodies.
5. IMGT Colliers de Perles are also of interest when **3D structures** are available. In IMGT/3Dstructure-DB (Ehrenmann et al. 2010b), "IMGT Collier de Perles on 2 layers" are displayed with hydrogen bonds for V and C domains. Clicking on a residue in "IMGT Collier de Perles on one layer" gives access to the corresponding IMGT Residue@Position card, which provides the atom contact types and atom contact categories for that amino acid. IMGT Colliers de Perles display the IMGT pMH contact sites for 3D structures with peptide/MH (pMH) complexes, which can be compared with the pMH contact sites available in IMGT/3Dstructure-DB (Kaas et al. 2004, 2008; Kaas and Lefranc 2005; Ehrenmann et al. 2010b).

The IMGT Colliers de Perles for the V, C, and G domains, based on the IMGT unique numbering, therefore, represent a major step forward for the comparative analysis of the sequences and structures of the IgSF and MhSF domains, for the study of their evolution, and for applications in antibody engineering, IG and TR repertoires in autoimmune diseases and leukemias, pMH contact analysis, and more generally, ligand-receptor interactions involving V, C, and/or G domains.

ACKNOWLEDGMENTS

We are grateful to Chantal Ginestoux, Véronique Giudicelli, Patrice Duroux, François Ehrenmann, Gérard Lefranc, and the IMGT team for their motivation and expertise.

REFERENCES

Bertrand G, Duprat E, Lefranc M-P, Marti J, Coste J. 2004. Characterization of human fcgr3b*02 (hna-1b, na2) cDNAs and imgt standardized description of fcgr3b alleles. *Tissue Antigens* **64:** 119–131.

Brochet X. 2008. Conception et intégration d'un système d'information dédié à l'analyse et à la gestion des séquences réarrangées des récepteurs d'antigènes au sein d'IMGT: application à la Leucémie Lymphoïde Chronique, *PhD Thesis*, Université Montpellier 1, Montpellier, France.

Brochet X, Lefranc M-P, Giudicelli V. 2008. IMGT/V-QUEST: The highly customized and integrated system for IG and TR standardized V-J and V-D-J sequence analysis. *Nucl Acids Res* **36:** W503–W508.

Duprat E, Kaas Q, Garelle V, Lefranc G, Lefranc M-P. 2004. IMGT standardization for alleles and mutations of the V-LIKE-DOMAINs and C-LIKE-DOMAINs of the immunoglobulin superfamily. In *Recent research developments in human genetics* (ed. SG Pandalai), Vol 2, pp. 111–136. Research Signpost, Trivandrum, Kerala, India.

Duroux P, Kaas Q, Brochet X, Lane J, Ginestoux C, Lefranc M-P, Giudicelli V. 2008. IMGT-Kaleidoscope, the Formal IMGT-ONTOLOGY paradigm. *Biochimie* **90:** 570–583.

Ehrenmann F, Lefranc M-P. 2011a. IMGT/DomainGapAlign: IMGT standardized analysis of amino acid sequences of variable, constant and groove domains (IG, TR, MH, IgSF, MhSF). *Cold Spring Harb Protoc* doi: 10.1101/pdb.prot5636.

Ehrenmann F, Lefranc M-P. 2011b. IMGT/3Dstructure-DB: Querying the IMGT database for 3D structures in immunology and immunoinformatics (IG or antibodies, TR, MH, RPI, and FPIA). *Cold Spring Harb Protoc* doi: 10.1101/pdb.prot5637.

Ehrenmann F, Duroux P, Giudicelli V, Lefranc M-P. 2010a. Standardized sequence and structure analysis of antibody using IMGT. In *Antibody engineering* 2nd ed. (ed. R Kontermann, S Dübel), Vol 2, pp. 11–31. Springer-Verlag, Berlin/Heidelberg, Germany.

Ehrenmann F, Kaas Q, Lefranc M-P. 2010b. IMGT/3Dstructure-DB and IMGT/DomainGapAlign: A database and a tool for immunoglobulins or antibodies, T cell receptors, MHC, IgSF and MhcSF. *Nucl Acids Res* **38:** D301–D307.

Ehrenmann F, Giudicelli V, Brochet X, Lefranc M-P. 2011. IMGT/Collier de Perles: IMGT standardized representation of domains (IG, TR, and IgSF variable and constant domains, MH and MhSF groove domains). *Cold Spring Harb Protoc* doi: 10.1101/pdb.prot5635.

Frigoul A, Lefranc M-P. 2005. MICA: standardized IMGT allele nomenclature, polymorphisms and diseases. In *Recent research*

developments in human genetics (ed. SG Pandalai), Vol 3, pp. 95–145. Research Signpost, Trivandrum, Kerala, India.

Garapati VP, Lefranc M-P. 2007. IMGT Colliers de Perles and IgSF domain standardization for T cell costimulatory activatory (CD28, ICOS) and inhibitory (CTLA4, PDCD1 and BTLA) receptors. *Dev Comp Immunol* **31:** 1050–1072.

Giudicelli V, Lefranc M-P. 1999. Ontology for Immunogenetics: The IMGT-ONTOLOGY. *Bioinformatics* **12:** 1047–1054.

Giudicelli V, Lefranc M-P. 2009. IMGT standardized analysis of immunoglobulin rearranged sequences. In *Immunoglobulin gene analysis in chronic lymphocytic leukemia* (ed. P Ghia et al.), pp. 33–52. Wolters Kluwer Health Italy Ltd, Milan, Italy.

Giudicelli V, Lefranc M-P. 2011. IMGT/JunctionAnalysis: IMGT standardized analysis of the V-J and V-D-J junctions of the rearranged immunoglobulins (IG) and T cell receptors (TR). *Cold Spring Harb Protoc* doi: 10.1101/pdb.prot5634.

Giudicelli V, Chaume D, Lefranc M-P. 2004. IMGT/V-QUEST, an integrated software for immunoglobulin and T cell receptor V-J and V-D-J rearrangement analysis. *Nucl Acids Res* **32:** W435–W440.

Giudicelli V, Brochet X, Lefranc M-P. 2011. IMGT/V-QUEST: IMGT standardized analysis of the immunoglobulin (IG) and T cell receptor (TR) nucleotide sequences. *Cold Spring Harb Protoc* doi: 10.1101/pdb.prot5633.

Kaas Q, Lefranc M-P. 2005. T cell receptor/peptide/MHC molecular characterization and standardized pMHC contact sites in IMGT/3Dstructure-DB. *In Silico Biol*, **5:** 505–528.

Kaas Q, Lefranc M-P. 2007. IMGT Colliers de Perles: Standardized sequence-structure representations of the IgSF and MhcSF superfamily domains. *Current Bioinformatics* **2:** 21–30.

Kaas Q, Ruiz M, Lefranc M-P. 2004. IMGT/3Dstructure-DB and IMGT/StructuralQuery, a database and a tool for immunoglobulin, T cell receptor and MHC structural data. *Nucl Acids Res* **32:** D208–D210.

Kaas Q, Ehrenmann F, Lefranc M-P. 2007. IG, TR, MHC, IgSf and MhcSF: What do we learn from the IMGT Colliers de Perles? *Brief Funct Genomic Proteomic* **6:** 253–264.

Kaas Q, Duprat E, Tourneur G, Lefranc M-P. 2008. IMGT standardization for molecular characterization of the T cell receptor/peptide/MHC complexes. In *Immunoinformatics* (ed. C Schoenbach et al.), pp. 19–49. Immunomics Reviews, Series of Springer Science and Business Media LLC, Springer, New York.

Lefranc M-P. 1997. Unique database numbering system for immunogenetic analysis. *Immunol Today* **18:** 509. doi: 10.1016/S0167-5699(97)01163-8.

Lefranc M-P. 1999. The IMGT unique numbering for Immunoglobulins, T cell receptors and Ig-like domains. *The Immunologist* **7:** 132–136.

Lefranc M-P. 2005. IMGT, the international ImMunoGeneTics information system: A standardized approach for immunogenetics and immunoinformatics. *Immunome Res* **1:** 3. doi: 10.1186/1745-7580-1-3.

Lefranc M-P. 2008a. IMGT, the international ImMunoGeneTics information system for immunoinformatics. Methods for querying IMGT databases, tools and Web resources in the context of immunoinformatics. *Mol Biotechnol* **40:** 101–111.

Lefranc M-P. 2008b. IMGT-ONTOLOGY, IMGT databases, tools and Web resources for Immunoinformatics. In *Immunoinformatics* (ed. C Schoenbach et al.), pp. 1–18. Immunomics Reviews, Series of Springer Science and Business Media LLC, Springer, New York.

Lefranc M-P. 2009. Antibody database and tools: The IMGT experience. In *Therapeutic monoclonal antibodies: from bench to clinic* (ed. A Zhiqiang), pp. 91–114. John Wiley Sons, Inc, Hoboken, NJ.

Lefranc M-P. 2011a. IMGT, the International ImMunoGeneTics Information System. *Cold Spring Harb Protoc* doi: 10.1101/pdb.top115.

Lefranc M-P. 2011b. From IMGT-ONTOLOGY IDENTIFICATION axiom to IMGT standardized keywords: For immunoglobulins (IG), T cell receptors (TR), and conventional genes. *Cold Spring Harb Protoc* doi: 10.1101/pdb.ip82.

Lefranc M-P. 2011c. From IMGT-ONTOLOGY DESCRIPTION axiom to IMGT standardized labels: For immunoglobulin (IG) and T cell receptor (TR) sequences and structures. *Cold Spring Harb Protoc* doi: 10.1101/pdb.ip83.

Lefranc M-P. 2011d. From IMGT-ONTOLOGY CLASSIFICATION axiom to IMGT standardized gene and allele nomenclature: For immunoglobulins (IG) and T cell receptors (TR). *Cold Spring Harb Protoc* doi: 10.1101/pdb.ip84.

Lefranc M-P. 2011e. IMGT unique numbering for the variable (V), constant (C), and groove (G) domains of IG, TR, MH, IgSF, and MhSF. *Cold Spring Harb Protoc* doi: 10.1101/pdb.ip85.

Lefranc M-P, Lefranc G. 2001a. *The Immunoglobulin FactsBook*, 1–458. Academic Press, London, UK.

Lefranc M-P, Lefranc G. 2001b. *The T cell receptor FactsBook*, 1–398. Academic Press, London, UK.

Lefranc M-P, Pommié C, Ruiz M, Giudicelli V, Foulquier E, Truong L, Thouvenin-Contet V, Lefranc G. 2003. IMGT unique numbering for immunoglobulin and T cell receptor variable domains and Ig superfamily V-like domains. *Dev Comp Immunol* **27:** 55–77.

Lefranc M-P, Giudicelli V, Ginestoux C, Bosc N, Folch G, Guiraudou D, Jabado-Michaloud J, Magris S, Scaviner D, Thouvenin V, et al. 2004. IMGT-ONTOLOGY for Immunogenetics and Immunoinformatics. *In Silico Biol* **4:** 17–29.

Lefranc M-P, Clément O, Kaas Q, Duprat E, Chastellan P, Coelho I, Combres K, Ginestoux C, Giudicelli V, Chaume D, et al. 2005a. IMGT-Choreography for Immunogenetics and Immunoinformatics. *In Silico Biol*, **5:** 45–60.

Lefranc M-P, Duprat E, Kaas Q, Tranne M, Thiriot A, Lefranc G. 2005b. IMGT unique numbering for MHC groove G-DOMAIN and MHC superfamily (MhcSF) G-LIKE-DOMAIN. *Dev Comp Immunol* **29:** 917–938.

Lefranc M-P, Pommié C, Kaas Q, Duprat E, Bosc N, Guiraudou D, Jean C, Ruiz M, Da Piedade I, Rouard M, et al. 2005c. IMGT unique numbering for immunoglobulin and T cell receptor constant domains and Ig superfamily C-like domains. *Dev Comp Immunol* **29:** 185–203.

Lefranc M-P, Giudicelli V, Regnier L, Duroux P. 2008. IMGT, a system and an ontology that bridge biological and computational spheres in bioinformatics. *Brief Bioinform* **9:** 263–275.

Lefranc M-P, Giudicelli V, Ginestoux C, Jabado-Michaloud J, Folch G, Bellahcene F, Wu Y, Gemrot E, Brochet X, Lane J, et al. 2009. IMGT, the international ImMunoGeneTics information system. *Nucl Acids Res* **37:** D1006–D1012.

Poiron C, Wu Y, Ginestoux C, Ehrenmann F, Duroux P, Lefranc M-P. 2010. IMGT/mAb-DB: The IMGT database for therapeutic monoclonal antibodies. JOBIM 2010 Poster 13. http://www.jobim2010.fr/?q=fr/node/55.

Pommié C, Levadoux S, Sabatier R, Lefranc G, Lefranc M-P. 2004. IMGT standardized criteria for statistical analysis of immunoglobulin V-REGION amino acid properties. *J Mol Recognit* **17:** 17–32.

Ruiz M, Lefranc M-P. 2002. IMGT gene identification and Colliers de Perles of human immunoglobulin with known 3D structures. *Immunogenetics* **53:** 857–883.

Protocol

IMGT/V-QUEST: IMGT Standardized Analysis of the Immunoglobulin (IG) and T Cell Receptor (TR) Nucleotide Sequences

Véronique Giudicelli, Xavier Brochet, and Marie-Paule Lefranc[1]

IMGT, the international ImMunoGeneTics information system, Laboratoire d'ImmunoGénétique Moléculaire LIGM, Université Montpellier 2, Institut de Génétique Humaine IGH, UPR CNRS 1142, 34396 Montpellier cedex 5, France

INTRODUCTION

IMGT/V-QUEST is the integrated online IMGT tool for the standardized analysis of the immunoglobulin (IG) or antibody and T cell receptor (TR) rearranged nucleotide sequences. The analysis of these antigen receptors represents a crucial challenge for the study of the adaptative immune response in normal and disease-related situations. The expressed IG and TR repertoires represent a potential of 10^{12} IG and 10^{12} TR per individual. This huge diversity results from mechanisms that occur at the DNA level during IG molecular synthesis, including the combinatorial rearrangement of the variable (V), diversity (D), and joining (J) genes, the N diversity (V, D, and J exonuclease trimming and addition at random of nucleotides by the terminal deoxynucleotidyl transferase during the V-(D)-J rearrangement) and, for IG, somatic hypermutations. IMGT/V-QUEST, as described here, identifies the V, D, J genes and alleles by alignment with the germline IG and TR gene and allele sequences of the IMGT reference directory. The tool describes the V-REGION mutations and identifies the hot spot positions in the closest germline V gene. IMGT/V-QUEST integrates IMGT/JunctionAnalysis for a detailed analysis of the V-J and V-D-J junctions and IMGT/Automat for a complete sequence annotation of the sequences, and provides IMGT Collier de Perles.

RELATED INFORMATION

IMGT/V-QUEST (Lefranc 2004; Giudicelli and Lefranc 2005, 2008; Brochet et al. 2008) is part of IMGT, the international ImMunoGeneTics information system, http://www.imgt.org (Lefranc et al. 2009). Standardization and IMGT Scientific chart rules are based on the IMGT-ONTOLOGY concepts of identification, classification, description (Giudicelli and Lefranc 1999; Lefranc et al. 2004, 2005a, 2008), and numerotation (Lefranc 1997, 1999; Lefranc et al. 2003, 2005b,c) generated from the axioms of the Formal IMGT-ONTOLOGY or IMGT-Kaleidoscope (Duroux et al. 2008). The concepts of identification led to the IMGT standardized keywords, the concepts of classification to the IMGT standardized gene and allele names, the concepts of description to the IMGT standardized labels, and the concepts of numerotation to the IMGT unique numbering.

A detailed description of IMGT is provided in **IMGT, the International ImMunoGeneTics Information System** (Lefranc 2011a). Information is also available on **From IMGT-ONTOLOGY IDENTIFICATION Axiom to IMGT Standardized Keywords: For Immunoglobulins (IG), T Cell Receptors (TR), and Conventional Genes** (Lefranc 2011b), **From IMGT-ONTOLOGY DESCRIPTION Axiom to IMGT Standardized Labels: For Immunoglobulin (IG) and T Cell Receptor (TR) Sequences and Structures** (Lefranc 2011c), **From IMGT-ONTOLOGY CLASSIFICATION Axiom to IMGT Standardized Gene and Allele Nomenclature: For Immunoglobulins (IG) and T Cell Receptors (TR)** (Lefranc 2011d), **IMGT Unique Numbering for the Variable (V), Constant (C), and Groove (G) Domains of IG, TR, MH, IgSF, and MhSF** (Lefranc 2011e), and **IMGT Collier de Perles for the Variable (V), Constant (C), and Groove (G) Domains of IG, TR, MH, IgSF, and MhSF** (Lefranc 2011f).

In addition, protocols are available for **IMGT/JunctionAnalysis: IMGT Standardized Analysis of the V-J and V-D-J Junctions of the Rearranged Immunoglobulins (IG) and T Cell Receptors (TR)** (Giudicelli and Lefranc 2011), **IMGT/Collier de Perles: IMGT Standardized Representation of Domains (IG, TR, and IgSF Variable and Constant Domains, MH and MhSF Groove Domains)** (Ehrenmann et al. 2011), **IMGT/DomainGapAlign: IMGT Standardized Analysis of Amino Acid Sequences of**

[1]Corresponding author (Marie-Paule.Lefranc@igh.cnrs.fr).
Cite as: Cold Spring Harb Protoc; 2011; doi:10.1101/pdb.prot5633

Variable, Constant, and Groove Domains (IG, TR, MH, IgSF, MhSF) (Ehrenmann and Lefranc 2011a), and **IMGT/3Dstructure-DB: Querying the IMGT Database for 3D Structures in Immunology and Immunoinformatics (IG or Antibodies, TR, MH, RPI, and FPIA)** (Ehrenmann and Lefranc 2011b).

MATERIALS

It is essential that you consult the appropriate Material Safety Data Sheets and your institution's Environmental Health and Safety Office for proper handling of equipment and hazardous materials used in this protocol.

Equipment

Computer (Internet-connected)

METHOD

1. Using any modern web browser, access the IMGT Home page, http://www.imgt.org and click, in the "IMGT tools" section, on the link to IMGT/V-QUEST.

   This gives access to the IMGT/V-QUEST Welcome page.

IMGT/V-QUEST Welcome Page

2. In the IMGT/V-QUEST Welcome page (Fig. 1), locate the following sections:

   i. "Analyze your immunoglobulin (IG) or antibody nucleotide sequences"

   This section shows the species for which the analysis of IG sequences is available.

   There is no need to specify more precisely the type of IG chain or locus, as the tool will automatically identify whether sequences are IGH, IGK, or IGL by selecting one species or taxon in that section. This works for species with complete IMGT reference directories, such as human (Lefranc 2000a; Lefranc and Lefranc 2001a) and mouse (Giudicelli et al. 2005a). This also holds for the other species or taxons with incomplete IMGT reference directories; however, in those cases, results should be analyzed considering the status of the IMGT reference directory (information on the updates on the web site).

   ii. "Analyze your T cell Receptor (TR) nucleotide sequences"

   This section shows the species for which the analysis of TR sequences is available.

   There is no need to specify more precisely the type of TR chain or locus, as the tool will automatically identify whether sequences are TRA, TRB, TRG, or TRD by selecting one species or taxon in that section. This works for species with complete IMGT reference directories, such as human (Lefranc 2000b; Lefranc and Lefranc 2001b) and mouse (Giudicelli et al. 2005a). This also holds true for the other species or taxons with incomplete IMGT reference directories; however, in those cases, results should be analyzed considering the status of the IMGT reference directory (information on the updates on the web site).

   iii. "Download IMGT reference directory in FASTA format (IG and TR)"

   This section shows the species for which the IMGT reference directory is available.

   Clicking on one of the links gives access to the data (overview and IMGT/V-QUEST reference directory per species or taxon), if needed.

3. Choose the IG or TR section (see Step 2), depending on your sequences.

4. Click on the species (e.g., "Human") or taxon that corresponds to the IMGT reference directory against which you want your sequences to be analyzed.

   This will open the IMGT/V-QUEST Search page. That page has a similar display regardless of the choice IG or TR (Step 3) and the species (Step 4).

IMGT/V-QUEST Search Page

5. In the IMGT/V-QUEST Search page (Fig. 2), and below the title that reflects your choice IG or TR (Step 3), locate the following.

   i. "Your selection" that recalls your selection, for example "Human" (Step 4)

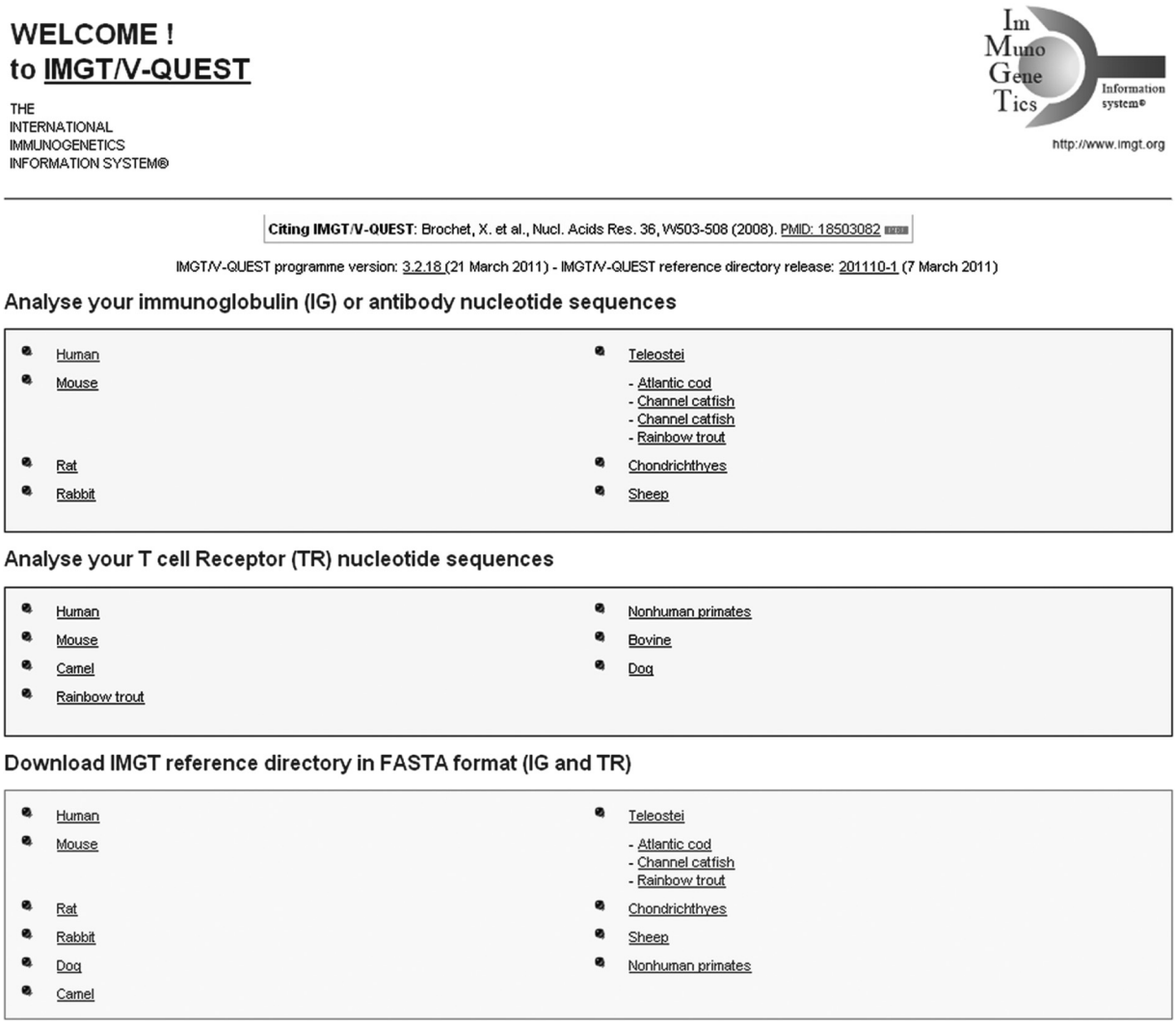


FIGURE 1. IMGT/V-QUEST Welcome page.

A note indicates the IMGT reference directory set against which your sequences will be analyzed, e.g., "Human (Homo sapiens) IG set," as determined by the tool from Steps 3 and 4. A link provides access to the IMGT reference directory sets.

A link for sequence sets allows one to test IMGT/V-QUEST, if needed (only for human IG sequences).

ii. The sequence submission section

That section comprises:

Two windows to either type (or copy/paste) your sequence(s) or upload a file, in FASTA format (proceed to Step 6 for entering your sequences).

The "Start" and "Clear the form" buttons.

iii. "Selection for results"

IMGT/V-QUEST provides the choice between three types of selection for the results:

"A. Detailed view." This displays the results of the analysis for each sequence, individually (proceed to Step 8).

"B. Synthesis view." This displays the alignments of sequences that express the same V gene and allele and allows one to compare the localization of the mutations and the composition of their junctions (proceed to Step 14 for choice of the parameters).

"C. Excel file." This allows the users to get the results of the analysis in a spreadsheet with the choice of 11 sheets (proceed to Step 21 for choice of the parameters).

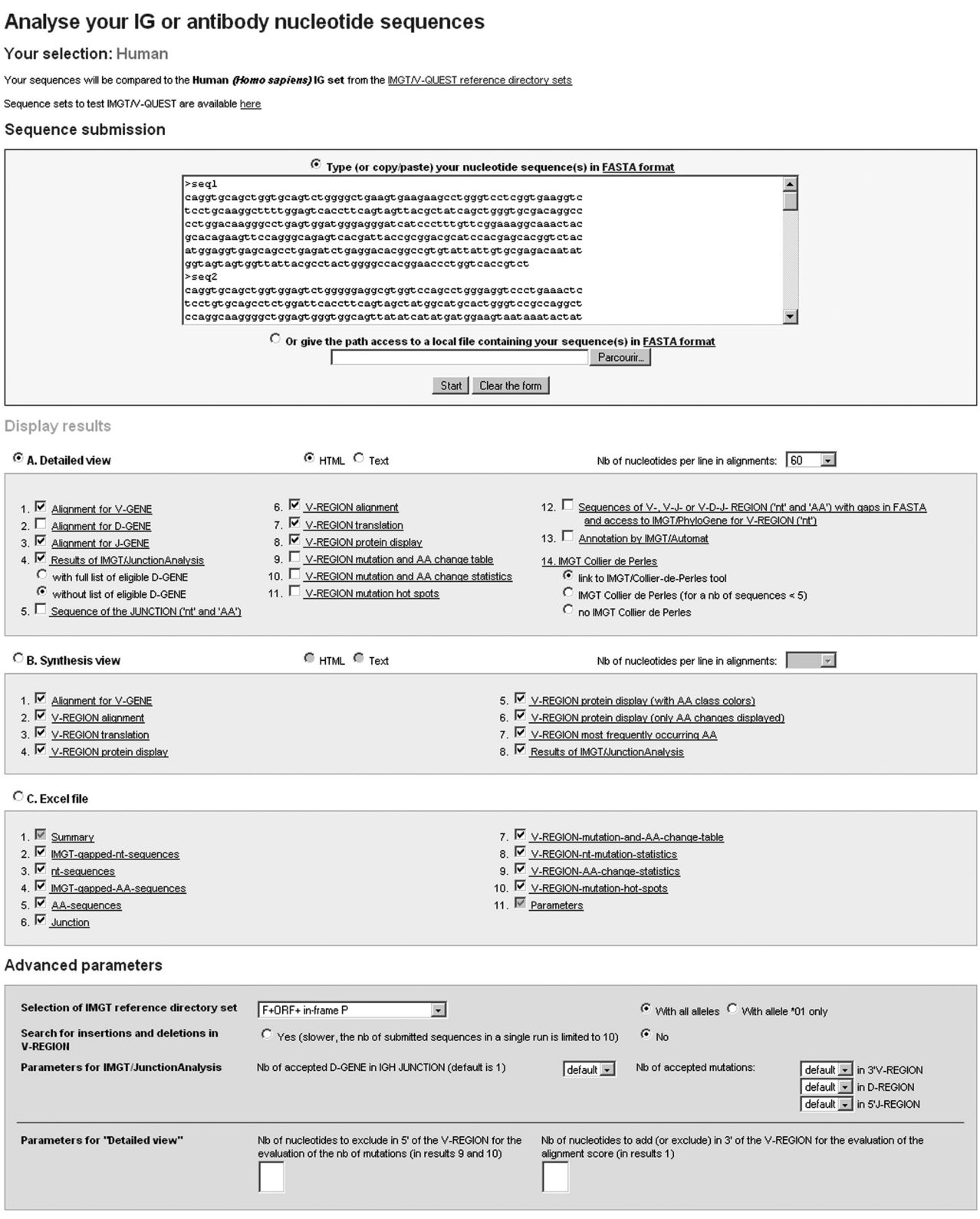


FIGURE 2. The IMGT/V-QUEST Search page. The title "Analyze your IG or antibody nucleotide sequences" and the selection "Human" reflect the user's choice.

iv. "Advanced parameters"

"Advanced parameters" allows one to modify the parameters for sophisticated queries or unusual sequences (proceed to Step 11 for choice of "Advanced parameters").

Entering Nucleotide Sequences for Analysis

6. Copy and paste your sequence(s) in the text area of "Nucleotide sequences," or upload a text file by selecting the option "Or give the path access to a local file containing your sequence(s) in FASTA format" and using the "Browse" button to select the file.

   The analyzed sequences are the V-J or V-D-J regions that encode the variable domain (V-DOMAIN) of the antigen receptors. The analysis is performed for sequences from genomic DNA (gDNA) or complementary DNA (cDNA) without need for the user to specify the molecule type. If the submitted sequences are in reverse orientation, the tool will automatically "complementary reverse" them and will provide the user sequences and the results in the sense orientation. The online IMGT/V-QUEST can analyze up to 50 sequences in a single run. The sequences must be formatted in FASTA format. Note that FASTA headers longer than 30 characters will be truncated in the results.

7. Depending on the selection you want for your results, proceed to:
   i. Step 8 for A. Detailed view
   ii. Step 14 for B. Synthesis view
   iii. Step 21 for C. Excel file

IMGT/V-QUEST Detailed View

8. Select the option "A. Detailed view" (selected by default).

9. Select the format of the results, "HTML" (by default) or "Text," and the "Nb of nucleotides per line in alignments" (60 by default).

10. Select, in the light orange box, the displays of the results to be shown.

    There is a choice of 14 different result displays for the analysis of the V-DOMAIN (Lefranc 2004; Giudicelli and Lefranc 2005, 2008; Brochet et al. 2008). The most commonly requested are already selected by default. You can check or uncheck the result displays as needed.

    i. Alignment for V-GENE
    ii. Alignment for D-GENE
    iii. Alignment for J-GENE
    iv. Results of IMGT/JunctionAnalysis with full list of eligible D-GENE or without list of eligible D-GENE
    v. Sequence of the JUNCTION (nucleotide ["nt"] and amino acid ["AA"])
    vi. V-REGION alignment
    vii. V-REGION translation
    viii. V-REGION protein display
    ix. V-REGION mutation and AA change table
    x. V-REGION mutation and AA change statistics
    xi. V-REGION mutation hot spots
    xii. Sequences of V-, V-J-, or V-D-J-REGION ("nt" and "AA") with gaps, in FASTA and access to IMGT/PhyloGene for V-REGION ("nt")
    xiii. Annotation by IMGT/Automat
    xiv. IMGT Colliers de Perles: link to IMGT/Collier de Perles tool **IMGT/Collier de Perles: IMGT Standardized Representation of Domains (IG, TR, and IgSF Variable and Constant Domains, MH and MhSF Groove Domains)** (Ehrenmann et al. 2011), IMGT Collier de Perles (for a number [nb] of sequences less than five), or no IMGT Collier de Perles

11. Go to "Advanced parameters" at the bottom of the search page and modify the parameters according to the particularities of the sequences, if needed.

    Default values of "Advanced parameters" have been set for classical analysis. They should be modified for sophisticated queries or for unusual sequences (for example, with insertions and/or deletions).

    i. "Selection of IMGT reference directory set," with a choice of four sets ("F+ORF," "F+ORF+in frame P" (by default), "F+ORF including orphons," or "F+ORF+in frame P including orphons") where F is functional, ORF is open reading frame, P is pseudogene. This allows sequences to be compared with only relevant sequences (e.g., orphon sequences are relevant for genomic, but not for expressed repertoire studies).

    *The selected set can also be chosen either "With all alleles" or "With allele *01 only."*

    ii. "Search for insertions and deletions in V-REGION" ("Yes" or "No").

    If "Yes," the search is slower and the number of submitted sequences in a single run is limited to 10.

    iii. "Parameters for IMGT/JunctionAnalysis:" "Nb of accepted D-GENE" (provided for the IGH junctions and for the TRB and TRD junctions, respectively) or "Nb of accepted mutations" in 3′V-REGION, D-REGION, and 5′J-REGION.

    iv. "Parameters for Detailed view:" "Nb of nucleotides to exclude in 5′ of the V-REGION for the evaluation of the nb of mutations" (to avoid, e.g., counting primer specific nucleotides) or "Nb of nucleotides to add (or exclude) in 3′ of the V-REGION for the evaluation of the alignment score" (e.g., in case of low or high exonuclease activity).

12. Click on the "Start" button to launch the analysis.

    "Detailed view" results are displayed in "A. Detailed results for the IMGT/V-QUEST analyzed sequences." The top of this page indicates the number of analyzed sequences with links to individual results (Fig. 3).

13. Use the link associated with a sequence name to go directly to the individual results.

    The results include the sequence in FASTA format and a Result summary (Fig. 4 and Fig. 5) and, if selected in Step 10:

    - *The Alignment for V, D, and J genes and alleles (Fig. 6)*
    - *The results of IMGT/JunctionAnalysis and the sequence of the JUNCTION ("nt" and "AA") (Fig. 7)*
    - *Different displays of the V-REGION (Fig. 8)*
    - *The analysis of the mutations (Fig. 9)*

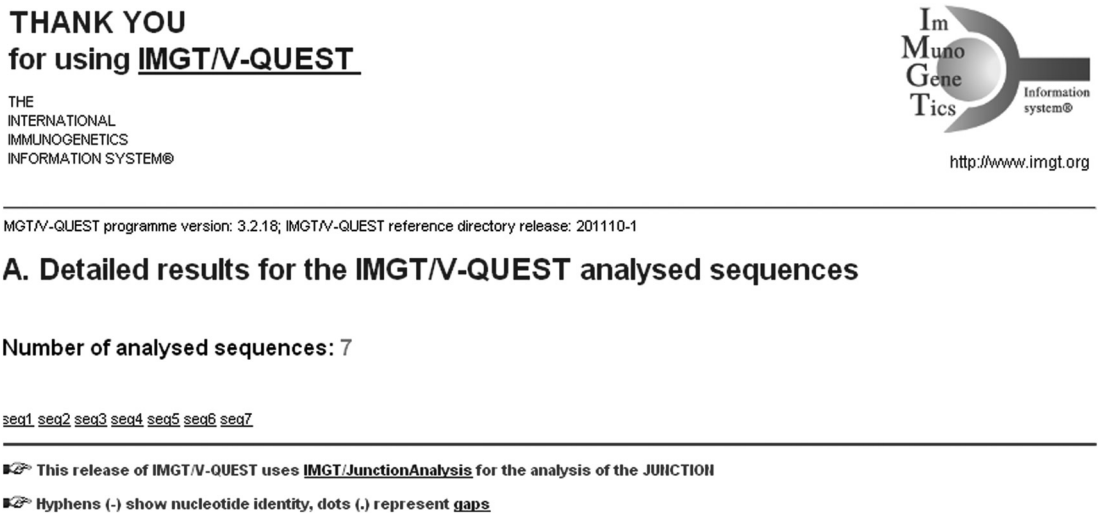


FIGURE 3. IMGT/V-QUEST "Detailed view" result page: *top* of the page. The number of analyzed sequences is indicated. The link associated to a sequence name allows one to go directly to the individual sequence result.

Sequence number 1: seq1

Sequence compared with the human IG set from the IMGT reference directory

```
>seq1
gaggtgcagctgttggagtctgggggaggcgtggtccagcctgggaggtccctgagactc
tcctgtatagcctctggattcaccttcagtagctatcctatgacctgggtccgccaggct
ccaggcaaggggctggagtgggtggcaagtatatcatatgacggaagttataaatataag
gtagactccatgaagggccgactcaccatctccagagacaattccaagaacacgctgtat
ttggaaatgaacagcctgacagctgaggacacggctgtgtattactgtgcgaggacagct
ttctttaacgcctatgacttctggggccagggaaccctggtcaccgtctcctcagcctcc
accaagggcccatcggtcttccccctggcaccctcctccaagagcacctctgggggcaca
gcggccctgggctgcctggtcaaggactacttccccgaaccggtgacggtgtcgtggaac
tcaggcgccctgaccagcggcgtgcacaccttcccggctgtcctacagtcctcaggactc
tactccctcagcagcgtggtgaccgtgccctccagcagcttgggcacccagacctacatc
tgcaacgtgaatcacaagcccagcaacaccaaggtggacaagaaagttgagcccaaatct
tgtgacaaaactcacaca
```

Result summary:	Productive IGH rearranged sequence (no stop codon and in-frame junction)		
V-GENE and allele	IGHV3-30*04	score = 1255	identity = 93,06% (268/288 nt)
J-GENE and allele	IGHJ4*02	score = 204	identity = 91,67% (44/48 nt)
D-GENE and allele by IMGT/JunctionAnalysis	IGHD5-5*01	D-REGION is in reading frame 1	
FR-IMGT lengths, CDR-IMGT lengths and AA JUNCTION	[25.17.38.11]	[8.8.11]	CARTAFFNAYDFW

FIGURE 4. IMGT/V-QUEST "Detailed view" result page: *top* part of an individual result with the sequence and the "Result summary." The *top* part of each individual result indicates the IMGT reference directory set (here, human IG set) against which the sequence was analyzed, and displays the user sequence in FASTA format (a sequence submitted in antisense orientation will be shown as a complementary reverse sequence, that is in V gene sense orientation). The "Result summary" provides a crucial feature that is the evaluation of the user sequence functionality performed by IMGT/V-QUEST: productive (if no stop codon and in-frame junction) or unproductive (if stop codons and/or out-of-frame junction). It also summarizes the main characteristics of the analyzed sequence, which includes (1) the names (with species in the six-letter abbreviation, and the functionality added in the version 3.2.19) of the closest "V-GENE and allele" (e.g., Homsap IGHV3-30*04 F) and "J-GENE and allele" (e.g., Homsap IGHJ4*02 F) with the alignment score, the percentage of identity and the ratio "number of identical nucleotides/number of aligned nucleotides," (2) the name (with species in the six-letter abbreviation, and the functionality) of the closest "D-GENE and allele" (e.g., Homsap IGHD5-5*01 F) determined by IMGT/JunctionAnalysis (Giudicelli and Lefranc 2011) with the D-REGION reading frame, (3) the framework region (FR)-IMGT lengths (e.g., [25.17.38.11]), the complementarity determining region (CDR)-IMGT lengths (e.g., [8.8.11]), and the amino acid (AA) JUNCTION sequence that characterizes a V domain. Information shown in orange (here, "Productive IGH rearranged sequence" and "93,06%" are used by clinicians analyzing the level of mutations in chronic lymphocytic leukemia (CLL) IGHV rearranged sequences as a prognostic criterion (Giudicelli and Lefranc 2008). Note that IMGT/V-QUEST provides warnings (data not shown) that appear as notes in red to alert the user if potential insertions or deletions are suspected in the V-REGION (sequences with <85% of identity and/or with different CDR1-IMGT and/or CDR2-IMGT lengths compared with the closest germline V-REGION), or whether other possibilities for the J gene and allele are identified. The seq1 accession number is AB012909 from the IMGT/LIGM-DB database (Giudicelli et al. 2006).

- *The localization of the mutation hot spots (Fig. 10)*
- *The sequences of V-, V-J-, or V-D-J- REGION ("nt" and "AA") with gaps in FASTA and access to IMGT/PhyloGene for V-REGION ("nt") (Fig. 11)*
- *The annotation by IMGT/Automat (Fig. 12)*
- *The IMGT Collier de Perles or, by default, the link to the IMGT Collier de Perles tool* ***IMGT/Collier de Perles: IMGT Standardized Representation of Domains (IG, TR, and IgSF Variable and Constant Domains, MH and MhSF Groove Domains)*** *(Fig. 13; Ehrenmann et al. 2011)*

IMGT/V-QUEST Synthesis View

14. Select the option "B. Synthesis view."
15. Select the format of the results, "HTML" (by default) or "Text," and the "Nb of nucleotides per line in alignments" (60 by default).

Sequence number 1: seq2

Sequence compared with the human IG set from the IMGT reference directory

>seq2
gaggtgaaggtggtgcagtctggagcagaggtgaaaaagcccggggagtctctgcacatc
tcctgtcagatttct**GCAGACGGT**ggagacgactttaccactctctggatcggctgggtg
cgccagatgcccgggaaagggctggagtggatggggatcatctggcctggtgactctgat
accacatacagtccgtcctttccaagggccacgtctccatttcagccgacaagtccaccagt
accgcctaccttcagtggagcagcctgaaggcctcggactccgccatgtattactgtgcg
accaacagtggttattaccccggggatggctttgattactggggccagggaaccctgatc
attgtctcctca

Result summary:	☞ Nucleotide insertions have been detected and automatically removed for this analysis: they are displayed as capital letters in the user submitted sequence above.					
	localization in V-REGION	nb of inserted nt	inserted nt	causing frameshift	from V-REGION codon	from nt position in user submitted sequence
	CDR1-IMGT	9	GCAGACGGT	no	27	76

IMGT/V-QUEST results after removal of the insertion(s): Potentially productive IGH rearranged sequence: no stop codon and in-frame junction (Check also your sequence with BLAST against IMGT/GENE-DB reference sequences to eventually identify out-of-frame pseudogenes)			
V-GENE and allele	IGHV5-51*01	score = 1201	identity = 90,97% (262/288 nt) [90,62% (261/288 nt)]
J-GENE and allele	IGHJ4*02	score = 172	identity = 85,11% (40/47 nt)
D-GENE and allele by IMGT/JunctionAnalysis	IGHD3-22*01	D-REGION is in reading frame 2	
FR-IMGT lengths, CDR-IMGT lengths and AA JUNCTION	[25.17.38.11]	[8.8.14]	CATNSGYYPGDGFDYW

FIGURE 5. IMGT/V-QUEST "Detailed view" result page: example of "Result summary" for a sequence with insertion analyzed with the option "Search for insertions and deletions." The detection and detailed description of insertions and/or deletions are shown in the "Result summary" first row to capture the user attention. Insertions appear as capital letters in the FASTA sequence. In the "V-GENE and allele" line of the "Result summary," the percentage of identity and the ratio are those obtained after removal of the insertion(s) and/or deletion(s) by IMGT/V-QUEST. Between brackets are shown the percentage of identity and ratio considering each insertion or deletion as one mutational event. This information is used by clinicians analyzing the level of mutations in CLL IGHV rearranged sequences (Giudicelli and Lefranc 2008). The seq2 accession number is L26531 from the IMGT/LIGM-DB database (Giudicelli et al. 2006).

16. Check, in the light orange box, the displays of the results to be shown.

    There are seven different result displays for the alignments of sequences that express the same V gene and allele, and the results of IMGT/JunctionAnalysis (Lefranc 2004; Giudicelli and Lefranc 2005, 2008; Brochet et al. 2008). They are all selected by default, but you can uncheck or check them as needed:

    Alignment for V-GENE
    V-REGION alignment
    V-REGION translation
    V-REGION protein display
    V-REGION protein display (with AA class colors)
    V-REGION protein display (only AA changes displayed)
    V-REGION most frequently occurring AA
    Results of IMGT/JunctionAnalysis

17. Go to "Advanced parameters" at the bottom of the search page and modify the parameters according to the particularities of the sequences, if needed (for details, see Step 11).

18. Click on "Start" to launch the analysis.

    "Synthesis view" results are displayed in "B. Synthesis for the IMGT/V-QUEST analyzed sequences." The top of this page indicates the number of analyzed sequences and displays a "Summary table" (Fig. 14).

19. Below the "Summary table," use the links associated with the locus name(s) (e.g., IGH) to display Results of IMGT/JunctionAnalysis for sequences identified by the tool as belonging to the same locus (Fig. 15).

20. Use the links associated with the gene and allele names to display the synthesis results of input sequences identified by the tool as using the same closest germline V gene and allele.

    The synthesis results provide seven different displays (if all were selected):

    - *"Alignment for V-GENE" (Fig. 16)*
    - *"V-REGION alignment according to the IMGT numbering" (Fig. 16)*
    - *"V-REGION translation"*
    - *V-REGION protein displays in three different formats (Fig. 17)*
    - *"V-REGION most frequently occurring AA per position and per FR-IMGT and CDR-IMGT" (Fig. 18)*

1. Alignment for V-GENE and allele identification

Closest V-REGIONs (evaluated from the V-REGION first nucleotide to the 2nd-CYS codon)

```
                                Score       Identity
L06615  IGHV3-30*04             1255        93,06% (268/288 nt)
M83134  IGHV3-30*01             1246        92,71% (267/288 nt)
M77300  IGHV3-30*09             1246        92,71% (267/288 nt)
M77324  IGHV3-30*14             1246        92,71% (267/288 nt)
X92283  IGHV3-30-3*01           1246        92,71% (267/288 nt)
```

Alignment with FR-IMGT and CDR-IMGT delimitations

```
                                <---------------------------------- FR1-IMGT ---------------------------------->___________
seq1                            gaggtgcagctgttggagtctgggggga...ggcgtggtccagcctgggaggtccctgagactctcctgtatagcctctggattcaccttc
L06615  IGHV3-30*04             c-----------g--------------...-----------------------------------------gc-------------------
M83134  IGHV3-30*01             c-----------g--------------...-----------------------------------------gc-------------------
M77300  IGHV3-30*09             c-----------g--------------...-----------------------------------------gc-------------------
M77324  IGHV3-30*14             c-----------g--------------...-----------------------------------------gc-------------------
X92283  IGHV3-30-3*01           c-----------g--------------...-----------------------------------------gc-------------------

                                __ CDR1-IMGT ___________<--------------------- FR2-IMGT -------------------->___________ CDR
seq1                            ............agtagctatcctatgacctgggtccgccaggctccaggcaaggggctggagtgggtggcaagtatatcatatgac...
L06615  IGHV3-30*04             ............---------g-----ca-----------------------------------------gt------------t...
M83134  IGHV3-30*01             ............---------g-----ca----------------------------a------------gt------------t...
M77300  IGHV3-30*09             ............---------g-----ca-----------------------------------------gt------------t...
M77324  IGHV3-30*14             ............---------g-----ca-----------------------------------------gt------------t...
X92283  IGHV3-30-3*01           ............---------g-----ca-----------------------------------------gt------------t...

                                2-IMGT ________<------------------------------------------------------ FR3-IMGT -----------
seq1                            ...ggaagttataaatataaggtagactccatgaag...ggccgactcaccatctccagagacaattccaagaacacgctgtatttggaa
L06615  IGHV3-30*04             ...------a-------ct-c-c-------g-----...------t-------------------------------------c--c--
M83134  IGHV3-30*01             ...------a-------ct-c-c-------g-----...------t-------------------------------------c--c--
M77300  IGHV3-30*09             ...------a-------ct-c-c-------g-----...------t--g----------------------------------c--c--
M77324  IGHV3-30*14             ...------a-------ct-c-c-------g-----...------t-------------------------------------c-tc--
X92283  IGHV3-30-3*01           ...-----ca-------ct-c-c-------g-----...------t-------------------------------------c--c--

                                ----------------------------------------->___________ CDR3-IMGT __________
seq1                            atgaacagcctgacagctgaggacacggctgtgtattactgtgcgaggacagctttctttaacgcctatgacttctggggccagggaacc
L06615  IGHV3-30*04             -------------g---------------------------------aga
M83134  IGHV3-30*01             -------------g---------------------------------aga
M77300  IGHV3-30*09             -------------g---------------------------------aga
M77324  IGHV3-30*14             -------------g---------------------------------aga
X92283  IGHV3-30-3*01           -------------g---------------------------------a


seq1                            ctggtcaccgtctcctcagcctccaccaagggcccatcggtcttcccccctggcaccctcctccaagagcacctctgggggcaca
L06615  IGHV3-30*04
M83134  IGHV3-30*01
M77300  IGHV3-30*09
M77324  IGHV3-30*14
X92283  IGHV3-30-3*01
```

---

3. Alignment for J-GENE and allele identification

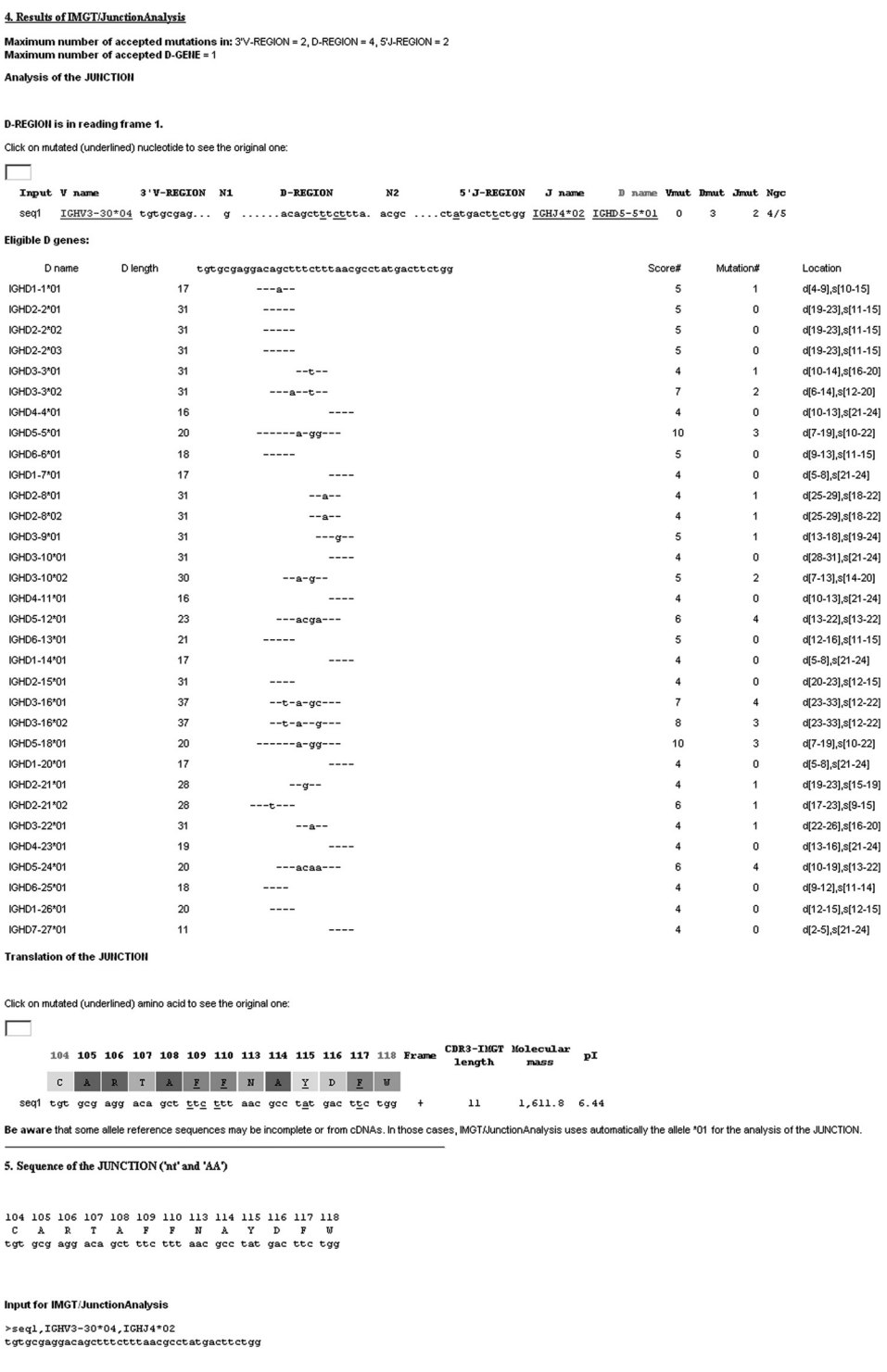
Closest J-REGIONs

```
                                Score       Identity
X86355  IGHJ4*02                204         91,67% (44/48 nt)
J00256  IGHJ4*01                195         89,58% (43/48 nt)
M25625  IGHJ4*03                177         85,42% (41/48 nt)
J00256  IGHJ5*01                174         82,35% (42/51 nt)
X86355  IGHJ5*02                174         82,35% (42/51 nt)
```

Alignment

```
seq1                            gacagctttctttaacgcctatgacttctggggccagggaaccctggtcaccgtctcctcagcctccaccaagggcccatcggtcttccc
X86355  IGHJ4*02                ..............--ta--t-----a-------------------------------------
J00256  IGHJ4*01                ..............--ta--t-----a---------a-----------------------
M25625  IGHJ4*03                ..............g-ta--t-----a---------a--g---------------------
J00256  IGHJ5*01                ...........ac---tgg-tc----c---------a-----------------------
X86355  IGHJ5*02                ...........ac---tgg-tc---cc-----------------------------------
```

FIGURE 6. IMGT/V-QUEST "Detailed view" result page: "Alignment for V-GENE and allele identification" and "Alignment for J -GENE and allele identification." If selected in the Search page, the alignments for the V- and J-GENE and allele identification are displayed with the alignment score, the percentage of identity for the five closest genes and alleles and the ratio number of identical nucleotides/number of aligned nucleotides. Dashes indicate identical nucleotides. Dots indicate gaps. FR-IMGT and CDR-IMGT are delimited according to the IMGT unique numbering (Lefranc 1997, 1999; Lefranc et al. 2003). The seq1 accession number is AB012909 from the IMGT/LIGM-DB database (Giudicelli et al. 2006).



4. Results of IMGT/JunctionAnalysis

Maximum number of accepted mutations in: 3'V-REGION = 2, D-REGION = 4, 5'J-REGION = 2
Maximum number of accepted D-GENE = 1

Analysis of the JUNCTION

D-REGION is in reading frame 1.

Click on mutated (underlined) nucleotide to see the original one:

Input	V name	3'V-REGION	N1	D-REGION	N2	5'J-REGION	J name	D name	Vmut	Dmut	Jmut	Ngc
seq1	IGHV3-30*04	tgtgcgag...	g	.....acagctttctttta.	acgc	....ctatgacttctgg	IGHJ4*02	IGHD5-5*01	0	3	2	4/5

Eligible D genes:

D name	D length	tgtgcgaggacagctttctttaacgcctatgacttctgg	Score#	Mutation#	Location
IGHD1-1*01	17	---a--	5	1	d[4-9],s[10-15]
IGHD2-2*01	31	-----	5	0	d[19-23],s[11-15]
IGHD2-2*02	31	-----	5	0	d[19-23],s[11-15]
IGHD2-2*03	31	-----	5	0	d[19-23],s[11-15]
IGHD3-3*01	31	--t--	4	1	d[10-14],s[16-20]
IGHD3-3*02	31	---a--t--	7	2	d[6-14],s[12-20]
IGHD4-4*01	16	----	4	0	d[10-13],s[21-24]
IGHD5-5*01	20	------a-gg---	10	3	d[7-19],s[10-22]
IGHD6-6*01	18	-----	5	0	d[9-13],s[11-15]
IGHD1-7*01	17	----	4	0	d[5-8],s[21-24]
IGHD2-8*01	31	--a--	4	1	d[25-29],s[18-22]
IGHD2-8*02	31	--a--	4	1	d[25-29],s[18-22]
IGHD3-9*01	31	---g--	5	1	d[13-18],s[19-24]
IGHD3-10*01	31	----	4	0	d[28-31],s[21-24]
IGHD3-10*02	30	--a-g--	5	2	d[7-13],s[14-20]
IGHD4-11*01	16	----	4	0	d[10-13],s[21-24]
IGHD5-12*01	23	---acga---	6	4	d[13-22],s[13-22]
IGHD6-13*01	21	-----	5	0	d[12-16],s[11-15]
IGHD1-14*01	17	----	4	0	d[5-8],s[21-24]
IGHD2-15*01	31	----	4	0	d[20-23],s[12-15]
IGHD3-16*01	37	--t-a-gc---	7	4	d[23-33],s[12-22]
IGHD3-16*02	37	--t-a--g---	8	3	d[23-33],s[12-22]
IGHD5-18*01	20	------a-gg---	10	3	d[7-19],s[10-22]
IGHD1-20*01	17	----	4	0	d[5-8],s[21-24]
IGHD2-21*01	28	--g--	4	1	d[19-23],s[15-19]
IGHD2-21*02	28	---t---	6	1	d[17-23],s[9-15]
IGHD3-22*01	31	--a--	4	1	d[22-26],s[16-20]
IGHD4-23*01	19	----	4	0	d[13-16],s[21-24]
IGHD5-24*01	20	---acaa---	6	4	d[10-19],s[13-22]
IGHD6-25*01	18	----	4	0	d[9-12],s[11-14]
IGHD1-26*01	20	----	4	0	d[12-15],s[12-15]
IGHD7-27*01	11	----	4	0	d[2-5],s[21-24]

Translation of the JUNCTION

Click on mutated (underlined) amino acid to see the original one:

	104	105	106	107	108	109	110	113	114	115	116	117	118	Frame	CDR3-IMGT length	Molecular mass	pI
	C	A	R	T	A	F	F	N	A	Y	D	F	W				
seq1	tgt	gcg	agg	aca	gct	ttc	ttt	aac	gcc	tat	gac	ttc	tgg	+	11	1,611.8	6.44

Be aware that some allele reference sequences may be incomplete or from cDNAs. In those cases, IMGT/JunctionAnalysis uses automatically the allele *01 for the analysis of the JUNCTION.

5. Sequence of the JUNCTION ('nt' and 'AA')

```
104 105 106 107 108 109 110 113 114 115 116 117 118
 C   A   R   T   A   F   F   N   A   Y   D   F   W
tgt gcg agg aca gct ttc ttt aac gcc tat gac ttc tgg
```

Input for IMGT/JunctionAnalysis

```
>seq1,IGHV3-30*04,IGHJ4*02
tgtgcgaggacagctttctttaacgcctatgacttctgg
```

FIGURE 7. IMGT/V-QUEST "Detailed view" result page: "Results of IMGT/JunctionAnalysis" and "Sequence of the JUNCTION ("nt" and "AA")." The "Results of IMGT/JunctionAnalysis" (Giudicelli and Lefranc 2011) comprise the detailed "Analysis of the JUNCTION" (Yousfi Monod et al. 2004) with the D-REGION reading frame, "Eligible D genes" if selected (D genes and alleles that match at least four or more nucleotides with the junction, allowing the user to visualize the result among other close solutions), and "Translation of the JUNCTION" with AA colored according to the 11 IMGT physicochemical classes (Pommié et al. 2004). The "Sequence of the JUNCTION ("nt" and "AA")" provides the JUNCTION in nucleotides and amino acids according to the IMGT numbering and in the FASTA format required as input by IMGT/JunctionAnalysis online. The seq1 accession number is AB012909 from the IMGT/LIGM-DB database (Giudicelli et al. 2006).

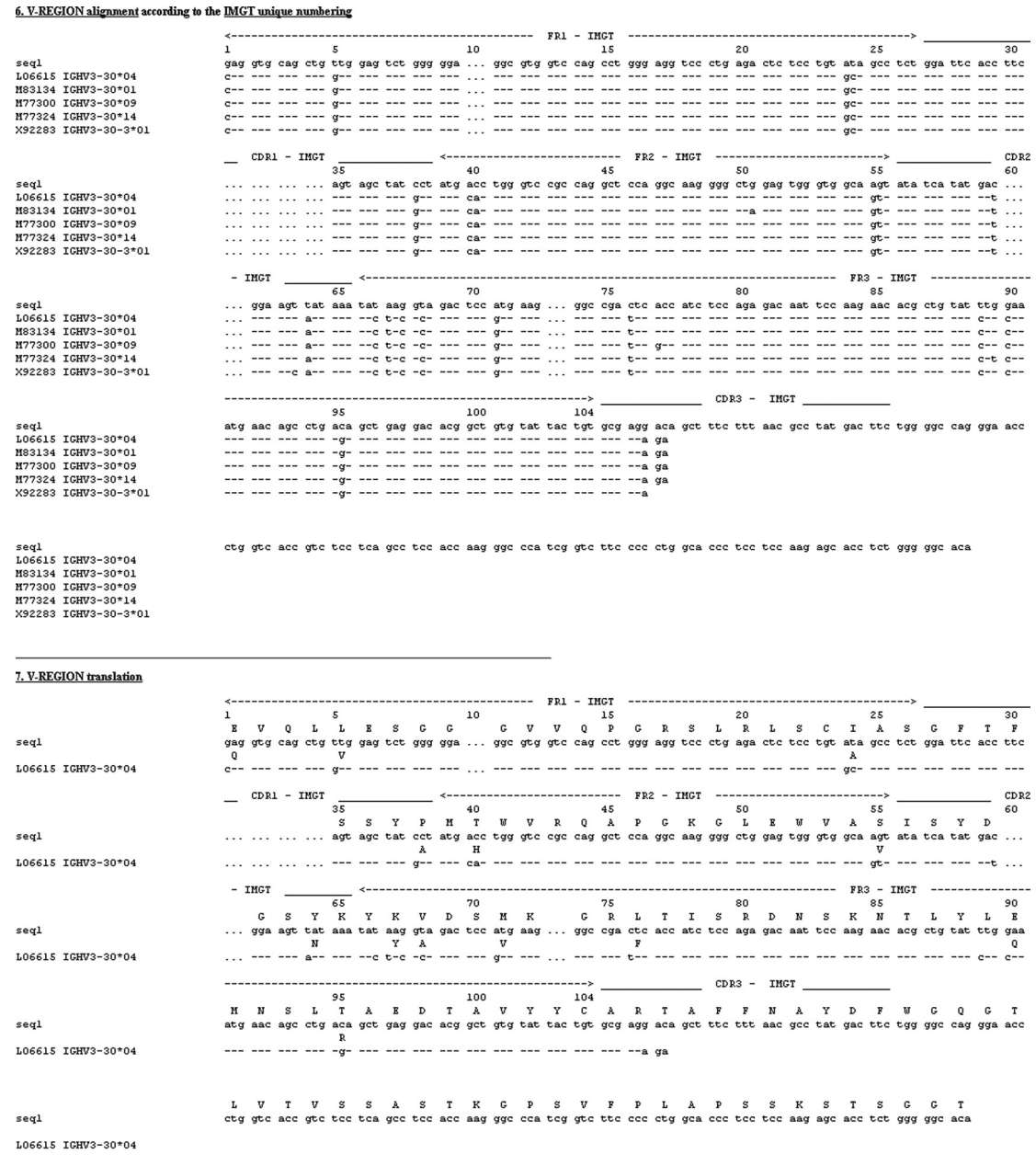


FIGURE 8. IMGT/V-QUEST "Detailed view" result page: three displays of the V-REGION. "V-REGION alignment according to the IMGT unique numbering" displays the nucleotide sequences with the FR-IMGT and CDR-IMGT delimitations according to the IMGT unique numbering (Lefranc 1997, 1999; Lefranc et al. 2003). The "V-REGION translation" displays the nucleotide sequence and deduced amino acid translation of the input sequence, aligned with the closest germline V-REGION, and with the FR-IMGT and CDR-IMGT delimitations. "V-REGION protein display" displays the deduced amino acid translation of the input sequence, aligned with the V-REGION of the closest germline V-GENE and with the FR-IMGT and CDR-IMGT delimitations, and on the third line of the alignment and shown in bold, the amino acids of the input sequences that are different from the closest germline V-REGION. The seq1 accession number is AB012909 from the IMGT/LIGM-DB database (Giudicelli et al. 2006).

9. V-REGION mutation and AA change table

FR1-IMGT	CDR1-IMGT	FR2-IMGT	CDR2-IMGT	FR3-IMGT	CDR3-IMGT
c1>g, Q1>E (+ + -) g13>t, V5>L (+ - +) g70>a, A24>I (+ - +) c71>t, A24>I (+ - +)	g112>c, A38>P (- - -)	c118>a, H40>T (+ - -) a119>c, H40>T (+ - -) g163>a, V55>S (- - -) t164>g, V55>S (- - -)	t177>c a190>t, N64>Y (- - -)	c198>t t199>a, Y67>K (- - -) c201>g, Y67>K (- - -) c203>t, A68>V (+ - +) g211>a, V71>M (+ - -) t226>c, F76>L (+ - -) c265>t c268>g, Q90>E (+ + -) g284>c, R95>T (- - -)	a318>g g319>a a320>c

10. V-REGION mutation and AA change statistics

Nucleotide (nt) mutations

IMGT labels		V-REGION	FR1-IMGT	CDR1-IMGT	FR2-IMGT	CDR2-IMGT	FR3-IMGT	CDR3-IMGT
Nb of positions including IMGT gaps (nt)		317 (320)	78	36	51	30	117	5 (8)
Nb of nucleotides		293 (296)	75	24	51	24	114	5 (8)
Nb of identical nucleotides		273	71	23	47	22	105	5
Nb of mutations		20 (23)	4	1	4	2	9	0 (3)
Mutations	Silent	3 (6)	0	0	0	1	2	0 (3)
	Nonsilent	17	4	1	4	1	7	0
Transitions	a>g	0 (1)	0	0	0	0	0	0 (1)
	g>a	3 (4)	1	0	1	0	1	0 (1)
	c>t	4	1	0	0	0	3	0
	t>c	2	0	0	0	1	1	0
Transversions	a>c	1 (2)	0	0	1	0	0	0 (1)
	c>a	1	0	0	1	0	0	0
	a>t	1	0	0	0	1	0	0
	t>a	1	0	0	0	0	1	0
	g>c	2	0	1	0	0	1	0
	c>g	3	1	0	0	0	2	0
	g>t	1	1	0	0	0	0	0
	t>g	1	0	0	1	0	0	0

Amino acid (AA) changes

IMGT labels			V-REGION	FR1-IMGT	CDR1-IMGT	FR2-IMGT	CDR2-IMGT	FR3-IMGT	CDR3-IMGT
Nb of positions including IMGT gaps (AA)			105 (106)	26	12	17	10	39	1 (2)
Nb of AA			97 (98)	25	8	17	8	38	1 (2)
Nb of identical AA			84 (85)	22	7	15	7	32	1 (2)
Nb of AA changes			13	3	1	2	1	6	0
AA changes	Very similar	(+ + +)	0	0	0	0	0	0	0
	Similar	(+ + -)	2	1	0	0	0	1	0
		(+ - +)	3	2	0	0	0	1	0
	Dissimilar	(+ - -)	3	0	0	1	0	2	0
		(- + -)	0	0	0	0	0	0	0
		(- - +)	0	0	0	0	0	0	0
	Very dissimilar	(- - -)	5	0	1	1	1	2	0

FIGURE 9. IMGT/V-QUEST "Detailed view" result page: "V-REGION mutation and AA change table," and "V-REGION mutation and AA change statistics." The "V-REGION mutation and AA change table" lists the mutations (nt and AA) of the analyzed sequence compared with the closest V-REGION allele. They are described for the V-REGION and for each FR-IMGT and CDR-IMGT, with their nucleotide and codon position according to the IMGT unique numbering (Lefranc 1997, 1999; Lefranc et al. 2003) and for the AA changes according to the IMGT AA classes (Pommié et al. 2004). For example, c1>g, Q1>E (++−) means that the nucleotide mutation (c>g) at nt 1 leads to an AA change (Q>E) at codon 1 with the same hydropathy (+) and volume (+) but with different physicochemical properties (−) classes (Pommié et al. 2004). It is the first time that such qualification of AA replacement is provided. This has led us to identify four types of AA changes: very similar (+++), similar (++−, +−+), dissimilar (−−+, −+−, +−−) and very dissimilar (−−−). The "V-REGION mutation and AA change statistics" comprises two tables, the first one for "Nucleotide (nt) mutations" described as silent/nonsilent, transitions/transversions, and the second one for "Amino acid (AA) changes" described as above. For both tables, results are given for the V-REGION and per FR-IMGT and CDR-IMGT. Statistics are calculated up to the 3′ end of the V-REGION identified in the input sequence (this includes the 3′ last two identical nucleotides with the closest germline V-REGION). The numbers in parentheses, in the V-REGION and CDR3-IMGT columns, correspond to the statistics calculated up to the 3′ end of the closest germline V-REGION, and therefore may include nucleotide and amino acid differences due to the junction diversity. The accession number of the analyzed sequence is AB012909 from the IMGT/LIGM-DB database (Giudicelli et al. 2006).

11. V-REGION mutation hot spots (motifs and localizations in germline V-REGION)

(a/t)a wa		t(a/t) tw		(a/g)g(c/t)(a/t) rgyw		(a/t)(a/g)c(c/t) wrcy	
Motif	Localization	Motif	Localization	Motif	Localization	Motif	Localization
ta	105-106 (CDR1)	tt	82-83 (CDR1)	agct	8-11 (FR1)	agct	8-11 (FR1)
ta	109-110 (CDR1)	tt	88-89 (CDR1)	agta	103-106 (CDR1)	agcc	41-44 (FR1)
ta	114-115 (CDR1)	ta	105-106 (CDR1)	ggct	132-135 (FR2)	agcc	72-75 (FR1)
aa	142-143 (FR2)	ta	109-110 (CDR1)	ggca	139-142 (FR2)	agct	106-109 (CDR1)
ta	165-166 (FR2-CDR2)	ta	114-115 (CDR1)	ggct	146-149 (FR2)	tgct	111-114 (CDR1)
ta	167-168 (CDR2)	tt	164-165 (FR2)	ggca	159-162 (FR2)	tact	196-199 (FR3)
ta	172-173 (CDR2)	ta	167-168 (CDR2)	agta	187-190 (CDR2)	agcc	277-280 (FR3)
aa	186-187 (CDR2)	ta	172-173 (CDR2)	agct	285-288 (FR3)	agct	285-288 (FR3)
ta	189-190 (CDR2)	ta	189-190 (CDR2)	ggct	297-300 (FR3)	tact	307-310 (FR3)
ta	192-193 (CDR2)	ta	192-193 (CDR2)				
aa	194-195 (CDR2)	ta	196-197 (FR3)				
ta	196-197 (FR3)	ta	199-200 (FR3)				
ta	199-200 (FR3)	tt	226-227 (FR3)				
aa	214-215 (FR3)	tt	246-247 (FR3)				
aa	244-245 (FR3)	ta	262-263 (FR3)				
aa	250-251 (FR3)	ta	304-305 (FR3)				
aa	253-254 (FR3)	tt	306-307 (FR3)				
ta	262-263 (FR3)						
aa	269-270 (FR3)						
aa	274-275 (FR3)						
ta	304-305 (FR3)						
ta	307-308 (FR3)						

FIGURE 10. IMGT/V-QUEST "Detailed view" result page: "V-REGION mutation hot spots (motifs and localization in germline V-REGION)." This table provides the hot spot motifs identified in the closest germline V-REGION and their localization. The hot spot patterns are (a/t)a and (a/g)g(c/t)(a/t), and their complementary reverse motifs t(a/t) and (a/t)(a/g)c(c/t). The accession number of the analyzed sequence is AB012909 from the IMGT/LIGM-DB database (Giudicelli et al. 2006).

12. 'seq1' V-REGION and V-D-J-REGION

V-REGION nucleotide sequence in FASTA format

```
>seq1
gaggtgcagctgttggagtctgggggga...ggcgtggtccagcctgggaggtccctgaga
ctctcctgtatagcctctggattcaccttc............agtagctatcctatgacc
tgggtccgccaggctccaggcaaggggctggagtgggtggcaagtatatcatatgac...
...ggaagttataaatataaggtagactccatgaag...ggccgactcaccatctccaga
gacaattccaagaacacgctgtatttggaaatgaacagcctgacagctgaggacacggct
gtgtattactgtgcgag
```

Analyse this sequence with IMGT/PhyloGene

V-REGION amino acid sequence in FASTA format

```
>seq1
EVQLLESGG.GVVQPGRSLRLSCIASGFTF....SSYPMTWVRQAPGKGLEWVASISYD.
.GSYKYKVDSMK.GRLTISRDNSKNTLYLEMNSLTAEDTAVYYCA
```

V-REGION amino acid sequence on one line

```
>seq1
EVQLLESGG.GVVQPGRSLRLSCIASGFTF....SSYPMTWVRQAPGKGLEWVASISYD..GSYKYKVDSMK.GRLTISRDNSKNTLYLEMNSLTAEDTAVYYCA
```

V-D-J-REGION nucleotide sequence in FASTA format

```
>seq1
gaggtgcagctgttggagtctgggggga...ggcgtggtccagcctgggaggtccctgaga
ctctcctgtatagcctctggattcaccttc............agtagctatcctatgacc
tgggtccgccaggctccaggcaaggggctggagtgggtggcaagtatatcatatgac...
...ggaagttataaatataaggtagactccatgaag...ggccgactcaccatctccaga
gacaattccaagaacacgctgtatttggaaatgaacagcctgacagctgaggacacggct
gtgtattactgtgcgaggacagctttctttaacgcctatgacttctggggccagggaacc
ctggtcaccgtctcctcag
```

V-D-J-REGION amino acid sequence in FASTA format

```
>seq1
EVQLLESGG.GVVQPGRSLRLSCIASGFTF....SSYPMTWVRQAPGKGLEWVASISYD.
.GSYKYKVDSMK.GRLTISRDNSKNTLYLEMNSLTAEDTAVYYCARTAFFNAYDFWGQGT
LVTVSS
```

V-D-J-REGION amino acid sequence on one line

```
EVQLLESGG.GVVQPGRSLRLSCIASGFTF....SSYPMTWVRQAPGKGLEWVASISYD..GSYKYKVDSMK.GRLTISRDNSKNTLYLEMNSLTAEDTAVYYCARTAFFNAYDFWGQGTLVTVSS
```

FIGURE 11. IMGT/V-QUEST "Detailed view" result page: sequences of V-REGION and V-J- or V-D-J-REGION. This section provides nucleotide and amino acid sequences with gaps according to the IMGT unique numbering (Lefranc 1997, 1999; Lefranc et al. 2003) and includes the V-REGION nucleotide sequence in FASTA format with access to the IMGT/PhyloGene tool (Elemento and Lefranc 2003), "V-REGION amino acid on one line," V-J or V-D-J-REGION nucleotide and amino acid sequences in FASTA format, V-J, or V-D-J-REGION AA sequence on one line.

13. Annotation by IMGT/Automat

```
       Label            Location/Qualifiers
V-D-J-REGION            1..355
                        /CDR_length="[8.8.11]"
                        /FR_length="[25.17.38.11]"
                        /nucleotide sequence
                        gaggtgcagctgttggagtctgggggaggcgtggtccagcctgggaggtccctgagactc
                        tcctgtatagcctctggattcaccttcagtagctatcctatgacctgggtccgccaggct
                        ccaggcaaggggctggagtgggtggcaagtatatcatatgacggaagttataaatataag
                        gtagactccatgaagggccgactcaccatctccagagacaattccaagaacacgctgtat
                        ttggaaatgaacagcctgacagctgaggacacggctgtgtattactgtgcgaggacagct
                        ttctttaacgcctatgacttctggggccagggaaccctggtcaccgtctcctcag
                        /translation
                        EVQLLESGGGVVQPGRSLRLSCIASGFTFSSYPMTWVRQAPGKGLEWVASISYDGSYKYK
                        VDSMKGRLTISRDNSKNTLYLEMNSLTAEDTAVYYCARTAFFNAYDFWGQGTLVTVSS
V-REGION                1..293
                        /allele="IGHV3-30*04"
                        /gene="IGHV3-30"
                        /nucleotide sequence
                        gaggtgcagctgttggagtctgggggaggcgtggtccagcctgggaggtccctgagactc
                        tcctgtatagcctctggattcaccttcagtagctatcctatgacctgggtccgccaggct
                        ccaggcaaggggctggagtgggtggcaagtatatcatatgacggaagttataaatataag
                        gtagactccatgaagggccgactcaccatctccagagacaattccaagaacacgctgtat
                        ttggaaatgaacagcctgacagctgaggacacggctgtgtattactgtgcgag
                        /translation
                        EVQLLESGGGVVQPGRSLRLSCIASGFTFSSYPMTWVRQAPGKGLEWVASISYDGSYKYK
                        VDSMKGRLTISRDNSKNTLYLEMNSLTAEDTAVYYCA
FR1-IMGT                1..75
                        /AA_IMGT="AA 1 to 26, AA 10 is missing"
                        /nucleotide sequence
                        gaggtgcagctgttggagtctgggggaggcgtggtccagcctgggaggtccctgagactc
                        tcctgtatagcctct
                        /translation
                        EVQLLESGGGVVQPGRSLRLSCIAS
1st-CYS                 64..66
                        /nucleotide sequence
                        tgt
                        /translation
                        C
CDR1-IMGT               76..99
                        /AA_IMGT="AA 27 to 38, AA 31, 32, 33, 34 are  missing"
                        /nucleotide sequence
                        ggattcaccttcagtagctatcct
                        /translation
                        GFTFSSYP
FR2-IMGT                100..150
                        /AA_IMGT="AA 39 to 55"
                        /nucleotide sequence
                        atgacctgggtccgccaggctccaggcaaggggctggagtgggtggcaagt
                        /translation
                        MTWVRQAPGKGLEWVAS
CONSERVED-TRP           106..108
                        /nucleotide sequence
                        tgg
                        /translation
                        W
CDR2-IMGT               151..174
                        /AA_IMGT="AA 56 to 65, AA 60, 61 are  missing"
                        /nucleotide sequence
                        atatcatatgacggaagttataaa
                        /translation
                        ISYDGSYK
```

FIGURE 12. IMGT/V-QUEST "Detailed view" result page: "Annotation by IMGT/Automat": the results of the analysis of the input sequence by IMGT/Automat (Giudicelli et al. 2003, 2005b) provide a full automatic annotation for the V-J-REGION or V-D-J-REGION using IMGT standardized labels.

IMGT/V-QUEST Excel File

21. Select the option "C. Excel file."

22. Select, in the light orange box, the displays of the results to be shown.

There are 11 sheets proposed for the result display. The "Summary" and "Parameters" sheets are always provided by default.

- *Summary*
- *IMGT-gapped-nt-sequences*

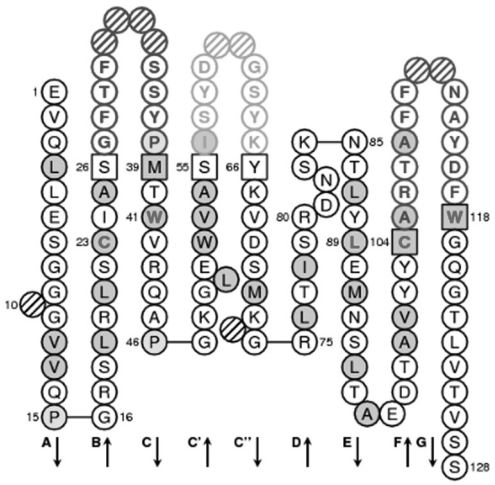


FIGURE 13. IMGT/V-QUEST "Detailed view" result page: IMGT Collier de Perles. This section recalls the identified region (here, V-D-J-REGION) and the CDR-IMGT lengths and FR-IMGT lengths. Depending on the choice selected in option 14 of the IMGT/V-QUEST Search page (see Fig. 2), the "IMGT Collier de Perles" (Ruiz and Lefranc 2002; Kaas et al. 2004, 2007; Kaas and Lefranc 2007) can be displayed either as a "link to IMGT/Collier de Perles tool" **IMGT/Collier de Perles: IMGT Standardized Representation of Domains (IG, TR, and IgSF Variable and Constant Domains, MH and MhSF Groove Domains)** (Ehrenmann et al. 2011) or as a direct "IMGT Collier de Perles (for a nb of sequences <5)" representation integrated into IMGT/V-QUEST results. The IMGT Collier de Perles shown was obtained by selecting this last option. Positions in blue mean that the amino acids of the user sequence at these positions is hydrophobic (hydropathy index with positive value) or is a tryptophan (W), as in 50% or more of analyzed V domains. Positions with red and bold letters indicate the five conserved positions of a V domain: 1st-CYS 23, CONSERVED-TRP 41, hydrophobic 89, 2nd-CYS 104, and J-TRP 118. Anchor positions are in squares. Hatched positions correspond to gaps according to the IMGT unique numbering for V domain (Lefranc 1997, 1999; Lefranc et al. 2003). Prolines are shown in yellow. Arrows indicate the β strands and their direction. The CDR-IMGT lengths of this domain are (8.8.11). The CDR-IMGT are colored as follows: CDR1-IMGT (red), CDR2-IMGT (orange), and CDR3-IMGT (purple) for the IGH (this figure), TRB and TRD V-D-J-REGION, and CDR1-IMGT (blue), CDR2-IMGT (green), and CDR3-IMGT (greenblue) (data not shown) for the IGK, IGL, TRA, and TRG V-J-REGION.

- *nt-sequences*
- *IMGT-gapped-AA-sequences*
- *AA-sequences*
- *Junction*
- *V-REGION-mutation-and-AA-change-table*
- *V-REGION-nt-mutation-statistics*
- *V-REGION-AA-change-statistics*
- *V-REGION-mutation-hot spot*
- *Parameters*

23. Go to "Advanced parameters" at the bottom of the search page and modify the parameters according to the particularities of the sequences, if needed (for details, see Step 11).

24. Click on "Start" to launch the analysis.
    A small window will appear on the screen, giving the choice to "open" or "save" the Excel file.

B. Synthesis for the IMGT/V-QUEST analysed sequences

Number of analysed sequences: 7

Sequences compared with the human IG set from the IMGT reference directory

Summary table:

Sequence ID	V-GENE and allele	Functionality	V-REGION score	V-REGION identity % (nt)	J-GENE and allele	D-GENE and allele	D-REGION reading frame	CDR-IMGT lengths	AA JUNCTION	JUNCTION frame
seq1	IGHV1-69*01	Productive	1273	93,75% (270/288 nt)	IGHJ4*01 or IGHJ4*02 or IGHJ4*03	IGHD3-22*01	2	[8.8.12]	CARQYGSSGYYAYW	in-frame
seq2	IGHV3-30*03, or IGHV3-30*18	Productive	1417	99,31% (286/288 nt)	IGHJ6*02	IGHD3-16*02	2	[8.8.23]	CAKDLDDYVWGSYRHSRYYYGMDFW	in-frame
seq3	IGHV3-30*04	Productive	1255	93,06% (268/288 nt)	IGHJ4*02	IGHD5-5*01	1	[8.8.11]	CARTAFFNAYDFW	in-frame
seq4	IGHV3-30*04	Productive	1435	100,00% (288/288 nt)	IGHJ4*02	IGHD6-19*01	2	[8.8.15]	CARDRSIAVAQYYFDYW	in-frame
seq5	IGHV3-30*04	Productive	1408	98,96% (285/288 nt)	IGHJ4*02	IGHD6-13*01	2	[8.8.17]	CARDSSYPPGIAAAGVXYW	in-frame
seq6	IGHV3-30*04	Productive	1426	99,65% (287/288 nt)	IGHJ1*01	IGHD3-10*01	2	[8.8.16]	CARGRTKGSGRPGYFQHW	in-frame
seq7	IGHV3-9*01	Unproductive (stop codons)	1417	99,31% (286/288 nt)	IGHJ4*02	IGHD3-10*01	2	[8.8.17]	CAKDIGGKYGSGGSGLD*W	in-frame

Results of IMGT/JunctionAnalysis for : IGH junctions

Alignment with the closest alleles:

The analysed sequences are aligned with the closest allele (with number of aligned sequences in parentheses):

IGHV1-69*01(1) IGHV3-30*03(1) IGHV3-30*04(4) IGHV3-9*01(1)

FIGURE 14. IMGT/V-QUEST "Synthesis view" result page: *top* of the page. The number of analyzed sequences is indicated. The Summary table displays one row for each input sequence, including: (1) the name of the sequence (Sequence ID), (2) the name of the closest V-GENE and allele (a note may appear with the V-GENE and allele name when the V-REGION score is very low (<200), and/or the percentage of identity is <85% and/or, compared with that of V, when input sequence has different CDR1-IMGT and/or CDR2-IMGT lengths: the alignment for this sequence has to be checked in "A. Detailed view," using the advanced parameters "Search for insertions and deletions in V-REGION"), (3) the functionality of the sequence (when found, the presence of stop codons is indicated), (4) the V-REGION score, (5) the V-REGION percentage of identity and ratio number of identical nucleotides (nt)/number of aligned nt, (6) the name of the closest J-GENE and allele (notes may appear with the J-GENE and allele name to indicate that other possibilities exist for the choice of the J-GENE and allele name), and (7) provided according to the IMGT/JunctionAnalysis results, the D-GENE and allele name, the D reading frame, the CDR-IMGT lengths, the AA JUNCTION, and the JUNCTION frame. In the absence of results of IMGT/JunctionAnalysis, only the AA JUNCTION defined by IMGT/V-QUEST is displayed. The seq1, seq2, seq3, seq4, seq5, seq6, and seq7 correspond to the DQ100777, AB021524, AB012909, AB021511, AB021516, AB021514, and AB021539 accession numbers, respectively, from the IMGT/LIGM-DB database (Giudicelli et al. 2006).

25. To see the results, click on "open."

The Excel file comprises 11 sheets. These can be viewed in the supplementary material (Results_Excel.doc).

DISCUSSION

IMGT/V-QUEST (http://www.imgt.org) provides a highly detailed and standardized analysis of the IG and TR rearranged sequences: it describes the main characteristics of user sequences according to the axioms and concepts of IMGT-ONTOLOGY (Giudicelli et al. 1999; Duroux et al. 2008). IMGT/V-QUEST is constantly updated based on the results of biological and clinical research. Improvements can be followed through the program version number and the reference directory release number indicated in the search pages and the list of the corresponding changes found in the corresponding documentations. New alleles annotated in IMGT are integrated in the IMGT reference directory as soon as they are confirmed experimentally and are publicly available in the generalist databases. IMGT/HighV-QUEST (Alamyar et al. 2010, 2011), the high-throughput version of IMGT/V-QUEST for Next Generation Sequencing (NGS) that analyzes up to 150,000 sequences per run, uses the IMGT/V-QUEST program and standards, and provides results with the same degree of resolution and high quality (results

Analysis of the JUNCTIONs

Click on mutated (underlined) nucleotide to see the original one:

Input	V name	3'V-REGION	P	N1	P	D-REGION	N2	P	5'J-REGION	J name	D name	Vmut	Dmut	Jmut	Ngc
seq1	IGHV1-69*01	tgtgcgaga..		caa		.....tatggtagtagtggttattac...	gc		..........ctactgg	IGHJ4*01	IGHD3-22*01	0	1	0	3/5
seq2	IGHV3-30*03	tgtgcgaaaga	tct	gg		.....atgattacgtttgggggagttatcgtcat...	tctagg		.....tactactacggtatggacttctgg	IGHJ6*02	IGHD3-16*02	1	1	1	5/8
seq3	IGHV3-30*04	tgtgcgag...		g		......acagctttcttta.	acgc		....ctatgacttctgg	IGHJ4*02	IGHD5-5*01	0	3	2	4/5
seq4	IGHV3-30*04	tgtgcgagaga	tc	gga		..gtatagcagtggc.....	cc	agt	actactttgactactgg	IGHJ4*02	IGHD6-19*01	0	0	0	4/5
seq5	IGHV3-30*04	tgtgcgagaga	t	tcctcctaccc ccc		gggtatagcagcagctgg...	ag		......ttgantactgg	IGHJ4*02	IGHD6-13*01	0	0	1	8/13
seq6	IGHV3-30*04	tgtgcgagag.		gaaggacgaag		..........ggttcggggag..........	acccggc		......tacttccagcactgg	IGHJ1*01	IGHD3-10*01	0	0	0	12/18
seq7	IGHV3-9*01	tgtgcaaaagata	t	tgggggcaaa		.....tatggttcgggg.....	ggctcgggac		......ttgactaatgg	IGHJ4*02	IGHD3-10*01	0	0	1	14/20

Translation of the JUNCTIONs

Click on mutated (underlined) amino acid to see the original one:

	104	105	106	107	108	109	110	111	111.1	111.2	111.3	111.4	111.5	112.5	112.4	112.3	112.2	112.1	112	113	114	115	116	117	118	Frame	CDR3-IMGT length	Molecular mass	pI
	C	A	R	Q	Y	G	S												S	G	Y	Y	A	Y	W				
#1 seq1	tgt	gcg	aga	caa	tat	ggt	agt	...	...	...	...	...	...	...	...	...	...	...	agt	ggt	tat	tac	gcc	tac	tgg	+	12	1,674.81	8.89
	C	A	K	D	L	D	D	Y	V	W	G	S	Y	R	H	S	R	Y	Y	Y	G	M	D	F	W				
#2 seq2	tgt	gcg	aaa	gat	ctg	gat	gat	tac	gtt	tgg	ggg	agt	tat	cgt	cat	tct	agg	tac	tac	tac	ggt	atg	gac	ttc	tgg	+	23	3,197.52	5.78
	C	A	R	T	A	F	F													N	A	Y	D	F	W				
#3 seq3	tgt	gcg	agg	aca	gct	ttc	ttt	...	...	...	...	...	...	...	...	...	...	...	...	aac	gcc	tat	gac	ttc	tgg	+	11	1,611.8	6.44
	C	A	R	D	R	S	I	A	V									A	Q	Y	Y	F	D	Y	W				
#4 seq4	tgt	gcg	aga	gat	cgg	agt	ata	gca	gtg	...	...	...	...	...	...	...	...	gcc	cag	tac	tac	ttt	gac	tac	tgg	+	15	2,127.37	6.58
	C	A	R	D	S	S	Y	P	P	G							I	A	A	A	G	V	X	Y	W				
#5 seq5	tgt	gcg	aga	gat	tcc	tcc	tac	ccc	ccg	ggt	...	...	...	...	...	...	ata	gca	gca	gct	gga	gtt	gan	tac	tgg	+	17	1,884.1	6.44
	C	A	R	G	R	T	K	G	S								G	R	P	G	Y	F	Q	H	W				
#6 seq6	tgt	gcg	aga	gga	agg	acg	aag	ggt	tcg	...	...	...	...	...	...	...	ggg	aga	ccc	ggc	tac	ttc	cag	cac	tgg	+	16	2,064.32	11.72
	C	A	K	D	I	G	G	K	Y	G							S	G	G	S	G	L	D	*	W				
#7 seq7	tgt	gca	aaa	gat	att	ggg	ggc	aaa	tat	ggt	...	...	...	...	...	...	tcg	ggg	ggc	tcg	gga	ctt	gac	taa	tgg	+	17	1,770.94	6.58

FIGURE 15. IMGT/V-QUEST "Synthesis view" result page: results of IMGT/JunctionAnalysis. This section displays the "Analysis of the JUNCTIONs" and the "Translation of the JUNCTIONs" per locus, as provided by the IMGT/JunctionAnalysis tool for all sequences identified by IMGT/V-QUEST as belonging to the same locus (IGH, IGK, or IGL locus for IG sequences, TRA, TRB, TRG, or TRD locus for TR sequences) (Yousfi Monod et al. 2004; Brochet et al. 2008; Giudicelli and Lefranc 2011).

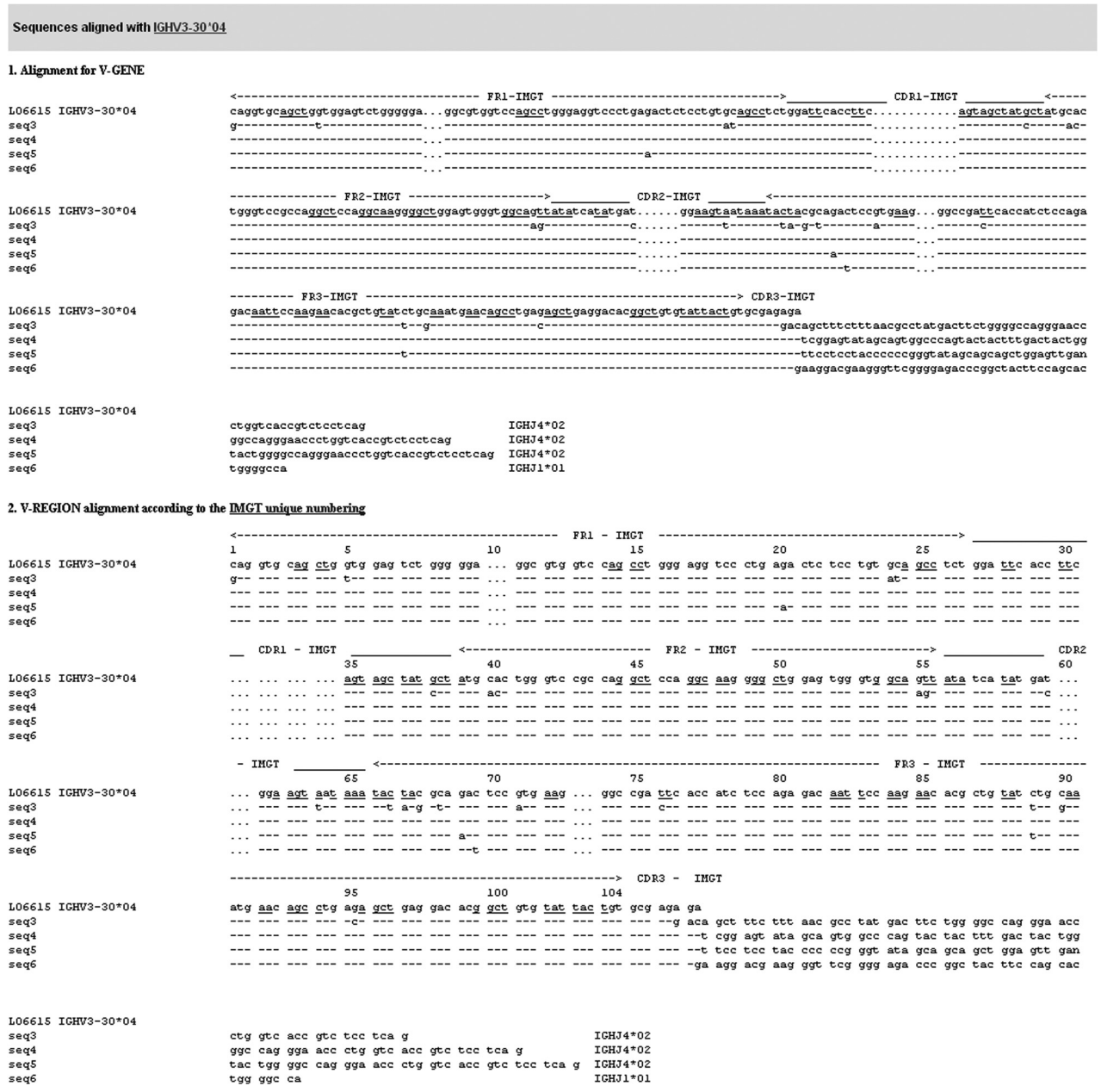


FIGURE 16. IMGT/V-QUEST "Synthesis view" result page: "Alignment for V-GENE" and "V-REGION alignment according to the IMGT unique numbering." The sequences are displayed with the delimitations of FR-IMGT and CDR-IMGT. Dashes indicate identical nucleotides. Dots indicate gaps. FR-IMGT and CDR-IMGT are delimited according to the IMGT unique numbering (Lefranc 1997, 1999; Lefranc et al. 2003). Displayed sequences are those of the input set, which have been identified as using the same closest germline gene and allele (here IGHV3-30*04). The sequence of the V-REGION of IGHV3-30*04 is shown on the first line and the input sequences are aligned, with nucleotide mutations indicated by comparison with IGHV3-30*04. Underlined nucleotides in IGHV3-30*04 correspond to mutation hot spot motif localizations.

similar to those of the Excel file, formatted in comma separated value [CSV]). Rules for sequence analysis are improved through scientific collaboration and with the constant feedback of scientists. IMGT/V-QUEST has been recommended by the European Research Initiative on chronic lymphocytic leukemia CLL (ERIC) for the analysis of the IGHV gene mutational status in CLL (Ghia et al. 2007). The level of mutations in IGHV genes of CLL patients is a prognostic criterion. Moreover, >20% of patients with

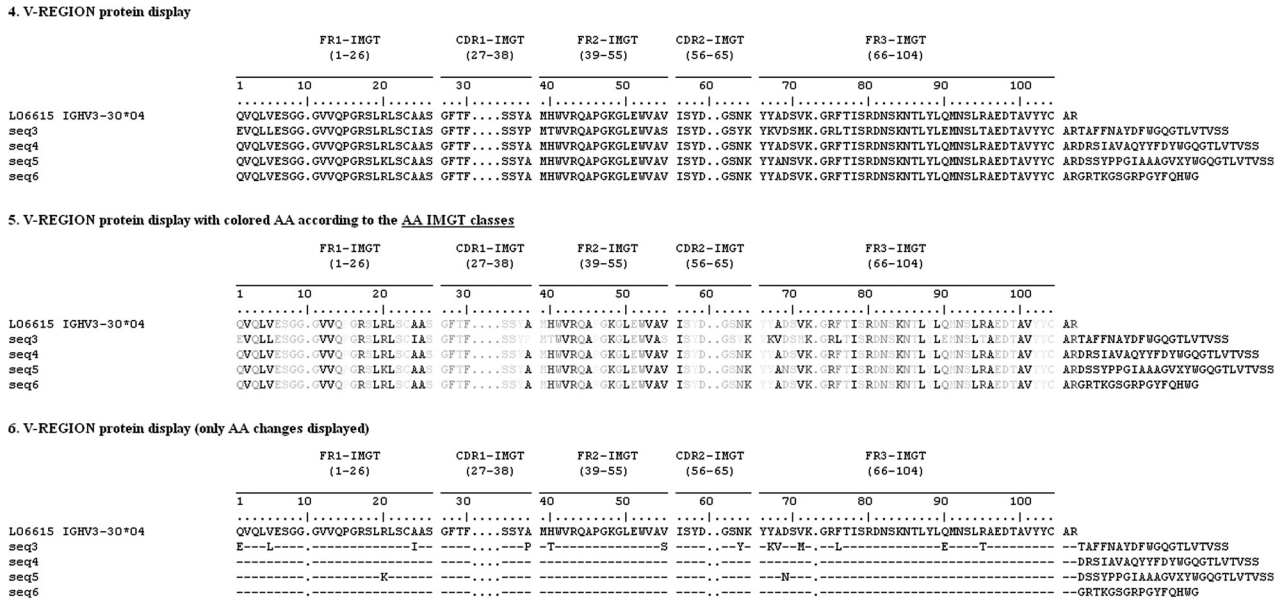


FIGURE 17. IMGT/V-QUEST "Synthesis view" result page: V-REGION protein displays. Three different displays are shown. "V-REGION protein display," "V-REGION protein display with colored AA according to the AA IMGT classes," and "V-REGION protein display (only AA changes displayed)." The first line shows the AA sequence of the closest V-REGION allele against which are aligned the translations of the input sequences identified by IMGT/V-QUEST as using the same gene and allele (here IGHV3-30*04).

CLL carry stereotyped receptors with pathogenetic implications and clinical correlations (Stamatopoulos et al. 2007). Results obtained by the users have important implications in basic and medical research, antibody and TR specificity, and in biotechnology related to antibody engineering (Giudicelli and Lefranc 2008; Lefranc 2009; Ehrenmann et al. 2010). IMGT standards are approved by the World Health Organization-International Union of Immunological Societies (WHO-IUIS) subcommittee for IG and TR (Lefranc 2007, 2008).

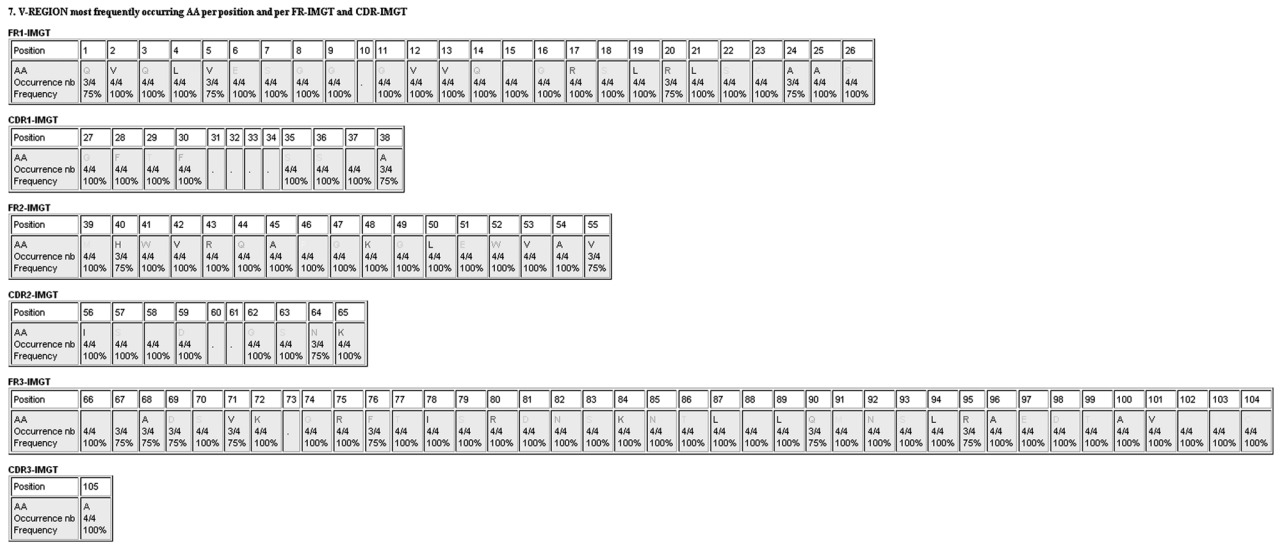


FIGURE 18. IMGT/V-QUEST "Synthesis view" result page: "V-REGION most frequently occurring AA per position and per FR-IMGT and CDR-IMGT." This section shows, for each FR-IMGT and CDR-IMGT, and for each codon (or AA), the position of the most frequently occurring AA.

ACKNOWLEDGMENTS

We thank François Ehrenmann and Patrice Duroux for checking the protocol. We are grateful to Gérard Lefranc and to the IMGT team for their motivation and expertise.

REFERENCES

Alamyar E, Giudicelli V, Duroux P, Lefranc M-P, 2010. IMGT/HighV-QUEST: A high-throughput system and Web portal for the analysis of rearranged nucleotide sequences of antigen receptors - High-throughput version of IMGT/V-QUEST. JOBIM 2010 Poster 60. http://www.jobim2010.fr/?q=fr/node/55.

Alamyar E, Duroux D, Lefranc M-P, Giudicelli V. 2011. IMGT tools for the nucleotide analysis of immunoglobulin (IG) and T cell receptor (TR) V-(D)-J repertoires, polymorphisms, and IG mutations: IMGT/V-QUEST and IMGT/HighV-QUEST for NGS. *Methods Mol Biol* (in press).

Brochet X, Lefranc M-P, Giudicelli V. 2008. IMGT/V-QUEST: The highly customized and integrated system for IG and TR standardized V-J and V-D-J sequence analysis. *Nucl Acids Res* **36:** W503–W508.

Duroux P, Kaas Q, Brochet X, Lane J, Ginestoux C, Lefranc M-P, Giudicelli V. 2008. IMGT-Kaleidoscope, the formal IMGT-ONTOLOGY paradigm. *Biochimie* **90:** 570–583.

Ehrenmann F, Lefranc M-P. 2011a. IMGT/DomainGapAlign: IMGT standardized analysis of amino acid sequences of variable, constant and groove domains (IG, TR, MH, IgSF, MhSF). *Cold Spring Harb Protoc* doi: 10.1101/pdb.prot5636.

Ehrenmann F, Lefranc M-P. 2011b. IMGT/3Dstructure-DB: Querying the IMGT database for 3D structures in immunology and immunoinformatics (IG or antibodies, TR, MH, RPI, and FPIA). *Cold Spring Harb Protoc* doi: 10.1101/pdb.prot5637.

Ehrenmann F, Duroux P, Giudicelli V, Lefranc M-P. 2010. Standardized sequence and structure analysis of antibody using IMGT. In *Antibody engineering* (ed. R Kontermann, S Dübel), 2nd ed, **Vol 2**, pp. 11–31. Springer-Verlag, Berlin/Heidelberg, Germany.

Ehrenmann F, Giudicelli V, Brochet X, Lefranc M-P. 2011. IMGT/Collier de Perles: IMGT standardized representation of domains (IG, TR, and IgSF variable and constant domains, MH and MhSF groove domains). *Cold Spring Harb Protoc* doi: 10.1101/pdb.prot5635.

Elemento O, Lefranc M-P. 2003. IMGT/PhyloGene: An on-line tool for comparative analysis of immunoglobulin and T cell receptor genes. *Dev Comp Immunol* **27:** 763–779.

Ghia P, Stamatopoulos K, Belessi C, Moreno C, Stilgenbauer S, Stevenson F, Davi F, Rosenquist R. 2007. ERIC recommendations on IGHV gene mutational status analysis in chronic lymphocytic leukemia. *Leukemia* **21:** 1–3.

Giudicelli V, Lefranc M-P. 1999. Ontology for immunogenetics: IMGT-ONTOLOGY. *Bioinformatics* **15:** 1047–1054.

Giudicelli V, Lefranc M-P. 2005. Interactive IMGT on-line tools for the analysis of immunoglobulin and T cell receptor repertoires. In *New research on immunology* (ed. BA Veskler), pp. 77–105. Nova Science, Hauppauge, NY.

Giudicelli V, Lefranc M-P. 2008. IMGT standardized analysis of immunoglobulin rearranged sequences. In *Immunoglobulin gene analysis in Chronic Lymphocytic Leukemia* (ed. P Ghia, R Rosenquist, F Davi), pp. 33–52. Wolters Kluwer Health Italy Ltd, Milan, Italy.

Giudicelli V, Lefranc M-P. 2011. IMGT/JunctionAnalysis: IMGT standardized analysis of the V-J and V-D-J junctions of the rearranged immunoglobulins (IG) and T cell receptors (TR). *Cold Spring Harb Protoc* doi: 10.1101/pdb.prot5634.

Giudicelli V, Protat C, Lefranc M-P. 2003. The IMGT strategy for the automatic annotation of IG and TR cDNA sequences: IMGT/Automat. In *Proceedings of the european conference on computational biology*, pp. 103–104. Paris, France.

Giudicelli V, Chaume D, Lefranc M-P. 2005a. IMGT/GENE-DB: A comprehensive database for human and mouse immunoglobulin and T cell receptor genes. *Nucleic Acids Res* **33:** D256–D261.

Giudicelli V, Chaume D, Jabado-Michaloud J, Lefranc M-P. 2005b. Immunogenetics sequence annotation: The strategy of IMGT based on IMGT-ONTOLOGY. *Stud Health Technol Inform* **116:** 3–8.

Giudicelli V, Duroux P, Ginestoux C, Folch G, Jabado-Michaloud J, Chaume D, Lefranc M-P. 2006. IMGT/LIGM-DB, the IMGT comprehensive database of immunoglobulin and T cell receptor nucleotide sequences. *Nucl Acids Res* **34:** D781–D784.

Kaas Q, Lefranc M-P. 2007. IMGT Colliers de Perles: Standardized sequence-structure representations of the IgSF and MhcSF superfamily domains. *Current Bioinformatics* **2:** 21–30.

Kaas Q, Ruiz M, Lefranc M-P. 2004. IMGT/3Dstructure-DB and IMGT/StructuralQuery, a database and a tool for immunoglobulin, T cell receptor and MHC structural data. *Nucl Acids Res* **32:** D208–D210.

Kaas Q, Ehrenmann F, Lefranc M-P. 2007. IG, TR and IgSf, MHC and MhcSF: What do we learn from the IMGT Colliers de Perles? *Brief. Funct Genomic Proteomic* **6:** 253–264.

Lefranc M-P. 1997. Unique database numbering system for immunogenetic analysis. *Immunol Today* **18:** 509. doi: 10.1016/S0167-5699(97)01163-8.

Lefranc M-P. 1999. The IMGT unique numbering for Immunoglobulins, T cell receptors and Ig-like domains. *The Immunologist* **7:** 132–136.

Lefranc M-P. 2000a. Nomenclature of the human immunoglobulin genes. In *Current protocols in immunology* (ed. JE Coligan, BE Bierer, DE Margulies, EM Shevach, W Strober), pp. A.1P.1–A.1P.37. John Wiley and Sons, Inc, Hoboken, NJ.

Lefranc M-P. 2000b. Nomenclature of the human T cell receptor genes. In *Current protocols in immunology* (ed. JE Coligan, BE Bierer, DE Margulies, EM Shevach, W Strober), pp. A.1O.1–A.1O.23. John Wiley and Sons, Inc, Hoboken, NJ.

Lefranc M-P. 2004. IMGT, the international ImMunoGenetics information system. In *Antibody engineering methods and protocols* (ed. BKC Lo), 2nd ed, **Vol 248**, pp. 27–49. Methods in Molecular Biology, Humana Press, Totowa, NJ.

Lefranc M-P. 2007. WHO-IUIS Nomenclature Subcommittee for Immunoglobulins and T cell receptors report. *Immunogenetics* **59:** 899–902.

Lefranc M-P. 2008. WHO-IUIS Nomenclature Subcommittee for Immunoglobulins and T cell receptors report August 2007, 13th International Congress of Immunology, Rio de Janeiro, Brazil. *Dev Comp Immunol* **32:** 461–463.

Lefranc M-P. 2009. Antibody database and tools: The IMGT experience. In *Therapeutic monoclonal antibodies: from Bench to Clinic* (ed. A Zhiqiang), pp. 91–114. John Wiley Sons, Inc, Hoboken, NJ.

Lefranc M-P. 2011a. IMGT, the International ImMunoGeneTics Information System. *Cold Spring Harb Protoc* doi: 10.1101/pdb.top115.

Lefranc M-P. 2011b. From IMGT-ONTOLOGY IDENTIFICATION axiom to IMGT standardized keywords: For immunoglobulins (IG), T cell receptors (TR), and conventional genes. *Cold Spring Harb Protoc* doi: 10.1101/pdb.ip82.

Lefranc M-P. 2011c. From IMGT-ONTOLOGY DESCRIPTION axiom to IMGT standardized labels: For immunoglobulin (IG) and T cell receptor (TR) sequences and structures. *Cold Spring Harb Protoc* doi: 10.1101/pdb.ip83.

Lefranc M-P. 2011d. From IMGT-ONTOLOGY CLASSIFICATION axiom to IMGT standardized gene and allele nomenclature: For immunoglobulins (IG) and T cell receptors (TR). *Cold Spring Harb Protoc* doi: 10.1101/pdb.ip84.

Lefranc M-P. 2011e. IMGT unique numbering for the variable (V), constant (C), and groove (G) domains of IG, TR, MH, IgSF, and MhSF. *Cold Spring Harb Protoc* doi: 10.1101/pdb.ip85.

Lefranc M-P. 2011f. IMGT Collier de Perles for the Variable (V), Constant (C), and Groove (G) Domains of IG, TR, MH, IgSF, and MhSF. *Cold Spring Harb Protoc* doi: 10.1101/pdb.ip86.

Lefranc M-P, Lefranc G. 2001a. *The Immunoglobulin FactsBook*. pp. 1–458. Academic Press, London, UK.

Lefranc M-P, Lefranc G. 2001b. *The T cell receptor FactsBook*. pp. 1–398. Academic Press, London, UK.

Lefranc M-P, Pommié C, Ruiz M, Giudicelli V, Foulquier E, Truong L, Thouvenin-Contet V, Lefranc G. 2003. IMGT unique numbering for immunoglobulin and T cell receptor variable domains and Ig superfamily V-like domains. *Dev Comp Immunol* **27:** 55–77.

Lefranc M-P, Giudicelli V, Ginestoux C, Bosc N, Folch G, Guiraudou D, Jabado-Michaloud J, Magris S, Scaviner D, Thouvenin V, et al. 2004. IMGT-ONTOLOGY for Immunogenetics and Immunoinformatics. *In Silico Biol*, Epub 2003 Nov 22; **4:** 17–29.

Lefranc M-P, Clément O, Kaas Q, Duprat E, Chastellan P, Coelho I, Combres K, Ginestoux C, Giudicelli V, Chaume D, et al. 2005a. IMGT-Choreography for Immunogenetics and Immunoinformatics. *In Silico Biol* **5:** 45–60.

Lefranc M-P, Pommié C, Kaas Q, Duprat E, Bosc N, Guiraudou D, Jean C, Ruiz M, Da Piedade I, Rouard M, et al. 2005b. IMGT unique numbering for immunoglobulin and T cell receptor constant domains and Ig superfamily C-like domains. *Dev Comp Immunol* **29:** 185–203.

Lefranc M-P, Duprat E, Kaas Q, Tranne M, Thiriot A, Lefranc G. 2005c. IMGT unique numbering for MHC groove G-DOMAIN and MHC superfamily (MhcSF) G-LIKE-DOMAIN. *Dev Comp Immunol* **29:** 917–938.

Lefranc M-P, Giudicelli V, Regnier L, Duroux P. 2008. IMGT, a system and an ontology that bridge biological and computational spheres in bioinformatics. *Brief Bioinform* **9:** 263–275.

Lefranc M-P, Giudicelli V, Ginestoux C, Jabado-Michaloud J, Folch G, Bellahcene F, Wu Y, Gemrot E, Brochet X, Lane J, et al. 2009. IMGT, the international ImMunoGeneTics information system. *Nucl Acids Res* **37:** D1006–D1012.

Pommié C, Levadoux S, Sabatier R, Lefranc M-P. 2004. IMGT standardized criteria for statistical analysis of immunoglobulin V-REGION amino acid properties. *J Mol Recognit* **17:** 17–32.

Ruiz M, Lefranc M-P. 2002. IMGT gene identification and Colliers de Perles of human immunoglobulins with known 3D structures. *Immunogenetics* **53:** 857–883.

Stamatopoulos K, Belessi C, Moreno C, Boudjograh M, Guida G, Smilevska T, Belhoul L, Stella S, Stavroyianni N, Crespo M, et al. 2007. Over 20% of patients with chronic lymphocytic leukemia carry stereotyped receptors: Pathogenetic implications and clinical correlations. *Blood* **109:** 259–270.

Yousfi Monod M, Giudicelli V, Chaume D, Lefranc M-P. 2004. IMGT/JunctionAnalysis: The first tool for the analysis of the immunoglobulin and T cell receptor complex V-J and V-D-J JUNCTIONs. *Bioinformatics* **20:** i379–i385.

Protocol

IMGT/JunctionAnalysis: IMGT Standardized Analysis of the V-J and V-D-J Junctions of the Rearranged Immunoglobulins (IG) and T Cell Receptors (TR)

Véronique Giudicelli and Marie-Paule Lefranc[1]

IMGT, the international ImMunoGeneTics information system, Laboratoire d'ImmunoGénétique Moléculaire LIGM, Université Montpellier 2, Institut de Génétique Humaine IGH, UPR CNRS 1142, 34396 Montpellier cedex 5, France

INTRODUCTION

IMGT/JunctionAnalysis is the online IMGT tool for the detailed and standardized analysis of the junctions between the variable (V), diversity (D), and joining (J) genes (V-J and V-D-J junctions) of the rearranged immunoglobulin (IG) or antibody and T cell receptor (TR) variable domains. The V-(D)-J junctions comprise the rearranged CDR3-IMGT and its anchors 2nd-CYS 104 and J-PHE or J-TRP 118. The diversity of the junctions that determines the antigen receptor specificity results from complex molecular mechanisms that occur at the DNA level during the IG and TR synthesis and create combinatorial diversity, N-diversity and, for IG, somatic hypermutations. The annotation of V-J or V-D-J junctions in rearranged IG and TR sequences represents a huge challenge due to its uniqueness and complexity. IMGT/JunctionAnalysis has been a major breakthrough by providing, for the first time, a very detailed and accurate analysis of the junctions. The tool, whose use is described here, identifies the D genes in the IGH, TRB, and TRD junctions, the trimmed nucleotides (nt) at the end of the genes which recombine, and the palindromic P regions in the absence of gene trimming. It delimits the N regions that result from the N-diversity, calculates the ratio of G+C nucleotides in the N regions, and evaluates the number of somatic hypermutations for each gene within the junction.

RELATED INFORMATION

IMGT/JunctionAnalysis (Lefranc 2004; Yousfi Monod et al. 2004; Giudicelli and Lefranc 2005, 2008) is part of IMGT, the international ImMunoGeneTics information system, http://www.imgt.org (Lefranc et al. 2009). A detailed description of IMGT is provided in **IMGT, the International ImMunoGeneTics Information System** (Lefranc 2011a). Standardization and IMGT Scientific chart rules are based on the IMGT-ONTOLOGY concepts of identification, classification, description (Giudicelli and Lefranc 1999; Lefranc et al. 2004, 2005a, 2008) and numerotation (Lefranc 1997, 1999; Lefranc et al. 2003, 2005b, c) generated from the axioms of the Formal IMGT-ONTOLOGY or IMGT-Kaleidoscope (Duroux et al. 2008). The concepts of identification led to the IMGT standardized keywords **(From IMGT-ONTOLOGY IDENTIFICATION Axiom to IMGT Standardized Keywords: For Immunoglobulins (IG), T Cell Receptors (TR), and Conventional Genes)** (Lefranc 2011b), the concepts of description to the IMGT standardized labels **(From IMGT-ONTOLOGY DESCRIPTION Axiom to IMGT Standardized Labels: For Immunoglobulin (IG) and T Cell Receptor (TR) Sequences and Structures)** (Lefranc 2011c), the concepts of classification to the IMGT standardized gene and allele names **(From IMGT-ONTOLOGY CLASSIFICATION Axiom to IMGT Standardized Gene and Allele Nomenclature: For Immunoglobulins (IG) and T Cell Receptors (TR))** (Lefranc 2011d), and the concepts of numerotation to the IMGT unique numbering **(IMGT Unique Numbering for the Variable (V), Constant (C), and Groove (G) Domains of IG, TR, MH, IgSF, and MhSF)** (Lefranc 2011e), and IMGT Colliers de Perles (**IMGT Collier de Perles for the Variable (V), Constant (C), and Groove (G) Domains of IG, TR, MH, IgSF, and MhSF)** (Lefranc 2011f).

In addition, protocols are available for **IMGT/V-QUEST: IMGT Standardized Analysis of the Immunoglobulin (IG) and T Cell Receptor (TR) Nucleotide Sequences** (Giudicelli et al. 2011),

[1]Corresponding author (Marie-Paule.Lefranc@igh.cnrs.fr).
Cite as: Cold Spring Harb Protoc; 2011; doi:10.1101/pdb.prot5634

IMGT/Collier de Perles: IMGT Standardized Representation of Domains (IG, TR, and IgSF Variable and Constant Domains, MH and MhSF Groove Domains) (Ehrenmann et al. 2011), **IMGT/DomainGapAlign: IMGT Standardized Analysis of Amino Acid Sequences of Variable, Constant, and Groove Domains (IG, TR, MH, IgSF, MhSF)** (Ehrenmann and Lefranc 2011a), and **IMGT/3Dstructure-DB: Querying the IMGT Database for 3D Structures in Immunology and Immunoinformatics (IG or Antibodies, TR, MH, RPI, and FPIA)** (Ehrenmann and Lefranc 2011b).

MATERIALS

It is essential that you consult the appropriate Material Safety Data Sheets and your institution's Environmental Health and Safety Office for proper handling of equipment and hazardous materials used in this protocol.

Equipment

Computer (Internet-connected)

METHOD

1. Using any modern web browser, access the IMGT Home page, http://www.imgt.org and click, in the "IMGT tools" section, on the link to IMGT/JunctionAnalysis.
   This gives access to the IMGT/JunctionAnalysis Welcome page.

IMGT/JunctionAnalysis Welcome Page

2. In the IMGT/JunctionAnalysis Welcome page (Fig. 1), locate at the top of the page:

   "Species" (drop-down list)
   "Locus" (radio buttons): IGH, IGK, IGL, TRA, TRB, TRG, TRD

   IMGT/JunctionAnalysis provides in a single run a standardized analysis of an unlimited number of IG and TR V-J and V-D-J junctions from the same locus and the same species. Analysis of IG and TR junctions of mice and humans can be performed exhaustively. Analysis of junctions of other species (rat, rabbit, trout) becomes progressively more available as genomic sequences are annotated in IMGT.

3. Select the species and the locus.

4. Locate the sequence submission section.

   That section comprises:

   Two windows to either type (or copy/paste) your sequence(s) or upload a file, in FASTA format.

   The "Start" and "Clear the form" buttons.

   Because the junctions include only a short part of the 3′ end of the V (3′V-REGION) and of the 5′ end of the J (5′J-REGION) (Bleakley et al. 2006, 2008), the tool cannot identify the V and J genes and alleles and, therefore, the corresponding gene and allele names must be provided and a specific format is required (detailed in Step 5).

5. Check the format of your junction nucleotide sequences. Keep the following points in mind:

   i. The required format is the FASTA format. Each JUNCTION nucleotide sequence must be preceded by the following information:

   - Identifier ("input"), with a 10-character maximum length. This identifier can be a sequence name, an accession number, a clone name, etc.
   - The name of the V-GENE and allele according to the IMGT gene name nomenclature (Lefranc 2000a,b, 2007, 2008; Lefranc and Lefranc 2001a,b; Giudicelli et al. 2005a).
   - The name of the J-GENE and allele according to the IMGT gene name nomenclature (Lefranc 2000a,b, 2007, 2008; Lefranc and Lefranc 2001a,b; Giudicelli et al. 2005a).

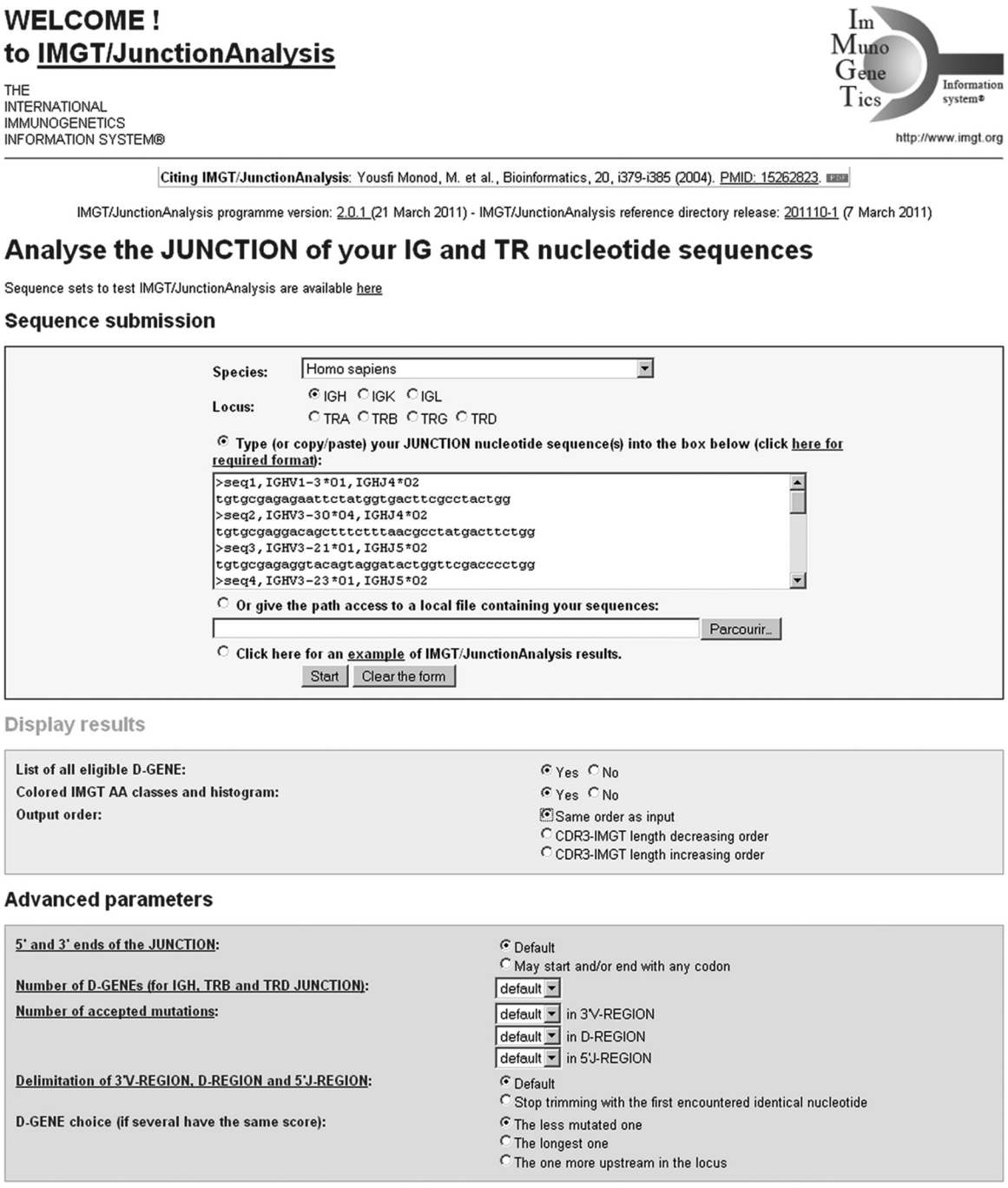


FIGURE 1. IMGT/JunctionAnalysis Welcome page.

ii. There is no limitation in the number of sequences to be analyzed in a single search, but a new line should be started for each sequence:

>Input1, V-GENE and allele name, J-GENE and allele name
nucleotide sequence (in uppercase or lowercase)
>Input2, V-GENE and allele name, J-GENE and allele name
nucleotide sequence (in uppercase or lowercase)
For instance,
>Seq1, IGHV7-4-1*02, IGHJ4*02

tgtgcgagagaagatagcaatggctacaaaatatttgactactgg
>Seq2, IGHV1-69*06, IGHJ5*02
tgtgcgagaggggggctaaggtcgaatttttggagtggtttcatgggtactggttcgacccctgg

iii. If the V-GENE allele or J-GENE allele is unknown, the IMGT/JunctionAnalysis tool accepts a "?" character instead of the allele number (ex: IGHV1-2*?) and will run the search against the allele *01 by default.

iv. If there are several proposed V-GENE and/or J-GENE, the different V-GENE and allele names, and/or J-GENE and allele names have to be separated by the "/" character (ex: IGHV1-2*01/IGHV1-3*?/IGHV1-18*02, IGHJ1*01/IGHJ2*01). The IMGT/JunctionAnalysis tool will run the search against the first V-GENE and allele and J-GENE and allele listed.

6. Select either "Type (or copy/paste) your JUNCTION nucleotide sequence(s) into the box below" and type (or copy and paste) the junction nucleotide sequences to be analyzed in the text area, or "Or give the path access to a local file containing your sequences" and upload a text file by using the "Browse" button to select the file.

7. For more information, locate "Click here for an example of IMGT/JunctionAnalysis" (option for first time users or for educational purposes):

   - Click on "example" to see the format for submission.
   - Select the radiobutton to automatically launch the analysis for the set of sequences indicated in the documentation.

Display Results

8. Locate the "Display results" section (Fig. 1). Perform Steps 9–11 for choosing options of the display, or if not needed, proceed to Step 12.

9. Choose, if preferred, not to display the "List of all eligible D-GENE."

   This option allows one to visualize all D genes that match a junction and to compare their score. It is displayed by default when only one junction is analyzed in the run, but can be eliminated.

10. Choose, if preferred, not to display the "Colored IMGT AA classes and histogram."

    A display of the "Colored IMGT AA classes and histogram," with colors of the AA according to the 11 IMGT physico-chemical AA classes (Pommié et al. 2004) (IMGT Aide-mémoire>Amino acids, http://www.imgt.org), is provided by default but can be eliminated.

11. Choose, if preferred, an "Output order" in "CDR3-IMGT length decreasing order" or in "CDR3-IMGT length increasing order."

    The results in "JUNCTION alignments with translation and IMGT AA classes" may be displayed in "Same order as input" (default), "CDR3-IMGT length decreasing order," or "CDR3-IMGT length increasing order."

Advanced Parameters

12. Locate the "Advanced parameters" section (Fig. 1). Perform Steps 13–17 for modifying parameters, or if not needed, proceed directly to Step 18.

    Advanced parameters are usually not needed with classical junctions. However, going through Steps 13–17 (even without modifying the parameters) will allow you to be aware of all the possibilities offered by the tool.

13. Check "5′ and 3′ ends of the JUNCTION" of your sequences and, if needed, select "May start and/or end with any codon."

    "Default" means that, as normally expected, your junction nucleotide sequences start in 5′ with 2nd-CYS 104 codon (V-REGION cysteine ["tgt" or "tgc"] codon) and end in 3′ with J-PHE or J-TRP 118 codon (J-REGION phenylalanine ["ttt" or "ttc"] or tryptophan ["tgg"]) codon). The conserved amino acids (AA) 2nd-CYS and J-PHE or J-TRP at positions 104 and 118, according to the IMGT unique numbering for V-DOMAIN (Lefranc 1997, 1999; Lefranc et al. 2003), are indeed the anchors of the CDR3-IMGT that delimit the JUNCTION.

    If one and/or the other of these anchor codon(s) is(are) mutated in some of your sequences, you can modify this parameter by selecting the other alternative "May start and/or end with any codon."

14. Modify, if needed, the "Number of D-GENE (for IGH, TRB, and TRD JUNCTION)" to be searched in the drop-down list.

Three loci IGH, TRB, and TRD loci have D genes in their V-D-J junction. "Default" means that the maximum number of D genes to be searched is one for IGH, one for TRB, and three for TRD.

You can modify the number of D genes to be searched in the junctions from 0 to 3 by selecting a number in the drop-down list.

15. Modify, if needed, the "Number of accepted mutations," in the corresponding 3′V-REGION, D-REGION, and 5′J-REGION drop-down lists.

The maximum number of accepted mutations in the 3′ V-REGION, D-REGION, and 5′ J-REGION depends on the locus. By default, no mutation is accepted for the TR junctions. For IGH, two, four, and two mutations are accepted in the 3′ V-REGION, D-REGION, and 5′ J-REGION, respectively. For the IGK and IGL loci, the maximum number of accepted mutations is seven in the 3′ V-REGION and seven in the 5′ J-REGION.

You can modify these numbers from 0 to 10 in the corresponding 3′ V-REGION, D-REGION, and 5′ J-REGION drop-down lists.

16. Select, if needed, for the "Delimitation of 3′V-REGION, D-REGION and 5′J-REGION," "Stop trimming with the first encountered identical nucleotide."

Rules have been defined to identify the trimmed (deleted) nucleotides at the 3′ end of the 3′ V-REGION, 5′ and 3′ ends of the D-REGION(s), and 5′ end of the 5′ J-REGION (Bleakley et al. 2006, 2008).

"Default" is "using patterns." This indicates that the tool identifies as trimmed the nucleotides that match the patterns "m," "m-" or "mm--" (at the 3′ end of the 3′ V-REGION and D-REGION), and the patterns "m," "-m," or "--mm" (at the 5′ end of the 5′ J-REGION and D-REGION) (where "m" indicates a mutation and "-" indicates an identical nucleotide by comparison with the corresponding closest germline alleles). The alternative is "Stop trimming with the first encountered identical nucleotide."

17. Select, if needed, for the "D-GENE choice (if several have the same score)," "The longest one," or "The one more upstream in the locus."

This criterion is used to discriminate between D regions with identical highest scores. "Default" is "The less mutated one." The other alternatives are "The longest one" or "The one more upstream in the locus."

18. Click on the "Start" button to launch the analysis.

This leads to the display of the IMGT/JunctionAnalysis Results page (Fig. 2 and Fig. 3).

See Discussion for further details.

DISCUSSION

The IMGT/JunctionAnalysis Results Page

The IMGT/JunctionAnalysis Results page provides the following:

1. A brief summary at the top of the page (Fig. 2).

It recalls the locus and species (as selected in Step 3).

It provides, for information, a link to the Locus representation in the IMGT Repertoire.

It recalls the values of the parameters used by the tool for the analysis:

- "Maximum number of accepted mutations" (see Step 15)
- "Deletion limits:" e.g., "using patterns" (default) (see Step 16)
- "Best D-GENE choice for a same score:" e.g., "less mutations" (default) (see Step 17)

2. "Analysis of the JUNCTION" (Fig. 2).

The "Analysis of the JUNCTION" provides the results of the analysis of the junctions at the nucleotide level. The junctions are displayed according to the order of the sequence submissions with the names of the input sequences and names of the V and J genes and alleles as provided by the user (in IMGT/JunctionAnalysis integrated in IMGT/V-QUEST (Lefranc 2004; Giudicelli and Lefranc 2005, 2008; Brochet et al. 2008) **(IMGT/V-QUEST: IMGT Standardized Analysis of the Immunoglobulin (IG) and T Cell Receptor (TR) Nucleotide Sequences** [Giudicelli et al. 2011], that information is from

WELCOME !
to IMGT/JunctionAnalysis

THE
INTERNATIONAL
IMMUNOGENETICS
INFORMATION SYSTEM®

http://www.imgt.org

Citing IMGT/JunctionAnalysis: Yousfi Monod, M. et al., Bioinformatics, 20, i379-i385 (2004). PMID: 15262823.

Analyse the JUNCTION of your IG and TR nucleotide sequences

Results

Locus IGH
Species Homo sapiens
IMGT Repertoire link Locus representation

Maximum number of accepted mutations:
V-REGION: 2; D-REGION: 4; J-REGION: 2

Deletion limits:
using patterns

Best D-GENE choice for a same score:
less mutations

Analysis of the JUNCTION

Click on mutated (underlined) nucleotide to see the original one:

	Input	V name	3'V-REGION	P	N1	D-REGION	N2	P	5'J-REGION	J name	D name	Vmut	Dmut	Jmut	Ngc
#1	seq1	IGHV1-3*01	tgtgcgagaga		att	...ctatggtg.....			...acttcgcctactgg	IGHJ4*02	IGHD4/OR15-4a*01	0	0	2	0/3
#2	seq2	IGHV3-30*04	tgtgcgag...		g	......acagcttttcttta.	acgc		....ctatgacttctgg	IGHJ4*02	IGHD5-5*01	0	3	2	4/5
#3	seq3	IGHV3-21*01	tgtgcgagag.			.gtacagtagga........	t		...actggttcgaccctgg	IGHJ5*02	IGHD1-26*01	0	2	0	0/1
#4	seq4	IGHV3-23*01	tgtgcga....		tcaacacaca	...gatataatggcta.....			..........cgacctctgg	IGHJ5*02	IGHD5-12*01	0	1	1	4/10
#5	seq5	IGHV3-30*03	tgtgcgaaag.		gggacgagcc	......cggggta.....			..............acggtatggacgtctgg	IGHJ6*02	IGHD4-23*01	1	1	0	8/10
#6	seq6	IGHV3-23*01	tgtgcgaaaga			......ctatgatagtagtggttat......	cc		.......tgactactgg	IGHJ4*02	IGHD3-22*01	0	0	0	2/2
#7	seq7	IGHV3-23*04	tgtgcgaaaga	t		.....ttaggttat...	cctat	t	actactttgactactgg	IGHJ4*02	IGHD3-16*01	0	2	0	2/5
#8	seq8	IGHV1-18*01	tgtgcgagag.		g	............agttggt..	tc		....ctttgactactgg	IGHJ4*02	IGHD6-13*01	0	1	0	2/3
#9	seq9	IGHV3-74*01	tgtgcaagaga	t	tcga	.....acgatttttggagtggttatta....	cggt		...........tactgg	IGHJ4*02	IGHD3-3*01	0	0	0	5/8
#10	seq10	IGHV3-30*03	tgtgcgaaaga	t	tc	......cggtgac...	ct		...tgcttttgatatctgg	IGHJ3*02	IGHD4-17*01	1	0	0	2/4
#11	seq11	IGHV1-69*09	tgtgcgaga		ccggaagcc	........gtaacagtaccagctcct.....	c		......ttgactactgg	IGHJ4*02	IGHD2-2*01	0	3	0	8/10

FIGURE 2. IMGT/JunctionAnalysis Results page: *top* of the page and "Analysis of the JUNCTION." Nucleotides of each region identified in a JUNCTION are displayed: 3′V-REGION, P regions (P), N regions (N1 and N2), D-REGION, and 5′J-REGION. Dots represent trimmed nucleotides. D name (IGH D-GENE and allele) is identified by IMGT/JunctionAnalysis. Underlined nucleotides are mutated compared with the germline region (the original nucleotide can be seen by clicking on the mutated one). Sequences seq1–seq11 correspond to the junction of sequences with accession numbers AJ399822, AB012909, AB027443, AF015123, AB021512, AB021521, AB063654, AB063683, AB021537, AB027433, AB027433, respectively, from the IMGT/LIGM-DB database (Giudicelli et al. 2006).

IMGT/V-QUEST analysis). Nucleotides of each region identified in a JUNCTION are displayed. Dots in 3′V-REGION, D-REGION, and 5′J-REGION indicate nucleotides trimmed in the rearranged sequence, by comparison to the corresponding germline 3′V-REGION, D-REGION, and 5′J-REGION.

Results provided by IMGT/JunctionAnalysis comprise:

- the display of the 3′V-REGION, D-REGION, and 5′J-REGION, with trimmed nucleotides shown by gaps;
- the D-GENE and allele name for IGH, TRB, and TRD loci;
- the N regions (N or N1, N2… if several, numbered from left to right) and P regions (P) (there is no numbering for the P-REGION);
- the number of mutations found in the 3′V-REGION, D-REGION, and 5′J-REGION (Vmut, Dmut, and Jmut, respectively), by comparison to the corresponding germline allele sequences;
- the ratio of the number of G+C nucleotides to the total number of N region nucleotides (Ngc).

You can click on a mutated (underlined) nucleotide to see the original one in the closest germline region.
Figure 3 shows the results obtained with an example of customization of Advanced parameters on one junction sequence.

3. "JUNCTION alignments with translation and IMGT AA classes" (Fig. 4).

The "JUNCTION alignments with translation and IMGT AA classes" (Fig. 4) provides the results of the analysis of the junctions at the AA level. Each JUNCTION nucleotide sequence is translated in amino

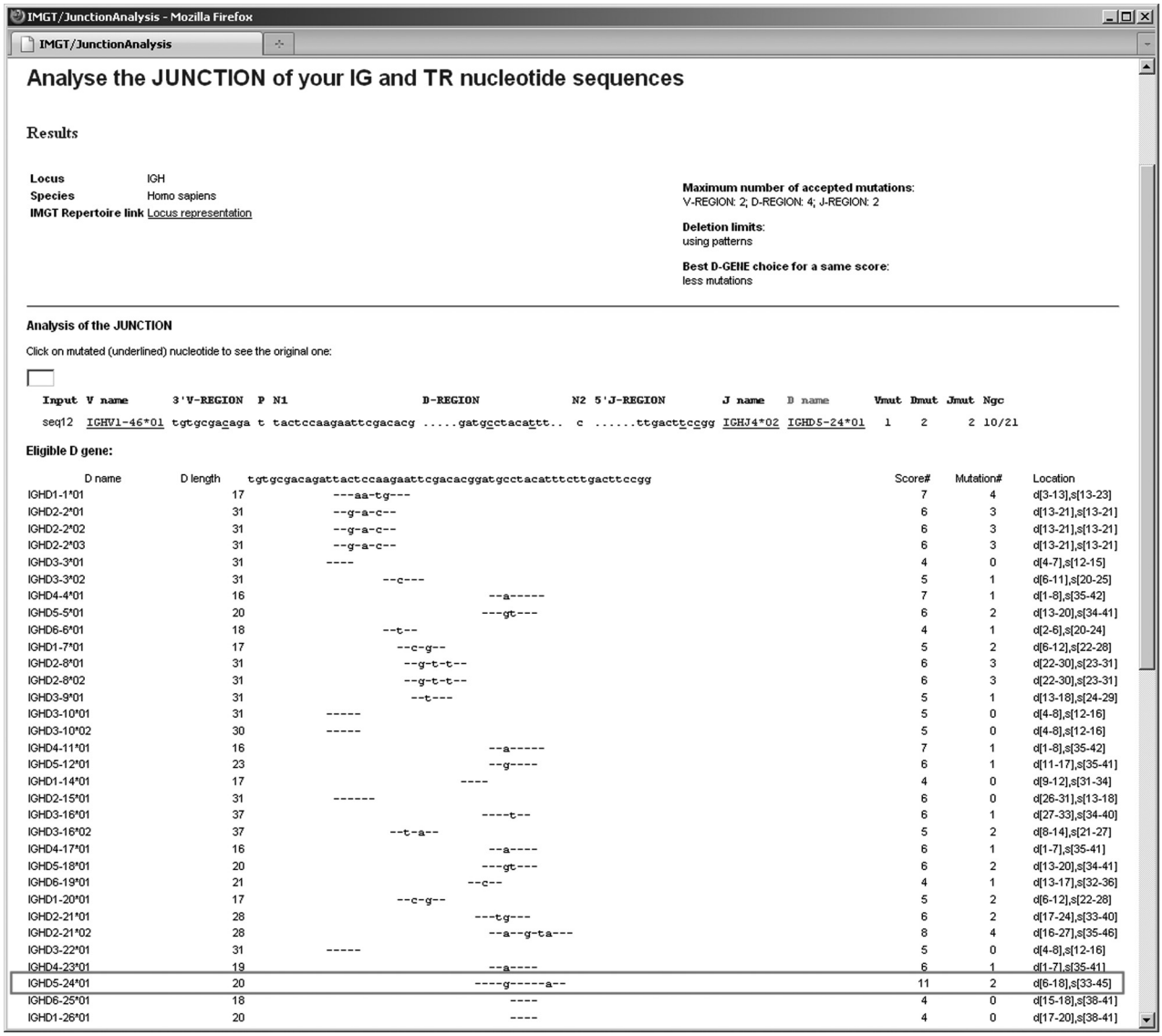


FIGURE 3. IMGT/JunctionAnalysis Results page "Analysis of the JUNCTION" with an example of customization of advanced parameters. In this example, the parameter "5′ and 3′ ends of the JUNCTION" has been set to "May start and/or end with any codon" to allow the analysis of that IGH junction that ends in 3′ with "cgg" instead of the classical "tgg" due to a mutation. Because there is only one sequence analyzed in that run, the list of eligible D genes is displayed by default. The D gene and allele with the best score is shown in the red rectangle. The location d[6-18],s[33-45] means that the nucleotides from position 6 to 18 of the germline IGHD5-24*01 D-REGION match the nucleotides of the junction sequence from position 33 to 45. Seq12 corresponds to the junction of the sequence with accession number AF006519 from the IMGT/LIGM-DB database (Giudicelli et al. 2006).

acid sequences. In the case of frameshifts, gaps indicated by one or two dots are inserted to maintain the J-REGION reading frame and to facilitate sequence comparison. Codons and AA are numbered according to the IMGT unique numbering for V-DOMAIN (Lefranc 1997, 1999; Lefranc et al. 2003). The numbering is made according to the longest JUNCTION obtained in the results. Colors of the AA classes are according to the 11 IMGT physicochemical AA classes (Pommié et al. 2004).
Note that:

- Gaps inserted in JUNCTION may split a D-REGION or a J-REGION, since the gaps are localized at the top of the CDR3-IMGT loops and depend on the CDR3-IMGT lengths and not on the sequence alignment.
- "*" indicates a STOP-CODON.

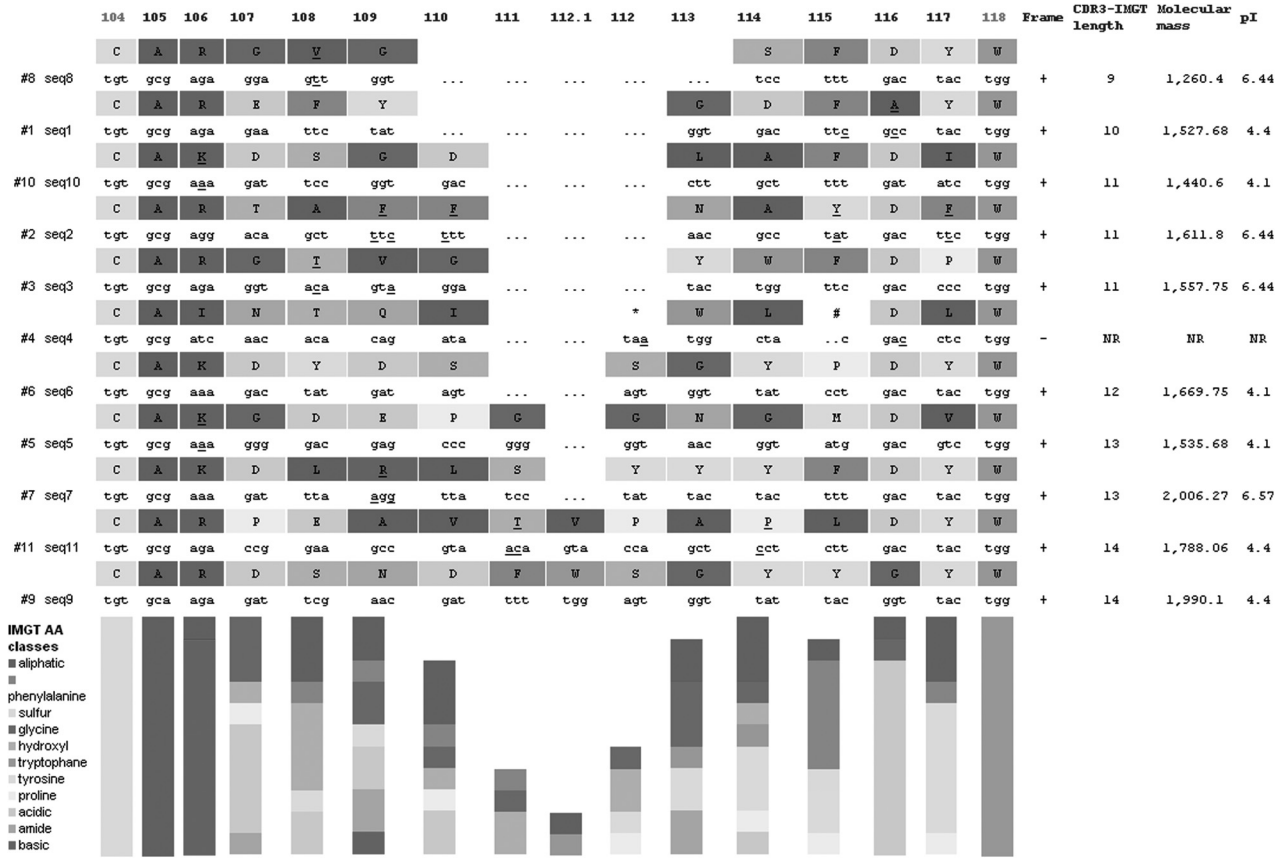


FIGURE 4. IMGT/JunctionAnalysis Results page: "JUNCTION alignments with translation and IMGT AA classes." On the *top*, the junctions are displayed (by "CDR3-IMGT length increasing order"). AA colors are according to the 11 IMGT physicochemical classes (Pommié et al. 2004). (*) a STOP-CODON (in seq4). In the "Frame" column: (+) an in-frame junction; (−) an out-of-frame junction. In the case of frameshifts, gaps indicated by one or two dots are inserted to maintain the J-REGION reading frame and to facilitate sequence comparison and the sign (#) is indicated in the AA sequence (in seq 4). Underlined nucleotides and amino acids are mutated compared with the germline genes (the original nucleotide or amino acids can be seen by clicking on the mutated one). For each junction, the length of the CDR3, the molecular mass, and the isoelectric point (pI) are shown on the *right*. NR indicates that the information is not relevant for out-of-frame junctions. On the *bottom*, a histogram represents the repartition of the 11 IMGT physicochemical AA classes (Pommié et al. 2004) in the junctions analyzed in the same run. Same sequences as in Figure 2.

- "#" indicates a frameshift.
- "+" and "–" at the end of the line indicates "in-frame" and "out-of-frame" JUNCTION, respectively.
- The CDR3-IMGT length, Molecular mass and pI values are indicated as "NR" (not relevant) in case of "out-of-frame" JUNCTION.

You can click on mutated nt or changed AA (underlined) to see the original nucleotide or amino acid in the closest germline region.

Uses of IMGT/JunctionAnalysis

IMGT/JunctionAnalysis (http://www.imgt.org) provides a highly detailed and standardized analysis of the IG and TR JUNCTION sequences: it describes precisely, at the nucleotide level, all characteristics of the junctions according to the complex mechanisms of IG and TR synthesis. IMGT/JunctionAnalysis is currently available for the analysis of junctions from species whose IG and TR loci are extensively studied (human, mouse). Other species (rat, rabbit, trout, etc.) are added as the genomic sequences are annotated in IMGT. Standardization of the results is based on the axioms and concepts of

IMGT-ONTOLOGY (Giudicelli and Lefranc 1999; Lefranc et al. 2004, 2005a, 2008; Duroux et al. 2008). IMGT/JunctionAnalysis has been integrated in IMGT/V-QUEST (Lefranc 2004; Giudicelli and Lefranc 2005, 2008; Brochet et al. 2008) **(IMGT/V-QUEST: IMGT Standardized Analysis of the Immunoglobulin (IG) and T Cell Receptor (TR) Nucleotide Sequences)** (Giudicelli et al. 2011) and in IMGT/HighV-QUEST (Alamyar et al. 2010, 2011). This allows the IMGT users to get, in a single tool, the identification of the V and J gene and allele names of IG and TR rearranged sequence, and the detailed analysis of the JUNCTION according to IMGT/JunctionAnalysis. In its version integrated in IMGT/V-QUEST and in IMGT/HighV-QUEST, IMGT/JunctionAnalysis used the default parameters. The benefit of the IMGT/JunctionAnalysis stand-alone version, available at the IMGT Home page (http://www.imgt.org), is that parameters can be modified by the users. Sequences of the junctions are provided by IMGT/V-QUEST (as "Input for IMGT/JunctionAnalysis" in "Sequences of the JUNCTION ("nt" and "AA")") for independent analysis in the stand-alone IMGT/JunctionAnalysis. Owing to the accuracy of the results, both tools are used by IMGT/Automat for the automatic annotation of the IG and TR cDNA sequences of IMGT/LIGM-DB (Giudicelli et al. 2003; 2005b). Results obtained by the users have important implications in basic and medical research (Stamatopoulos et al. 2007; Giudicelli and Lefranc 2008), antibody and TR specificity, and in biotechnology related to antibody engineering (Lefranc et al. 2008; Lefranc 2009; Ehrenmann et al. 2010).

ACKNOWLEDGMENTS

We are grateful to the IMGT team for their motivation and expertise.

REFERENCES

Alamyar E, Giudicelli V, Duroux P, Lefranc M-P. 2010. IMGT/HighV-QUEST: A high-throughput system and Web portal for the analysis of rearranged nucleotide sequences of antigen receptors - High-throughput version of IMGT/V-QUEST. JOBIM 2010 Poster 60. http://www.jobim2010.fr/?q=fr/node/55.

Alamyar E, Duroux D, Lefranc M-P, Giudicelli V. 2011. IMGT tools for the nucleotide analysis of immunoglobulin (IG) and T cell receptor (TR) V-(D)-J repertoires, polymorphisms, and IG mutations: IMGT/V-QUEST and IMGT/HighV-QUEST for NGS. *Methods Mol Biol* (in press).

Bleakley K, Giudicelli V, Wu Y, Lefranc M-P, Biau G. 2006. IMGT standardization for statistical analyses of T cell receptor junctions: The TRAV-TRAJ example. *In Silico Biol* **6:** 573–588.

Bleakley K, Lefranc M-P, Biau G. 2008. Recovering probabilities for nucleotide trimming processes for T cell receptor TRA and TRG V-J junctions analyzed with IMGT tools. *BMC Bioinformatics* **9:** 408. doi: 10.1186/1471-2105-9-408.

Brochet X, Lefranc M-P, Giudicelli V. 2008. IMGT/V-QUEST: The highly customized and integrated system for IG and TR standardized V-J and V-D-J sequence analysis. *Nucl Acids Res* **36:** W503–W508.

Duroux P, Kaas Q, Brochet X, Lane J, Ginestoux C, Lefranc M-P, Giudicelli V. 2008. IMGT-Kaleidoscope, the Formal IMGT-ONTOLOGY paradigm. *Biochimie* **90:** 570–583.

Ehrenmann F, Lefranc M-P. 2011a. IMGT/DomainGapAlign: IMGT standardized analysis of amino acid sequences of variable, constant and groove domains (IG, TR, MH, IgSF, MhSF). *Cold Spring Harb Protoc* doi: 10.1101/pdb.prot5636.

Ehrenmann F, Lefranc M-P. 2011b. IMGT/3Dstructure-DB: Querying the IMGT database for 3D structures in immunology and immunoinformatics (IG or antibodies, TR, MH, RPI, and FPIA). *Cold Spring Harb Protoc* doi: 10.1101/pdb.prot5637.

Ehrenmann F, Duroux P, Giudicelli V, Lefranc M-P. 2010. Standardized sequence and structure analysis of antibody using IMGT. In *Antibody engineering* 2nd ed. (ed. R Kontermann, S Dübel), Vol 2, pp. 11–31. Springer-Verlag, Berlin/Heidelberg, Germany.

Ehrenmann F, Giudicelli V, Brochet X, Lefranc M-P. 2011. IMGT/Collier de Perles: IMGT standardized representation of domains (IG, TR, and IgSF variable and constant domains, MH and MhSF groove domains). *Cold Spring Harb Protoc* doi: 10.1101/pdb.prot5635.

Giudicelli V, Lefranc M-P. 1999. Ontology for immunogenetics: IMGT-ONTOLOGY. *Bioinformatics* **15:** 1047–1054.

Giudicelli V, Lefranc M-P. 2005. Interactive IMGT on-line tools for the analysis of immunoglobulin and T cell receptor repertoires. In *New Research on Immunology* (ed. BA Veskler), pp. 77–105. Nova Science, Hauppauge, NY.

Giudicelli V, Lefranc M-P. 2008. IMGT standardized analysis of immunoglobulin rearranged sequences. In *Immunoglobulin gene analysis in Chronic Lymphocytic Leukemia* (ed. P Ghia, R Rosenquist, F Davi), pp. 33–52. Wolters Kluwer Health Italy, Italy.

Giudicelli V, Protat C, Lefranc M-P. 2003. The IMGT strategy for the automatic annotation of IG and TR cDNA sequences: IMGT/Automat. In *Proceedings of the European Conference on Computational Biology*, pp. 103–104. Paris, France.

Giudicelli V, Chaume D, Lefranc M-P. 2005a. IMGT/GENE-DB: A comprehensive database for human and mouse immunoglobulin and T cell receptor genes. *Nucleic Acids Res* **33:** D256–D261.

Giudicelli V, Chaume D, Jabado-Michaloud J, Lefranc M-P. 2005b. Immunogenetics sequence annotation: The strategy of IMGT based on IMGT-ONTOLOGY. *Stud Health Technol Inform* **116:** 3–8.

Giudicelli V, Duroux P, Ginestoux C, Folch G, Jabado-Michaloud J, Chaume D, Lefranc M-P. 2006. IMGT/LIGM-DB, the IMGT comprehensive database of immunoglobulin and T cell receptor nucleotide sequences. *Nucl Acids Res* **34:** D781–D784.

Giudicelli V, Brochet X, Lefranc M-P. 2011. IMGT/V-QUEST: IMGT standardized analysis of the immunoglobulin (IG) and T cell receptor (TR) nucleotide sequences. *Cold Spring Harb Protoc* doi: 10.1101/pdb.prot5633.

Lefranc M-P. 1997. Unique database numbering system for immunogenetic analysis. *Immunol Today* **18:** 509. doi: 10.1016/S0167-5699(97)01163-8.

Lefranc M-P. 1999. The IMGT unique numbering for Immunoglobulins, T cell receptors and Ig-like domains. *The Immunologist* **7:** 132–136.

Lefranc M-P. 2000a. Nomenclature of the human immunoglobulin genes. In *Current Protocols in Immunology* (ed. JE Coligan et al.), pp. A.1P.1–A.1P.37. John Wiley and Sons, Hoboken, NJ.

Lefranc M-P. 2000b. Nomenclature of the human T cell receptor genes. In *Current Protocols in Immunology* (ed. JE Coligan et al.), pp. A.1O.1–A.1O.23. John Wiley and Sons, Hoboken, NJ.

Lefranc M-P. 2004. IMGT, the International ImMunoGenetics Information System. In *Antibody engineering methods and protocols* 2nd ed. (ed. BKC Lo), Vol 248, pp. 27–49. Methods in Molecular Biology, Humana Press, Totowa, NJ.

Lefranc M-P. 2007. WHO-IUIS Nomenclature Subcommittee for Immunoglobulins and T cell receptors report. *Immunogenetics* **59:** 899–902.

Lefranc M-P. 2008. WHO-IUIS Nomenclature Subcommittee for Immunoglobulins and T cell receptors report August 2007, 13th International Congress of Immunology, Rio de Janeiro, Brazil. *Dev Comp Immunol* **32:** 461–463.

Lefranc M-P. 2009. Antibody database and tools: The IMGT experience. In *Therapeutic monoclonal antibodies: From bench to clinic* (ed. A Zhiqiang), pp. 91–114. John Wiley and Sons, Hoboken, NJ.

Lefranc M-P. 2011a. IMGT, the International ImMunoGeneTics Information System. *Cold Spring Harb Protoc* doi: 10.1101/pdb.top115.

Lefranc M-P. 2011b. From IMGT-ONTOLOGY IDENTIFICATION axiom to IMGT standardized keywords: For immunoglobulins (IG), T cell receptors (TR), and conventional genes. *Cold Spring Harb Protoc* doi: 10.1101/pdb.ip82.

Lefranc M-P. 2011c. From IMGT-ONTOLOGY DESCRIPTION axiom to IMGT standardized labels: For immunoglobulin (IG) and T cell receptor (TR) sequences and structures. *Cold Spring Harb Protoc* doi: 10.1101/pdb.ip83.

Lefranc M-P. 2011d. From IMGT-ONTOLOGY CLASSIFICATION axiom to IMGT standardized gene and allele nomenclature: For immunoglobulins (IG) and T cell receptors (TR). *Cold Spring Harb Protoc* doi: 10.1101/pdb.ip84.

Lefranc M-P. 2011e. IMGT unique numbering for the variable (V), constant (C), and groove (G) domains of IG, TR, MH, IgSF, and MhSF. *Cold Spring Harb Protoc* doi: 10.1101/pdb.ip85.

Lefranc M-P. 2011f. IMGT Collier de Perles for the Variable (V), Constant (C), and Groove (G) Domains of IG, TR, MH, IgSF, and MhSF. *Cold Spring Harb Protoc* doi: 10.1101/pdb.ip86.

Lefranc M-P, Lefranc G. 2001a. *The Immunoglobulin FactsBook*, 1–458. Academic Press, London, UK.

Lefranc M-P, Lefranc G. 2001b. *The T cell receptor FactsBook*, 1–398. Academic Press, London, UK.

Lefranc M-P, Pommié C, Ruiz M, Giudicelli V, Foulquier E, Truong L, Thouvenin-Contet V, Lefranc G. 2003. IMGT unique numbering for immunoglobulin and T cell receptor variable domains and Ig superfamily V-like domains. *Dev Comp Immunol* **27:** 55–77.

Lefranc M-P, Giudicelli V, Ginestoux C, Bosc N, Folch G, Guiraudou D, Jabado-Michaloud J, Magris S, Scaviner D, Thouvenin V, et al. 2004. IMGT-ONTOLOGY for Immunogenetics and Immunoinformatics. *In Silico Biology* **4:** 17–29.

Lefranc M-P, Clément O, Kaas Q, Duprat E, Chastellan P, Coelho I, Combres K, Ginestoux C, Giudicelli V, Chaume D, et al. 2005a. IMGT-Choreography for Immunogenetics and Immunoinformatics. *In Silico Biology* **5:** 45–60.

Lefranc M-P, Pommié C, Kaas Q, Duprat E, Bosc N, Guiraudou D, Jean C, Ruiz M, Da Piedade I, Rouard M, et al. 2005b. IMGT unique numbering for immunoglobulin and T cell receptor constant domains and Ig superfamily C-like domains. *Dev Comp Immunol* **29:** 185–203.

Lefranc M-P, Duprat E, Kaas Q, Tranne M, Thiriot A, Lefranc G. 2005c. IMGT unique numbering for MHC groove G-DOMAIN and MHC superfamily (MhcSF) G-LIKE-DOMAIN. *Dev Comp Immunol* **29:** 917–938.

Lefranc M-P, Giudicelli V, Regnier L, Duroux P. 2008. IMGT, a system and an ontology that bridge biological and computational spheres in bioinformatics. *Brief Bioinform* **9:** 263–275.

Lefranc M-P, Giudicelli V, Ginestoux C, Jabado-Michaloud J, Folch G, Bellahcene F, Wu Y, Gemrot E, Brochet X, Lane J, et al. 2009. IMGT, the international ImMunoGeneTics information system. *Nucl Acids Res* **37:** D1006–D1012.

Pommié C, Levadoux S, Sabatier R, Lefranc M-P. 2004. IMGT standardized criteria for statistical analysis of immunoglobulin V-REGION amino acid properties. *J Mol Recognit* **17:** 17–32.

Stamatopoulos K, Belessi C, Moreno C, Boudjograh M, Guida G, Smilevska T, Belhoul L, Stella S, Stavroyianni N, Crespo M, et al. 2007. Over 20% of patients with chronic lymphocytic leukemia carry stereotyped receptors: Pathogenetic implications and clinical correlations. *Blood* **109:** 259–270.

Yousfi Monod M, Giudicelli V, Chaume D, Lefranc M-P. 2004. IMGT/JunctionAnalysis: The first tool for the analysis of the immunoglobulin and T cell receptor complex V-J and V-D-J JUNCTIONs. *Bioinformatics* **20:** i379–i385.

Protocol

IMGT/Collier de Perles: IMGT Standardized Representation of Domains (IG, TR, and IgSF Variable and Constant Domains, MH and MhSF Groove Domains)

François Ehrenmann, Véronique Giudicelli, Patrice Duroux, and Marie-Paule Lefranc[1]

IMGT, the international ImMunoGeneTics information system, Laboratoire d'ImmunoGénétique Moléculaire LIGM, Université Montpellier 2, Institut de Génétique Humaine IGH, UPR CNRS 1142, 34396 Montpellier cedex 5, France

INTRODUCTION

IMGT/Collier de Perles is the IMGT online tool that allows one to draw standardized IMGT two-dimensional (2D) graphical representations of protein domains, or IMGT Colliers de Perles, starting from the user's own domain amino acid (AA) sequences. The use of this tool is described here. IMGT Colliers de Perles can be drawn for three domain types: the variable (V) domain and constant (C) domain of the immunoglobulin (IG) and T cell receptor (TR) and other members of the immunoglobulin superfamily (IgSF), and the groove (G) domain of the major histocompatibility (MH) and other members of the MH superfamily (MhSF). Sequences have to be gapped according to the IMGT unique numbering, using, for example, IMGT/DomainGapAlign. Resulting IMGT Colliers de Perles allow one to quickly visualize amino acids, which are important for the three-dimensional (3D) structural configuration, and to delimit the standardized framework regions (FR-IMGT) and complementarity determining regions (CDR-IMGT) of the IG and TR V-DOMAIN. The length of the strands, loops, and turns and, for the G type, the length of the helix represented in IMGT Colliers de Perles become crucial information in the domain characterization.

RELATED INFORMATION

IMGT/Collier de Perles (Ruiz and Lefranc 2002; Kaas and Lefranc 2007; Kaas et al. 2007) is a tool of IMGT http://www.imgt.org (Lefranc et al. 2009) using standardization based on the IMGT-ONTOLOGY concepts of identification, classification, description (Giudicelli and Lefranc 1999; Lefranc et al. 2004, 2005a, 2008) and numerotation (Lefranc 1997, 1999; Lefranc et al. 2003, 2005b,c), generated from the axioms of the Formal IMGT-ONTOLOGY or IMGT-Kaleidoscope (Duroux et al. 2008). The concepts of identification led to the IMGT standardized keywords, the concepts of classification to the IMGT standardized gene and allele names, the concepts of description to the IMGT standardized labels, and the concepts of numerotation to the IMGT unique numbering.

A detailed description of IMGT is provided in **IMGT, the International ImMunoGeneTics Information System** (Lefranc 2011a). Information is also available on **From IMGT-ONTOLOGY IDENTIFICATION Axiom to IMGT Standardized Keywords: For Immunoglobulins (IG), T Cell Receptors (TR), and Conventional Genes** (Lefranc 2011b), **From IMGT-ONTOLOGY DESCRIPTION Axiom to IMGT Standardized Labels: For Immunoglobulin (IG) and T Cell Receptor (TR) Sequences and Structures** (Lefranc 2011c), **From IMGT-ONTOLOGY CLASSIFICATION Axiom to IMGT Standardized Gene and Allele Nomenclature: For Immunoglobulins (IG) and T Cell Receptors (TR)** (Lefranc 2011d), **IMGT Unique Numbering for the Variable (V), Constant (C), and Groove (G) Domains of IG, TR, MH, IgSF, and MhSF** (Lefranc 2011e), and **IMGT Collier de Perles for the Variable (V), Constant (C), and Groove (G) Domains of IG, TR, MH, IgSF, and MhSF** (Lefranc 2011f).

In addition, protocols are available for **IMGT/V-QUEST: IMGT Standardized Analysis of the Immunoglobulin (IG) and T Cell Receptor (TR) Nucleotide Sequences** (Giudicelli et al. 2011), **IMGT/JunctionAnalysis: IMGT Standardized Analysis of the V-J and V-D-J Junctions of the Rearranged Immunoglobulins (IG) and T Cell Receptors (TR)** (Giudicelli and Lefranc 2011),

[1]Corresponding author (Marie-Paule.Lefranc@igh.cnrs.fr).
Cite as: Cold Spring Harb Protoc; 2011; doi:10.1101/pdb.prot5635

IMGT/DomainGapAlign: IMGT Standardized Analysis of Amino Acid Sequences of Variable, Constant, and Groove Domains (IG, TR, MH, IgSF, MhSF) (Ehrenmann and Lefranc 2011a), and **IMGT/3Dstructure-DB: Querying the IMGT Database for 3D Structures in Immunology and Immunoinformatics (IG or Antibodies, TR, MH, RPI, and FPIA)** (Ehrenmann and Lefranc 2011b).

MATERIALS

It is essential that you consult the appropriate Material Safety Data Sheets and your institution's Environmental Health and Safety Office for proper handling of equipment and hazardous materials used in this protocol.

Equipment

Computer (Internet-connected).

METHOD

The IMGT/Collier de Perles tool web application at the IMGT website (http://www.imgt.org) allows one to "Make Your Own IMGT Colliers de Perles" for V domain, C domain, or G domain starting from a user's amino acid sequence. This protocol describes how to draw "IMGT Colliers de Perles" using the IMGT/Collier de Perles tool.

IMGT Collier de Perles for V Domain

1. Using any modern web browser, go to the IMGT Home page http://www.imgt.org.
2. Click on the "IMGT/Collier de Perles" link in the IMGT tools section of the IMGT Home page.
   This will take you to the IMGT/Collier de Perles tool web application (Fig. 1). By default, the web application is configured for V type domains.
3. Locate the "Domain type" drop-down list: "Variable (V)" is displayed.
   "Variable (V)" allows one to draw IMGT Colliers de Perles for variable domains, which include the V-DOMAIN of IG and TR and the V-LIKE-DOMAIN of IgSF other than IG and TR.

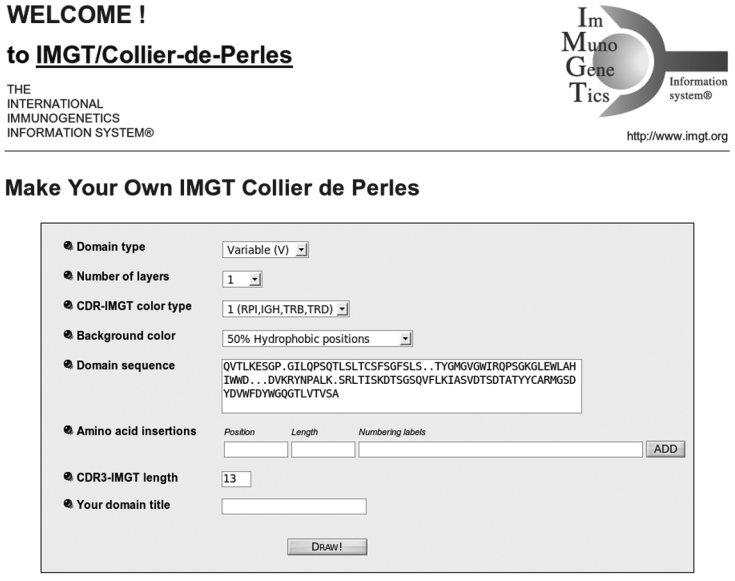


FIGURE 1. IMGT/Collier de Perles Welcome page. By default the page is configured for a V domain.

4. Locate the "Number of layers" drop-down list. Select "1" or "2," depending on the number of layers you want for the display of the IMGT Collier de Perles. Default is "1."

   This choice is only available for V and C domains.

5. Locate the "CDR-IMGT color type" drop-down list, and select "1 (RPI,IGH,TRB,TRD)" or "2 (IGK,IGL, TRA,TRG)," depending on the locus to which the sequence belongs.

   This choice is only available for V domains.

   For "1 (RPI,IGH,TRB,TRD)," CDR-IMGT regions are colored as follows: CDR1-IMGT (red), CDR2-IMGT (orange), and CDR3-IMGT (purple).

   For "2 (IGK,IGL,TRA,TRG)," CDR-IMGT regions are colored as follows: CDR1-IMGT (blue), CDR2-IMGT (green), and CDR3-IMGT (greenblue).

6. Locate the "Background color" drop-down list, and select one of the 10 lines, depending on the property you want to highlight in your IMGT Collier de Perles, by comparison with previously analyzed domains of the same locus (Pommié et al. 2004).

   This choice is only available for V domains. Ten selections are provided (the first one by default):

   - *"50% Hydrophobic positions:" positions that have a hydrophobic amino acid (hydropathy index with positive value) or a tryptophan (W) like 50% or more of analyzed V domains (blue positions in IMGT Collier de Perles).*
   - *"IGH 80% hydropathy classes:" positions with amino acids that have the same hydropathy class, like 80% or more of analyzed IGH V domains (three classes).*
   - *"IGK 80% hydropathy classes:" positions with amino acids that have the same hydropathy class, like 80% or more of analyzed IGK V domains (three classes).*
   - *"IGL 80% hydropathy classes:" positions with amino acids that have the same hydropathy class, like 80% or more of analyzed IGL V domains (three classes).*
   - *"IGH 80% volume classes:" positions with amino acids that have the same volume class, like 80% or more of anaylzed IGH V domains (five classes).*
   - *"IGK 80% volume classes:" positions with amino acids that have the same volume class, like 80% or more of analyzed IGK V domains (five classes).*
   - *"IGL 80% volume classes:" positions with amino acids that have the same volume class, like 80% or more of analyzed IGL V domains (five classes).*
   - *"IGH 80% physicochemical classes:" positions with amino acids that have the same physicochemical class, like 80% or more of analyzed IGH V domains (11 classes).*
   - *"IGK 80% physicochemical classes:" positions with amino acids that have the same physicochemical class, like 80% or more of analyzed IGK V domains (11 classes).*
   - *"IGL 80% physicochemical classes" positions with amino acids that have the same physicochemical class, like 80% or more of analyzed IGL V domains (11 classes).*

7. Locate the "Domain sequence" text area, and enter your gapped sequence obtained, for example, from **IMGT/DomainGapAlign (IMGT/DomainGapAlign: IMGT Standardized Analysis of Amino Acid Sequences of Variable, Constant, and Groove Domains (IG, TR, MH, IgSF, MhSF)** (Ehrenmann and Lefranc 2011a).

   A complete IMGT Collier de Perles of an IG or TR V-DOMAIN includes the CDR3-IMGT and the FR4-IMGT. The FR4-IMGT is composed of at least nine or 10 AA beyond the phenylalanine F (J-PHE 118) or tryptophan W (J-TRP 118) of the motif F/W-G-X-G that characterizes the J-REGION.

   If the length of the CDR3-IMGT of your sequence is less than 13, you need to add gaps at the top of the loop (Lefranc et al. 2003; Lefranc 2011e) http://www.imgt.org/textes/IMGTScientificChart/Numbering/IMGTIGVLsuperfamily.html. If the CDR3-IMGT is equal to or greater than 13, the tool will automatically adjust the number of positions to the length (provided that you fill in the required field in Step 9).

8. Locate the "Amino acid insertions" text area, and enter, if needed, amino acid insertions.

   This option is only needed if you deal with an unusual sequence characterized by insertions compared with the reference sequences. This can be the case for the IgSF V-LIKE-DOMAINs that have amino acid insertions compared with the IG and TR V-DOMAIN. The program for renumbering is based on the Smith and Waterman alignment, which uses IMGT

insertions as new amino acids and strongly penalizes the opening of gaps in the closest reference sequence. To add an insertion, enter the position that precedes the insertion, the length (that is the number of amino acids to add) and the numbering label(s) of the novel positions: for examples "84,1,84A" means that, following 84, one additional amino acid labeled 84A has to be added; "84,3,84A,84B,84C" means that, following 84, three additional amino acids labeled 84A, 84B, and 84C have to be added. Then click "ADD." Repeat the process if several insertions have to be entered.

9. Locate the "CDR3-IMGT length" text area, and enter the length of the CDR3-IMGT (required).

10. (Optional) Locate the "Your domain title" text area, and enter a name that you would like to be displayed in the figure.

11. Press the "Draw !" button to visualize the IMGT Collier de Perles for V domain.

    This will return a page showing the IMGT Collier de Perles taking into account your choices (for examples, see Figs. 2–6). In the case of unusual sequences, it is possible to proceed by trial and error, adding or deleting gaps manually, in the "Domain sequence" text area (Step 7).

IMGT Collier de Perles for C Domain

12. Follow Steps 1 and 2 above.

13. Locate the "Domain type" drop-down list, and select the domain type "Constant (C)." This will change the page of the web application (Fig. 7).

    "Constant (C)" allows one to draw IMGT Colliers de Perles for constant domains, which include the C-DOMAIN of IG and TR and the C-LIKE-DOMAIN of IgSF other than IG and TR.

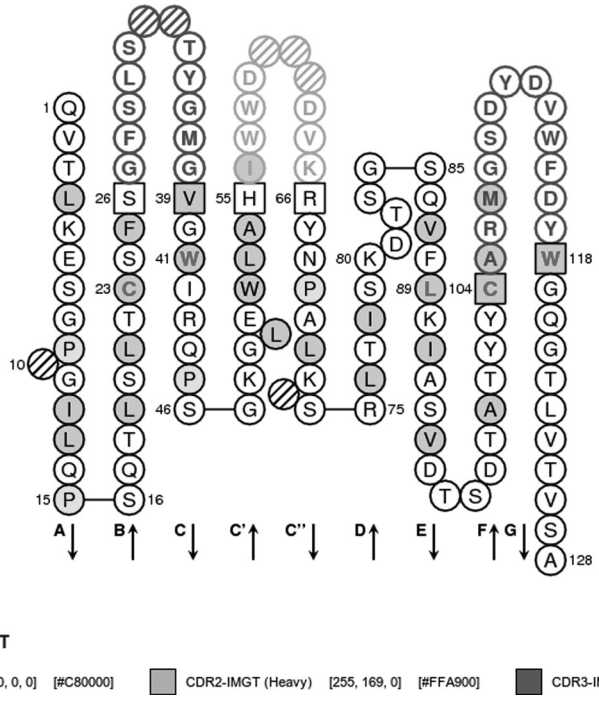


FIGURE 2. IMGT Collier de Perles of a V domain, on one layer, obtained after selection of "1 (RPI,IGH,TRB,TRD)" in CDR-IMGT color type, the domain belonging to an IGH chain, and by default "50% Hydrophobic positions." Positions in blue mean that the amino acid of the user sequence at these positions is hydrophobic (hydropathy index with positive value) or is a tryptophan (W), like in 50% or more of analyzed V domains. Positions with red and bold letters indicate the five conserved positions of a V domain: 1st-CYS 23, CONSERVED-TRP 41, hydrophobic 89, 2nd-CYS 104, and J-TRP 118. Anchor positions are in squares. Hatched positions correspond to gaps according to the IMGT unique numbering for V domain (Lefranc 1997, 1999; Lefranc et al. 2003). Prolines are shown in yellow. Arrows indicate the β strands and their direction. The CDR-IMGT lengths of this domain are [10.7.13].

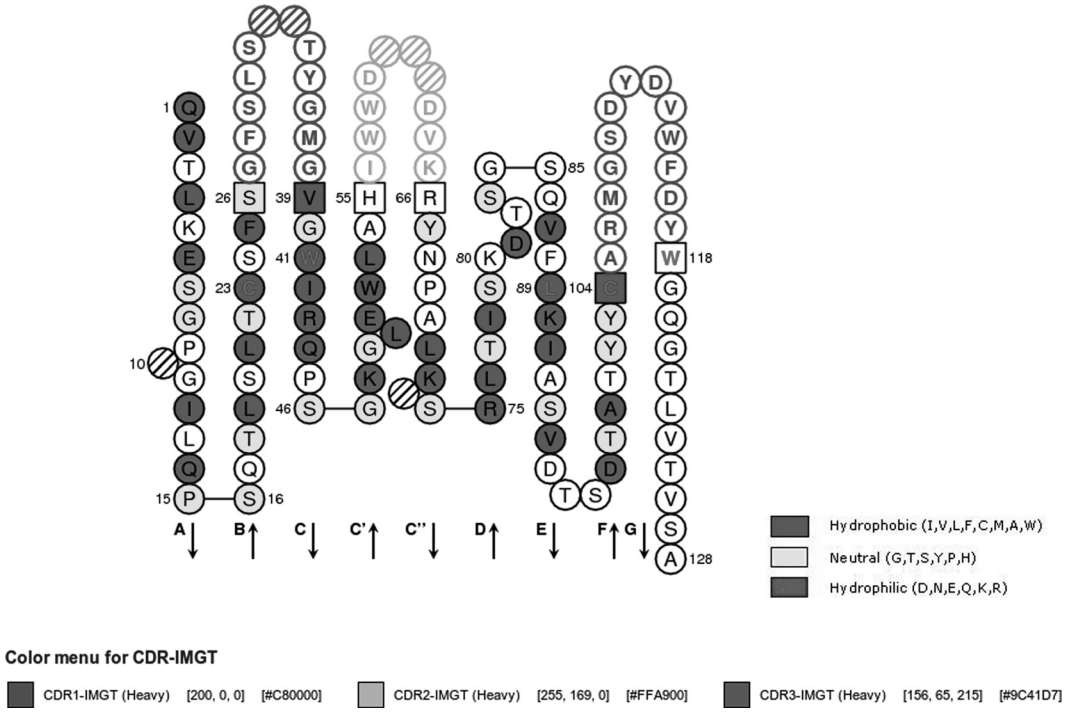


FIGURE 3. IMGT Collier de Perles of a V domain with "IGH 80% hydropathy classes" positions. Colored positions (three classes) mean that the property of the amino acids of the user sequence at these positions is found in >80% of the analyzed IGH V domains.

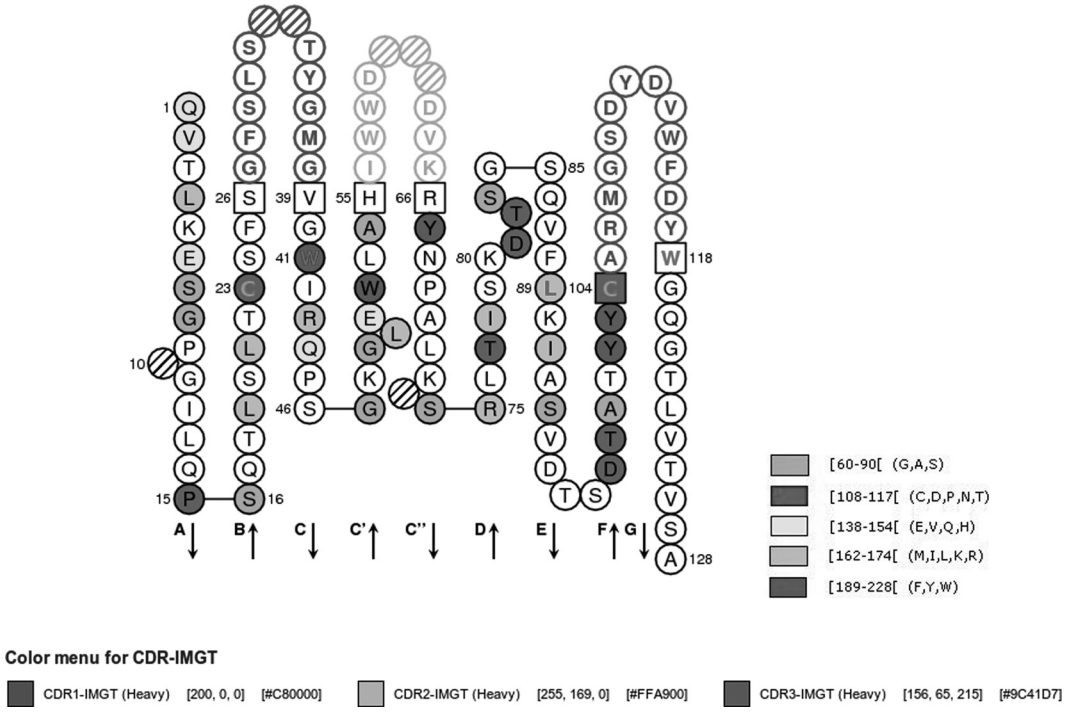


FIGURE 4. IMGT Collier de Perles of a V domain with "IGH 80% volume classes" positions. Colored positions (five classes) mean that the property of the amino acids of the user sequence at these positions is found in >80% of the analyzed IGH V domains.

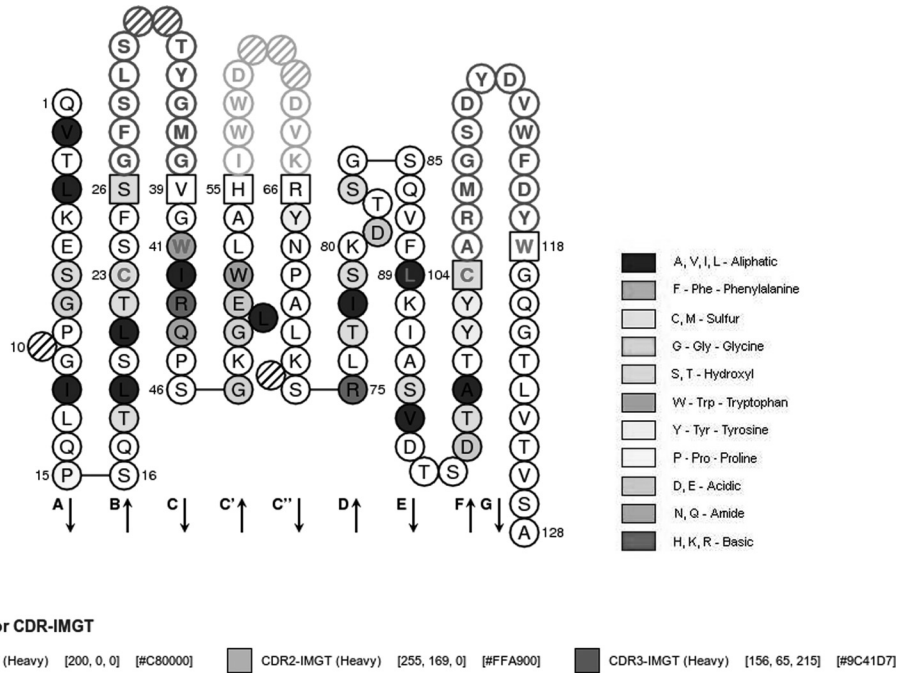


FIGURE 5. IMGT Collier de Perles of a V domain with "IGH 80% physicochemical classes" positions. Colored positions (11 classes) mean that the property of the amino acids of the user sequence at these positions is found in >80% of the analyzed IGH V domains.

14. Locate the "Number of layers" drop-down list. Select "1" or "2," depending on the number of layers you want for the display of the IMGT Collier de Perles. Default is "1."

    This choice is only available for V and C domains.

15. Locate the "Domain sequence" text area, and enter your gapped sequence obtained, for example, from IMGT/DomainGapAlign.

16. Locate the "Amino acid insertions" text area, and enter, if needed, amino acid insertions.

    This option is only needed if you deal with an unusual sequence characterized by insertions compared with the reference sequences. This can be the case for IgSF C-LIKE-DOMAIN that have amino acid insertions compared with the IG and TR C-DOMAIN. See Step 8 for more information.

17. (Optional) Locate the "Your domain title" text area, and enter a name that you would like to be displayed in the figure.

18. Press the "Draw !" button to visualize the IMGT Collier de Perles for C domain.

    This will return a page showing the IMGT Collier de Perles, taking into account your choices (e.g., see Figs. 8 and 9). In the case of unusual sequences, it is possible to proceed by trial and error, adding or deleting gaps manually in the "Domain sequence" text area (see Step15).

IMGT Collier de Perles for G Domain

19. Follow Steps 1 and 2 above.

20. Locate the "Domain type" drop-down list, and select the domain type "Groove (G)." This will change the page of the web application (Fig. 10).

    "Groove (G)" allows one to draw IMGT Colliers de Perles for groove domains, which include the G-DOMAIN of MH class I (or MH1) and class II (or MH2) and the G-LIKE-DOMAIN of MhSF other than MH (or related proteins of the immune system [RPI]-MH1Like). As a groove is made of 2 G domains, the sequences of the two domains for the IMGT Colliers de Perles are entered in the same page.

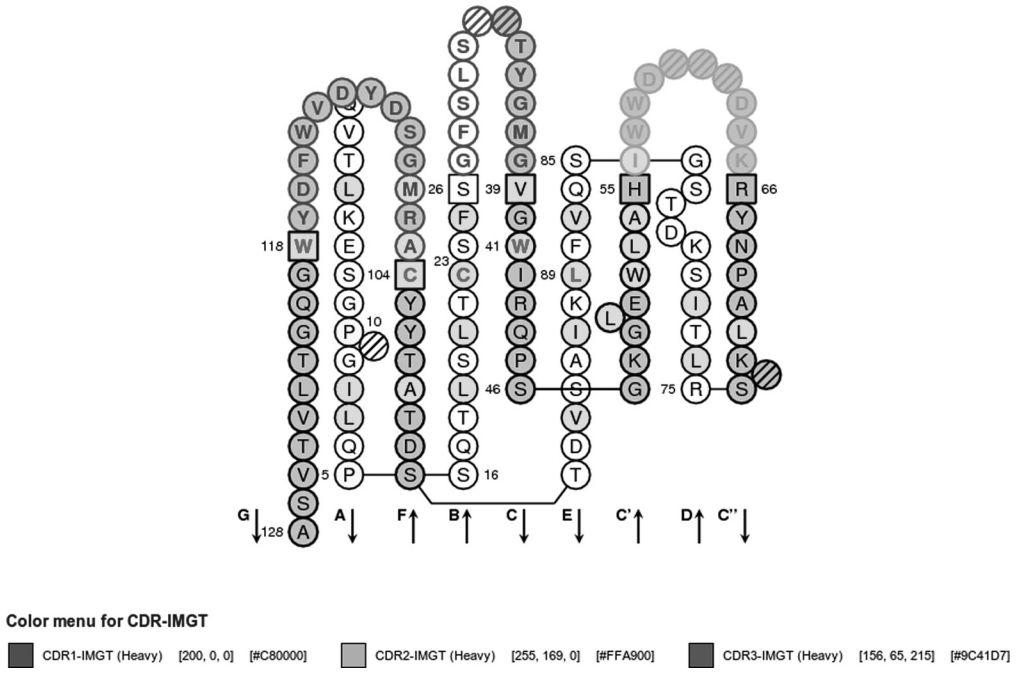


FIGURE 6. IMGT Collier de Perles of a V domain, on two layers, obtained after selection of "2" in Number of layers, "1 (RPI, IGH,TRB,TRD)" in CDR-IMGT color type, the domain belonging to an IGH chain, and by default "50% Hydrophobic positions." Positions in blue mean that the amino acid of the user sequence at these positions is hydrophobic (hydropathy index with positive value) or is a tryptophan (W), like in 50% or more of analyzed V domains. Positions with red and bold letters indicate the five conserved positions of a V domain: 1st-CYS 23, CONSERVED-TRP 41, hydrophobic 89, 2nd-CYS 104, and J-TRP 118. Anchor positions are in squares. Hatched positions correspond to gaps according to the IMGT unique numbering for V domain (Lefranc 1997, 1999; Lefranc et al. 2003). Prolines are shown in yellow. Arrows indicate the β strands and their direction. The CDR-IMGT lengths of this domain are [10.7.13].

21. Locate "Link G domains" and click on the button if your sequences are domains of MH1 or RPI-MH1Like, to have them linked.

22. Locate the "Domain sequence 1" text area, and enter the gapped sequence obtained, for example, from IMGT/DomainGapAlign, for the first G domain.

    For MH1, the sequence is that of the G-ALPHA1, for MH2, that of G-ALPHA, and for RPI-MH1Like, that of G-ALPHA1-LIKE.

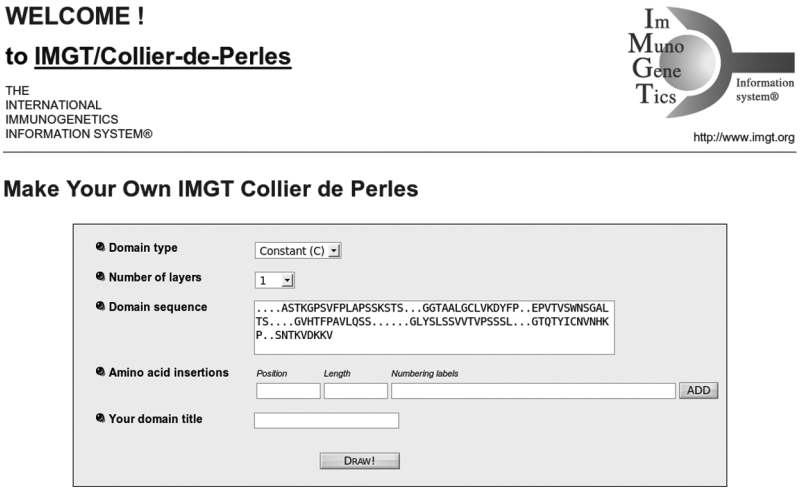


FIGURE 7. IMGT/Collier de Perles page for a C domain.

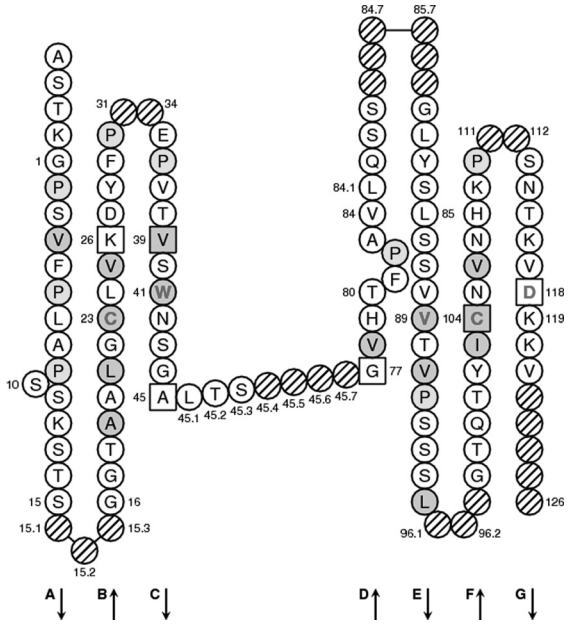


FIGURE 8. IMGT Collier de Perles of a C domain on one layer. The default is "50% Hydrophobic positions." Positions in blue mean that the amino acid of the user sequence at these positions is hydrophobic (hydropathy index with positive value) or is a tryptophan (W), as in 50% or more of analyzed C domains. Positions with red and bold letters are by analogy with a V domain. In a C domain they correspond to the two conserved cysteines, 1st-CYS 23 and 2nd-CYS 104, and to the usually hydrophobic positions 41 and 89, whereas position 118 is not conserved. Anchor positions are in squares. Hatched positions correspond to gaps according to the IMGT unique numbering for C domain (Lefranc et al. 2005b). Prolines are shown in yellow. Arrows indicate the β strands and their direction.

23. Locate the "Amino acid insertions" text area, and enter, if needed, AA insertions.

This option is only needed if you deal with unusual sequences characterized by insertions compared with the reference sequences. This can be the case for RPI-MH1Like G-LIKE-DOMAIN that have amino acid insertions compared with the MH G-DOMAIN. See Step 8 for more information.

Note that position 7A of G-ALPHA1 and positions 61A, 61B, and 72A of G-ALPHA2 are entered by default, as these insertions are characteristic of these domains, respectively.

24. (Optional) Locate the "Your domain title 1" text area, and enter a name that you would like to be displayed in the figure for the first G domain.

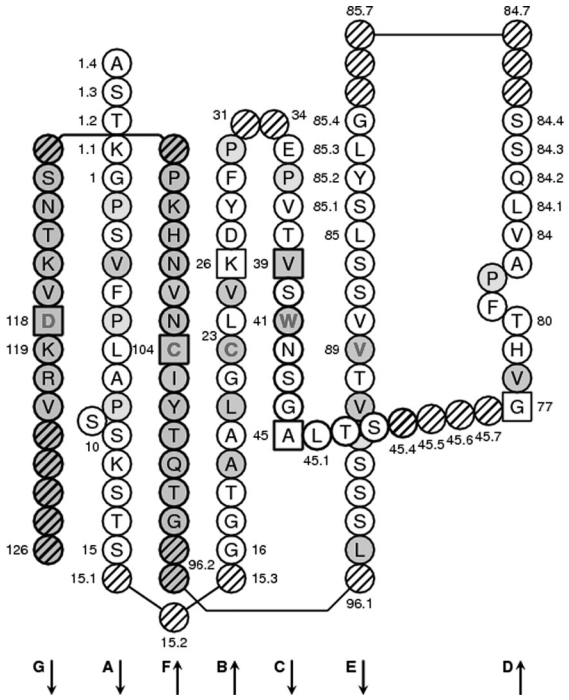


FIGURE 9. IMGT Collier de Perles of a C domain on two layers. The default is "50% Hydrophobic positions." Positions in blue mean that the amino acid of the user sequence at these positions is hydrophobic (hydropathy index with positive value) or is a tryptophan (W), like in 50% or more of analyzed C domains. Positions with red and bold letters indicate four conserved positions of a C domain: 1st-CYS 23, CONSERVED-TRP 41, hydrophobic 89, and 2nd-CYS 104, whereas position 118 (not conserved in a C domain) is only shown by analogy with a V domain. Anchor positions are in squares. Hatched positions correspond to gaps according to the IMGT unique numbering for C domain (Lefranc et al. 2005b). Prolines are shown in yellow. Arrows indicate the β strands and their direction.

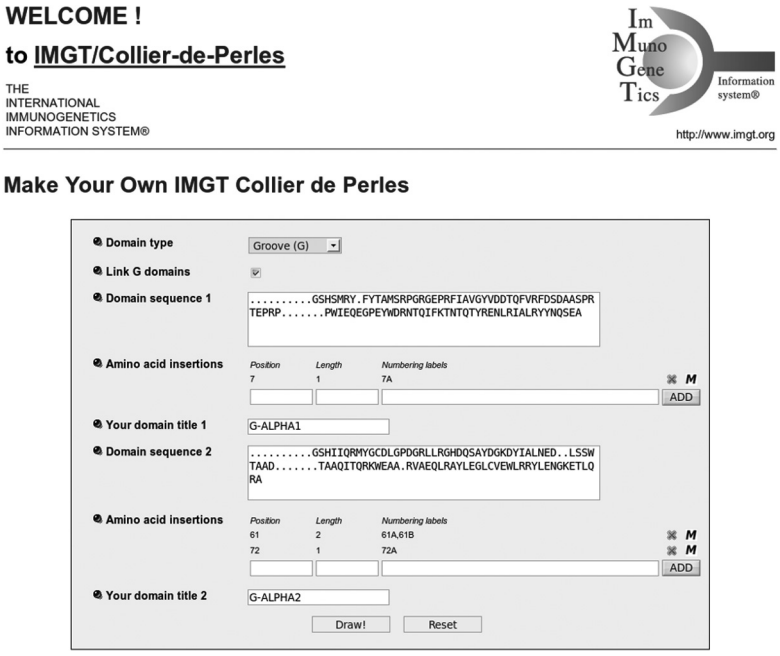


FIGURE 10. Page: IMGT/Collier de Perles for G domains.

25. Locate the "Domain sequence 2" text area, and enter the gapped sequence obtained, for example, from IMGT/DomainGapAlign, for the second G domain.

For MH1, the sequence is that of the G-ALPHA2, for MH2, that of G-BETA, and for RPI-MH1 Like, that of G-ALPHA2-LIKE.

26. Locate the "Amino acid insertions" text area, and enter, if needed, AA insertions.

See Step 23 for more information.

27. (Optional) Locate the "Your domain title 2" text area, and enter a name that you would like to be displayed in the figure for the second G domain.

28. Press the "Draw !" button to visualize the IMGT Collier de Perles for G domain.

This will return a page showing the IMGT Collier de Perles taking into account your choices (Fig. 11). In the case of unusual sequences, it is possible to proceed by trial and error, adding or deleting gaps manually, in the "Domain sequence" text area (see Step 22).

DISCUSSION

The IMGT Colliers de Perles allow a precise visualization of the interspecies differences for the IgSF V and C domain strands and loops and MhSF G domain strands and helix, even in the absence of 3D structures (Lefranc et al. 2008). The IMGT Colliers de Perles are particularly useful in molecular engineering and antibody humanization design based on CDR grafting to visualize the CDR-IMGT (Lefranc 2009; Ehrenmann et al. 2010a) and to easily compare the amino acid sequences of the four FR-IMGT between the murine or nonhuman primate and the closest human V-DOMAIN (Pelat et al. 2008). A recent analysis performed on humanized antibodies used in oncology underlines the importance of correct delimitation of the CDR regions to be grafted (Magdelaine-Beuzelin et al. 2007). The IMGT Colliers de Perles also allow a comparison to the IMGT Collier de Perles statistical profiles for the human expressed IGHV, IGKV, and IGLV repertoires (Pommié et al. 2004).

The IMGT Colliers de Perles for the V, C, and G domains, based on the IMGT unique numbering (Lefranc 1997, 1999; Lefranc et al. 2003, 2005b,c) therefore represent a major step forward for the comparative analysis of the sequences and structures of the IgSF and MhSF domains (Kaas et al. 2004; **IMGT/DomainGapAlign: IMGT Standardized Analysis of Amino Acid Sequences of Variable, Constant, and Groove Domains (IG, TR, MH, IgSF, MhSF)** (Ehrenmann and Lefranc 2011a; **IMGT/3Dstructure-DB: Querying the IMGT Database for 3D Structures in Immunology and**

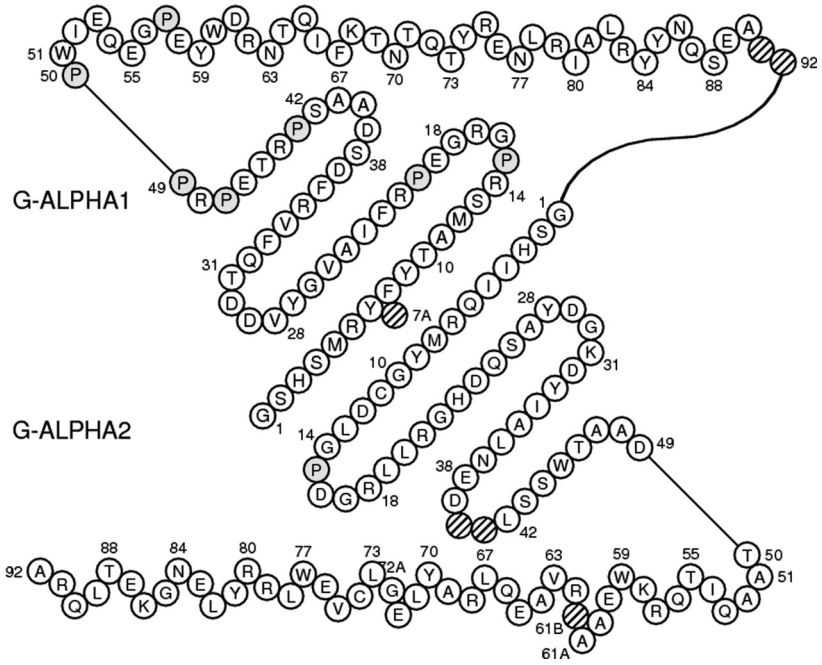


FIGURE 11. IMGT Colliers de Perles of G domains of MH1. The domains are linked as selected by the user. Hatched positions correspond to gaps according to the IMGT unique numbering for G domain (Lefranc et al. 2005c). Prolines are shown in yellow.

Immunoinformatics (IG or Antibodies, TR, MH, RPI, and FPIA) (Ehrenmann and Lefranc 2011b; Ehrenmann et al. 2010b) and the study of their evolution. In basic and clinical research, the IMGT Colliers de Perles provide new insights in the analysis of the IG and TR repertoires in autoimmune diseases and leukemias (Stamatopoulos et al. 2007), in the study of interactions between IG and antigen (IG/Ag) and between TR, peptide and MH (TR/pMH), and in antibody engineering and humanization for therapeutic applications (Lefranc 2009; Ehrenmann et al. 2010a).

ACKNOWLEDGMENTS

We thank Christophe le Roy for checking the protocol. We are grateful to Gérard Lefranc and to the IMGT team for their motivation and expertise.

REFERENCES

Duroux P, Kaas Q, Brochet X, Lane J, Ginestoux C, Lefranc M-P, Giudicelli V. 2008. IMGT-Kaleidoscope, the Formal IMGT-ONTOLOGY paradigm. *Biochimie* **90:** 570–583.

Ehrenmann F, Lefranc M-P. 2011a. IMGT/DomainGapAlign: IMGT standardized analysis of amino acid sequences of variable, constant and groove domains (IG, TR, MH, IgSF, MhSF). *Cold Spring Harb Protoc* doi: 10.1101/pdb.prot5636.

Ehrenmann F, Lefranc M-P. 2011b. IMGT/3Dstructure-DB: Querying the IMGT database for 3D structures in immunology and immunoinformatics (IG or antibodies, TR, MH, RPI, and FPIA). *Cold Spring Harb Protoc* doi: 10.1101/pdb.prot5637.

Ehrenmann F, Duroux P, Giudicelli V, Lefranc M-P. 2010a. Standardized sequence and structure analysis of antibody using IMGT. In *Antibody engineering* 2nd ed. (ed. R Kontermann, S Dübel), Vol 2, 11–31. Springer-Verlag, Berlin/Heidelberg, Germany.

Ehrenmann F, Kaas Q, Lefranc M-P. 2010b. IMGT/3Dstructure-DB and IMGT/DomainGapAlign: A database and a tool for immunoglobulins or antibodies, T cell receptors, MHC, IgSF and MhcSF. *Nucleic Acids Res* **38:** D301–D307.

Giudicelli V, Lefranc M-P. 1999. Ontology for immunogenetics: IMGT-ONTOLOGY. *Bioinformatics* **15:** 1047–1054.

Giudicelli V, Lefranc M-P. 2011. IMGT/JunctionAnalysis: IMGT standardized analysis of the V-J and V-D-J junctions of the rearranged immunoglobulins (IG) and T cell receptors (TR). *Cold Spring Harb Protoc* doi: 10.1101/pdb.prot5634.

Giudicelli V, Brochet X, Lefranc M-P. 2011. IMGT/V-QUEST: IMGT standardized analysis of the immunoglobulin (IG) and T cell receptor (TR) nucleotide sequences. *Cold Spring Harb Protoc* doi: 10.1101/pdb.prot5633.

Kaas Q, Lefranc M-P. 2007. IMGT Colliers de Perles: Standardized sequence-structure representations of the IgSF and MhcSF superfamily domains. *Current Bioinformatics* **2:** 21–30.

Kaas Q, Ruiz M, Lefranc M-P. 2004. IMGT/3Dstructure-DB and IMGT/StructuralQuery, a database and a tool for immunoglobulin, T cell receptor and MHC structural data. *Nucl Acids Res* **32:** D208–D210.

Kaas Q, Ehrenmann F, Lefranc M-P. 2007. IG, TR and IgSf, MHC and MhcSF: What do we learn from the IMGT Colliers de Perles? *Brief Funct Genomic Proteomic* **6:** 253–264.

Lefranc M-P. 1997. Unique database numbering system for immunogenetic analysis. *Immunology Today* **18:** 509. doi: 10.1016/S0167-5699(97)01163-8.

Lefranc M-P. 1999. The IMGT unique numbering for Immunoglobulins, T cell receptors and Ig-like domains. *The Immunologist* **7:** 132–136.

Lefranc M-P. 2009. Antibody database and tools: The IMGT experience. In *Therapeutic monoclonal antibodies: From bench to clinic* (ed. A Zhiqiang), pp. 91–114. John Wiley Sons, Inc, Hoboken, NJ.

Lefranc M-P. 2011a. IMGT, the International ImMunoGeneTics Information System. *Cold Spring Harb Protoc* doi: 10.1101/pdb.top115.

Lefranc M-P. 2011b. From IMGT-ONTOLOGY IDENTIFICATION axiom to IMGT standardized keywords: For immunoglobulins (IG), T cell receptors (TR), and conventional genes. *Cold Spring Harb Protoc* doi: 10.1101/pdb.ip82.

Lefranc M-P. 2011c. From IMGT-ONTOLOGY DESCRIPTION axiom to IMGT standardized labels: For immunoglobulin (IG) and T cell receptor (TR) sequences and structures. *Cold Spring Harb Protoc* doi: 10.1101/pdb.ip83.

Lefranc M-P. 2011d. From IMGT-ONTOLOGY CLASSIFICATION axiom to IMGT standardized gene and allele nomenclature: For immunoglobulins (IG) and T cell receptors (TR). *Cold Spring Harb Protoc* doi: 10.1101/pdb.ip84.

Lefranc M-P. 2011e. IMGT unique numbering for the variable (V), constant (C), and groove (G) domains of IG, TR, MH, IgSF, and MhSF. *Cold Spring Harb Protoc* doi: 10.1101/pdb.ip85.

Lefranc M-P. 2011f. IMGT Collier de Perles for the Variable (V), Constant (C), and Groove (G) Domains of IG, TR, MH, IgSF, and MhSF. *Cold Spring Harb Protoc* doi: 10.1101/pdb.ip86.

Lefranc M-P, Pommié C, Ruiz M, Giudicelli V, Foulquier E, Truong L, Thouvenin-Contet V, Lefranc G. 2003. IMGT unique numbering for immunoglobulin and T cell receptor variable domains and Ig superfamily V-like domains. *Dev Comp Immunol* **27:** 55–77.

Lefranc M-P, Giudicelli V, Ginestoux C, Bosc N, Folch G, Guiraudou D, Jabado-Michaloud J, Magris S, Scaviner D, Thouvenin V, et al. 2004. IMGT-ONTOLOGY for Immunogenetics and Immunoinformatics. *In Silico Biol* **4:** 17–29.

Lefranc M-P, Clément O, Kaas Q, Duprat E, Chastellan P, Coelho I, Combres K, Ginestoux C, Giudicelli V, Chaume D, et al. 2005a. IMGT-Choreography for Immunogenetics and Immunoinformatics. *In Silico Biol* **5:** 45–60.

Lefranc M-P, Pommié C, Kaas Q, Duprat E, Bosc N, Guiraudou D, Jean C, Ruiz M, Da Piedade I, Rouard M, et al. 2005b. IMGT unique numbering for immunoglobulin and T cell receptor constant domains and Ig superfamily C-like domains. *Dev Comp Immunol* **29:** 185–203.

Lefranc M-P, Duprat E, Kaas Q, Tranne M, Thiriot A, Lefranc G. 2005c. IMGT unique numbering for MHC groove G-DOMAIN and MHC superfamily (MhcSF) G-LIKE-DOMAIN. *Dev Comp Immunol* **29:** 917–938.

Lefranc M-P, Giudicelli V, Regnier L, Duroux P. 2008. IMGT, a system and an ontology that bridge biological and computational spheres in bioinformatics. *Brief Bioinform* **9:** 263–275.

Lefranc M-P, Giudicelli V, Ginestoux C, Jabado-Michaloud J, Folch G, Bellahcene F, Wu Y, Gemrot E, Brochet X, Lane J, et al. 2009. IMGT, the international ImMunoGeneTics information system. *Nucl Acids Res* **37:** D1006–D1012.

Magdelaine-Beuzelin C, Kaas Q, Wehbi V, Ohresser M, Jefferis R, Lefranc M-P, Watier H. 2007. Structure-function relationships of the variable domains of monoclonal antibodies approved for cancer treatment. *Crit Rev Oncol Hematol* **64:** 210–225.

Pelat T, Bedouelle H, Rees AR, Crennell SJ, Lefranc M-P, Thullier P. 2008. Germline humanization of a non-human Primate antibody that neutralizes the anthrax toxin, by in vitro and in silico engineering. *J Mol Biol* **384:** 1400–1407.

Pommié C, Levadoux S, Sabatier R, Lefranc M-P. 2004. IMGT standardized criteria for statistical analysis of immunoglobulin V-REGION amino acid properties. *J Mol Recognit* **17:** 17–32.

Ruiz M, Lefranc M-P. 2002. IMGT gene identification and Colliers de Perles of human immunoglobulins with known 3D structures. *Immunogenetics* **53:** 857–883.

Stamatopoulos K, Belessi C, Moreno C, Boudjograh M, Guida G, Smilevska T, Belhoul L, Stella S, Stavroyianni N, Crespo M, et al. 2007. Over 20% of patients with chronic lymphocytic leukemia carry stereotyped receptors: Pathogenetic implications and clinical correlations. *Blood* **109:** 259–270.

Protocol

IMGT/DomainGapAlign: IMGT Standardized Analysis of Amino Acid Sequences of Variable, Constant, and Groove Domains (IG, TR, MH, IgSF, MhSF)

François Ehrenmann and Marie-Paule Lefranc[1]

IMGT, the international ImMunoGeneTics information system, Laboratoire d'ImmunoGénétique Moléculaire LIGM, Université Montpellier 2, Institut de Génétique Humaine IGH, UPR CNRS 1142, 34396 Montpellier cedex 5, France

INTRODUCTION

IMGT/DomainGapAlign is the IMGT online tool for the analysis of amino acid (AA) sequences and two-dimensional (2D) structures of domains. The use of this tool is described here. Three domain types can be analyzed: the variable (V) and constant (C) domains of the immunoglobulins (IG) and T cell receptors (TR) and other immunoglobulin superfamily (IgSF) proteins, and the groove (G) domain of major histocompatibility (MH) and other MH superfamily (MhSF) proteins. IMGT/DomainGapAlign creates gaps in the user amino acid sequences and delimits, for the V and C domains, the framework regions (FR-IMGT) and complementarity determining regions (CDR-IMGT), and, for the G domains, the strands and helix, according to the IMGT unique numbering. The tool provides alignments with the closest sequences from the IMGT domain directory of IMGT/DomainDisplay. For the IG and TR V domains that result from V-(D)-J rearrangement, the tool provides alignments with the translation of the IMGT/GENE-DB germline V and joining (J) genes. IMGT/DomainGapAlign displays the amino acid changes, highlights them in IMGT Colliers de Perles, and provides tables with their detailed description, according to the IMGT rules. Several amino acid sequences can be analyzed simultaneously, provided that they belong to the same domain type.

RELATED INFORMATION

IMGT/DomainGapAlign (Ehrenmann et al. 2010b) is part of IMGT, the international ImMunoGeneTics information system, http://www.imgt.org (Lefranc et al. 2009). Standardization and IMGT Scientific chart rules are based on the IMGT-ONTOLOGY concepts of identification, classification, description (Giudicelli and Lefranc 1999; Lefranc et al. 2004, 2005a, 2008) and numerotation (Lefranc 1997, 1999; Lefranc et al. 2003, 2005b,c) generated from the axioms of the Formal IMGT-ONTOLOGY or IMGT-Kaleidoscope (Duroux et al. 2008). The concepts of identification led to the IMGT standardized keywords, the concepts of classification to the IMGT standardized gene and allele names, the concepts of description to the IMGT standardized labels, and the concepts of numerotation to the IMGT unique numbering.

The IMGT/DomainGapAlign tool web application at the IMGT website allows one to align the user amino acid domain sequences, to identify, in terms of sequence identity, the closest V, C, and G domains in the IMGT domain reference directory (Lefranc 2010d; Giudicelli et al. 2005), to create gaps according to the IMGT unique numbering (Lefranc 1997, 1999; Lefranc et al. 2003, 2005b,c), to highlight differences with the closest reference(s), and to generate IMGT Colliers de Perles (**IMGT/Collier de Perles: IMGT Standardized Representation of Domains (IG, TR, and IgSF Variable and Constant Domains, MH and MhSF Groove Domains))** (Ehrenmann et al. 2011). Several amino acid sequences can be analyzed simultaneously, provided that they belong to the same domain type (V, C, or G).

A detailed description of IMGT is provided in **IMGT, the International ImMunoGeneTics Information System** (Lefranc 2011a). Information is also available on **From IMGT-ONTOLOGY IDENTIFICATION Axiom to IMGT Standardized Keywords: For Immunoglobulins (IG), T Cell Receptors (TR), and Conventional Genes** (Lefranc 2011b), **From IMGT-ONTOLOGY DESCRIPTION Axiom to IMGT Standardized Labels: For Immunoglobulin (IG) and T Cell Receptor (TR) Sequences and Structures** (Lefranc 2011c), **From IMGT-ONTOLOGY CLASSIFICATION Axiom to IMGT Standardized**

[1]Corresponding author (Marie-Paule.Lefranc@igh.cnrs.fr).
Cite as: Cold Spring Harb Protoc; 2011; doi:10.1101/pdb.prot5636

Gene and Allele Nomenclature: For Immunoglobulins (IG) and T Cell Receptors (TR) (Lefranc 2011d), **IMGT Unique Numbering for the Variable (V), Constant (C), and Groove (G) Domains of IG, TR, MH, IgSF, and MhSF** (Lefranc 2011e), and **IMGT Collier de Perles for the Variable (V), Constant (C), and Groove (G) Domains of IG, TR, MH, IgSF, and MhSF** (Lefranc 2011f).

In addition, protocols are available for **IMGT/V-QUEST: IMGT Standardized Analysis of the Immunoglobulin (IG) and T Cell Receptor (TR) Nucleotide Sequences** (Giudicelli et al. 2011), **IMGT/Junction Analysis: IMGT Standardized Analysis of the V-J and V-D-J Junctions of the Rearranged Immunoglobulins (IG) and T Cell Receptors (TR)** (Giudicelli and Lefranc 2011), and **IMGT/3Dstructure-DB: Querying the IMGT Database for 3D Structures in Immunology and Immunoinformatics (IG or Antibodies, TR, MH, RPI, and FPIA)** (Ehrenmann and Lefranc 2011).

MATERIALS

It is essential that you consult the appropriate Material Safety Data Sheets and your institution's Environmental Health and Safety Office for proper handling of equipment and hazardous materials used in this protocol.

Equipment

Computer (Internet-connected)

METHOD

Querying IMGT/DomainGapAlign Tool

1. Using any modern web browser, go to the IMGT Home page http://www.imgt.org.
2. Click on the "IMGT/DomainGapAlign" link in the "IMGT tools" section of the IMGT Home page.
   This will take you to the IMGT/ DomainGapAlign tool web application (Fig. 1)
3. Locate the text area at the top of the page. Paste your amino acid sequences in FASTA format. Alternatively, you can upload a file.
   A precise delimitation of the domain sequences is not required; however, if the sequence contains several domains, the sequence should be split between the different domains. Several domains can be analyzed simultaneously, but each

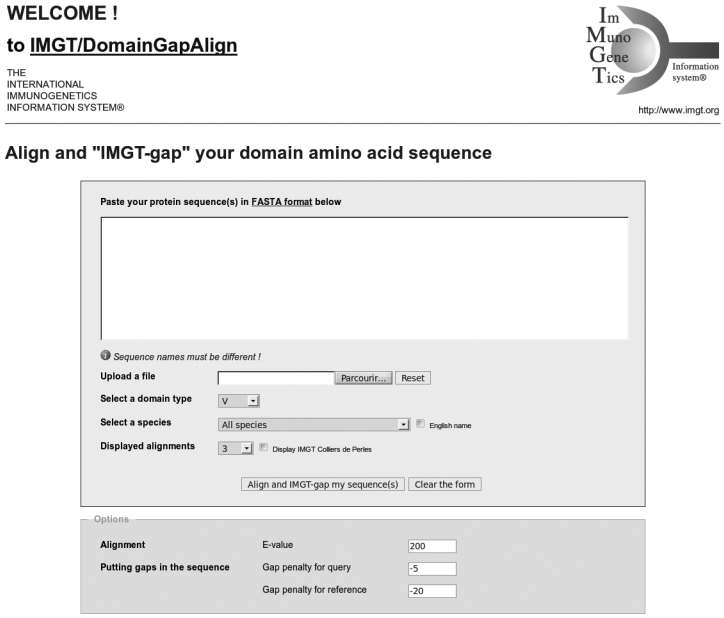


FIGURE 1. IMGT/DomainGapAlign Welcome page.

sequence must have a distinct name. If the limits and the numbers of domains of an amino acid sequence are unknown, you can progressively analyze the protein, shortening the sequence once a domain has been identified by the tool. (Note that the first domain identified by the tool is not necessarily the first one in the protein).

4. Locate the "Domain type" drop-down list, and select the domain type (V, C, or G) corresponding to the sequences to analyze.
   - *V for "variable" domain (V-DOMAIN of IG and TR and V-LIKE-DOMAIN of IgSF other than IG and TR)*
   - *C for "constant" domain (C-DOMAIN of IG and TR and C-LIKE-DOMAIN of IgSF other than IG and TR)*
   - *G for "groove" domain (G-DOMAIN of MH and G-LIKE-DOMAIN of MhSF other than MH)*
5. Locate the "Species" drop-down list, and select a Species.
   If the selection is "All species," the IMGT/DomainGapAlign tool will propose the result(s), taking into account the best Smith and Waterman scores and percentage of identity. If the selection is forced with a species, such as "Homo sapiens," the IMGT/DomainGapAlign tool will propose the result(s), taking into account the species, the best Smith and Waterman score(s), and percentage of identity.
6. Locate the "Displayed alignments" drop-down list and select the number of alignments to display (by default 3). Tick off the checkbox if you want to display the IMGT Colliers de Perles.
7. Check the "Options" section if you would like to modify the parameters.
   The alignment is performed by a modified Smith-Waterman algorithm with an asymmetrical management of insertions between the two sequences. By default, the gap penalty is –5 for the user sequence and –20 for the reference sequence.
8. Press the "Align and IMGT-gap my sequence(s)" button to launch the analysis.
   This will return the results page.

V Domain Analysis Using IMGT/DomainGapAlign

9. Follow Steps 1–8 above, selecting "V" in Step 4.
   The results display is similar for a V-DOMAIN or a V-LIKE-DOMAIN, despite the huge difference in term of biosynthesis. This strengthens the performance of the tool that respects the structural similarities, while taking into account, for a V-DOMAIN, the particularities of the IG and TR rearrangement and delimitations of the V-(D)-J-REGION.

The results page for a V-DOMAIN shows the following.

i. At the top (Fig. 2):
   - The sequence name (as provided by the user).
   - The Closest reference gene and allele(s) from the IMGT domain directory (gene and allele name, Species, Domain number, Smith-Waterman Score, V-REGION percentage of identity, and Overlap score). Note that if several closest alleles are identified (same percentage of identity) or if several alignments to display are chosen, the user can select the display of each corresponding alignment.
   - The Alignment(s) with the domain of the closest gene and allele from the IMGT domain directory.
   - The Region(s) and domain(s) identified in your sequence (by comparison with the closest genes and alleles) using IMGT color menu.
   - Sequence with or without gaps in FASTA format (HTML page or downloadable).

ii. Then, the Results summary (by comparison with the closest gene and allele) (Fig. 3):
   - Sequence name, V-REGION identity percentage, CDR-IMGT lengths, Number of different AA in CDR1- and CDR2-IMGT, FR-IMGT lengths, Number of different AA in FR-IMGT, Total number of AA changes in V-DOMAIN.
   - The characteristics of the amino acid changes in FR-IMGT (FR1–FR4) and CDR-IMGT (CDR1 and CDR2) (Lefranc 1999, 2011e; Lefrance et al. 2003), according to the IMGT physicochemical classes (Pommié et al. 2004). Note that the CDR3 that is not relevant for that analysis is not included.

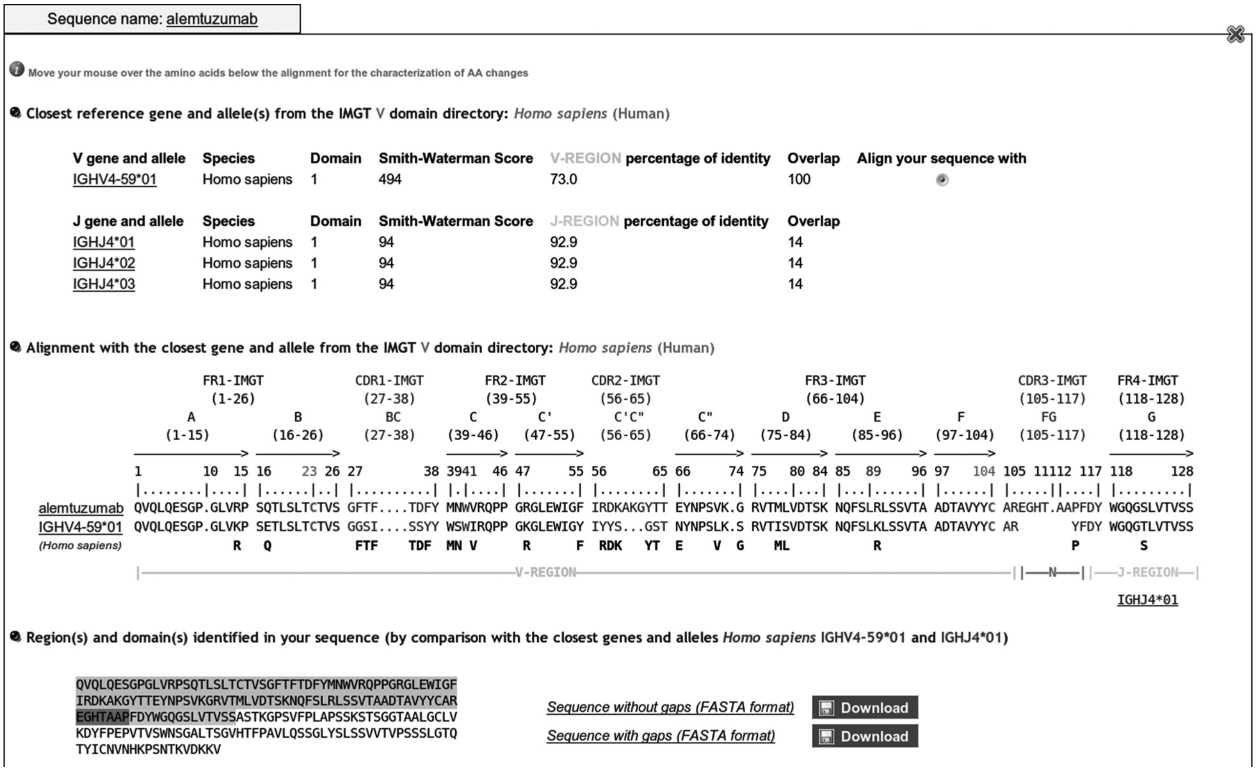

Sequence name: alemtuzumab
Move your mouse over the amino acids below the alignment for the characterization of AA changes
Closest reference gene and allele(s) from the IMGT V domain directory: Homo sapiens (Human)
V gene and allele | Species | Domain | Smith-Waterman Score | V-REGION percentage of identity | Overlap | Align your sequence with
IGHV4-59*01 | Homo sapiens | 1 | 494 | 73.0 | 100
J gene and allele | Species | Domain | Smith-Waterman Score | J-REGION percentage of identity | Overlap
IGHJ4*01 | Homo sapiens | 1 | 94 | 92.9 | 14
IGHJ4*02 | Homo sapiens | 1 | 94 | 92.9 | 14
IGHJ4*03 | Homo sapiens | 1 | 94 | 92.9 | 14
Alignment with the closest gene and allele from the IMGT V domain directory: Homo sapiens (Human)
FR1-IMGT (1-26) CDR1-IMGT (27-38) FR2-IMGT (39-55) CDR2-IMGT (56-65) FR3-IMGT (66-104) CDR3-IMGT (105-117) FR4-IMGT (118-128)
A (1-15) B (16-26) BC (27-38) C (39-46) C' (47-55) C'C" (56-65) C" (66-74) D (75-84) E (85-96) F (97-104) FG (105-117) G (118-128)
1 10 15 16 23 26 27 38 3941 46 47 55 56 65 66 74 75 80 84 85 89 96 97 104 105 11112 117 118 128
alemtuzumab QVQLQESGP.GLVRP SQTLSLTCTVS GFTF....TDFY MNWVRQPP GRGLEWIGF IRDKAKGYTT EYNPSVK.G RVTMLVDTSK NQFSLRLSSVTA ADTAVYYC AREGHT.AAPFDY WGQGSLVTVSS
IGHV4-59*01 QVQLQESGP.GLVKP SETLSLTCTVS GGSI....SSYY WSWIRQPP GKGLEWIGY IYYS...GST NYNPSLK.S RVTISVDTSK NQFSLKLSSVTA ADTAVYYC AR YFDY WGQGTLVTVSS
(Homo sapiens) R Q FTF TDF MN V R F RDK YT E V G ML R P S
V-REGION N J-REGION
IGHJ4*01
Region(s) and domain(s) identified in your sequence (by comparison with the closest genes and alleles Homo sapiens IGHV4-59*01 and IGHJ4*01)
QVQLQESGPGLVRPSQTLSLTCTVSGFTFTDFYMNWVRQPPGRGLEWIGF
IRDKAKGYTTEYNPSVKGRVTMLVDTSKNQFSLRLSSVTAADTAVYYCAR
EGHTAAPFDYWGQGSLVTVSSASTKGPSVFPLAPSSKSTSGGTAALGCLV
KDYFPEPVTVSWNSGALTSGVHTFPAVLQSSGLYSLSSVVTVPSSSLGTQ
TYICNVNHKPSNTKVDKKV
Sequence without gaps (FASTA format) Download
Sequence with gaps (FASTA format) Download

FIGURE 2. IMGT/DomainGapAlign *top* of the results page for a V-DOMAIN. The user V-DOMAIN AA sequence is aligned with the closest germline V-REGION and J-REGION, with IMGT gaps and delimitations of the FR-IMGT and CDR-IMGT according to the IMGT unique numbering (Lefranc et al. 2003). In this example, the user sequence is the V-DOMAIN of the heavy chain (VH) of the monoclonal antibody alemtuzumab. The V-REGION and J-REGION of the alemtuzumab VH is identified as having 73% and 92.9% identity with the *Homo sapiens* IGHV4-59*01 and IGHJ4*01, respectively.

Results summary (by comparison with the closest genes and alleles IGHV4-59*01 and IGHJ4*01)

Sequence name	V-REGION identity percentage	CDR-IMGT lengths	Number of different AA in CDR1- and CDR2-IMGT	FR-IMGT lengths	Number of different AA in FR-IMGT	Total number of AA changes in V-DOMAIN
alemtuzumab	73.0%	[8,10,12]	11	[25,17,38,11] = 91 AA	14	25

CDR-IMGT	Number of different AA	AA changes
CDR1-IMGT (27-38)	6	G28>F (- - -) very dissimilar S29>T (+ - +) similar I30>F (+ - -) dissimilar S35>T (+ - +) similar S36>D (- - -) very dissimilar Y37>F (- + -) dissimilar
CDR2-IMGT (56-65)	5	Y57>R (- - -) very dissimilar Y58>D (- - -) very dissimilar S59>K (- - -) very dissimilar G63>Y (+ - -) dissimilar S64>T (+ - +) similar

FR-IMGT	Number of different AA	AA changes
FR1-IMGT (1-26)	2	K14>R (+ + +) very similar E17>Q (+ + -) similar
FR2-IMGT (39-55)	5	W39>M (+ - -) dissimilar S40>N (- - -) very dissimilar I42>V (+ - +) similar K48>R (+ + +) very similar Y55>F (- + -) dissimilar
FR3-IMGT (66-104)	6	N66>E (+ - -) dissimilar L71>V (+ - +) similar S74>G (+ + -) similar I78>M (+ + -) similar S79>L (- - -) very dissimilar K90>R (+ + +) very similar
FR4-IMGT (118-129)	1	T122>S (+ - +) similar

FIGURE 3. IMGT/DomainGapAlign Results summary for a V-DOMAIN.

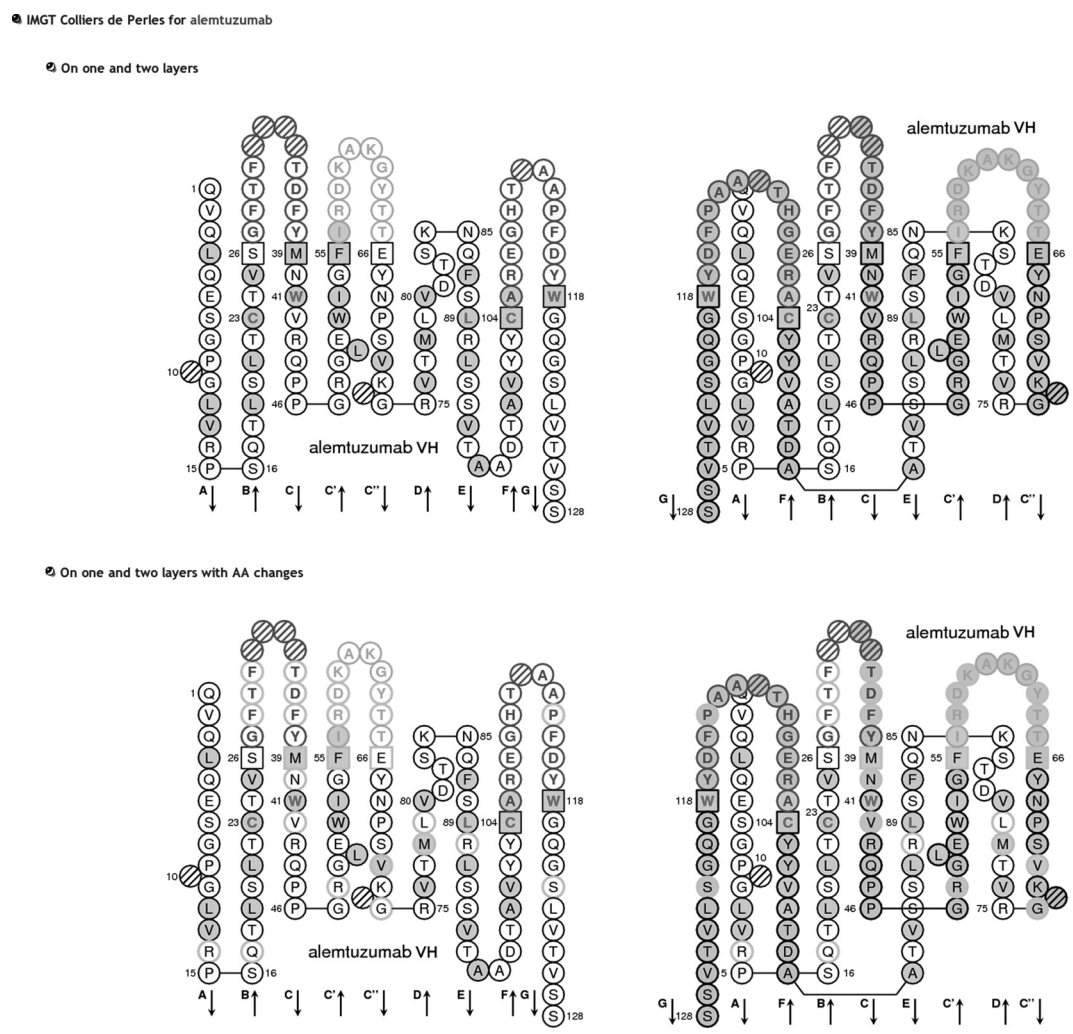


FIGURE 4. IMGT/DomainGapAlign IMGT Colliers de Perles for a V-DOMAIN. IMGT Collier de Perles on one and two layers are on the *left* and *right* side of the figure, respectively. A display of amino acid changes (pink border) is shown in the *bottom* section of the figure. Positions in blue mean that the amino acid of the user sequence at these positions is hydrophobic (hydropathy index with positive value) or is a tryptophan (W), like in 50% or more of analyzed V domains. Positions with red and bold letters indicate the five conserved positions of a V domain: 1st-CYS 23, CONSERVED-TRP 41, hydrophobic 89, 2nd-CYS 104, and J-TRP 118. Anchor positions are in squares. Hatched positions correspond to gaps according to the IMGT unique numbering for V domain (Lefranc 1997, 1999; Lefranc et al. 2003). Prolines are shown in yellow. Arrows indicate the β strands and their direction. The CDR-IMGT lengths of this domain are [8.10.12]. IMGT/DomainGapAlign identified the domain as a VH, and therefore used the corresponding IMGT color menu for CDR-IMGT: red (CDR1-IMGT), orange (CDR2-IMGT), and purple (CDR3-IMGT).

iii. If selected in Step 6, IMGT Colliers de Perles on one and two layers, and with or without AA changes (Fig. 4).

*The results page for a V-LIKE-DOMAIN shows a similar display. As an example, Figure 5 shows the top of the results page for the human CTLA4 aligned with the mouse CTLA4 (selection of "*Mus musculus*" in Step 5). The main difference is that the percentage of identity is calculated for the complete V domain (instead of being calculated for the V-REGION in the case of a V-DOMAIN).*

C Domain Analysis Using IMGT/DomainGapAlign

10. Follow Steps 1–8 above, selecting "C" in Step 4.

The results page for a C domain is similar for a C-DOMAIN or a C-LIKE-DOMAIN.

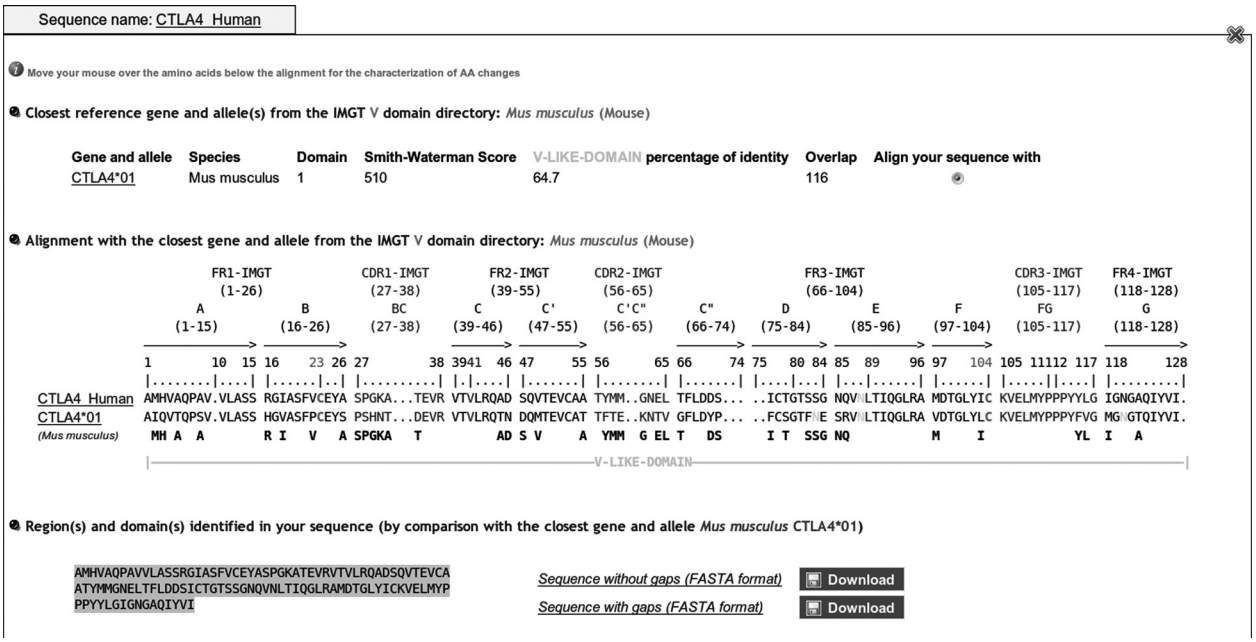


FIGURE 5. IMGT/DomainGapAlign *top* of the results page for a V-LIKE-DOMAIN.

The results page shows the following.

i. At the top (Fig. 6):

- The sequence name (as provided by the user).
- The Closest reference gene and allele(s) from the IMGT domain directory (gene and allele name, Species, Domain number, percentage of identity, C-DOMAIN, or C-LIKE-DOMAIN

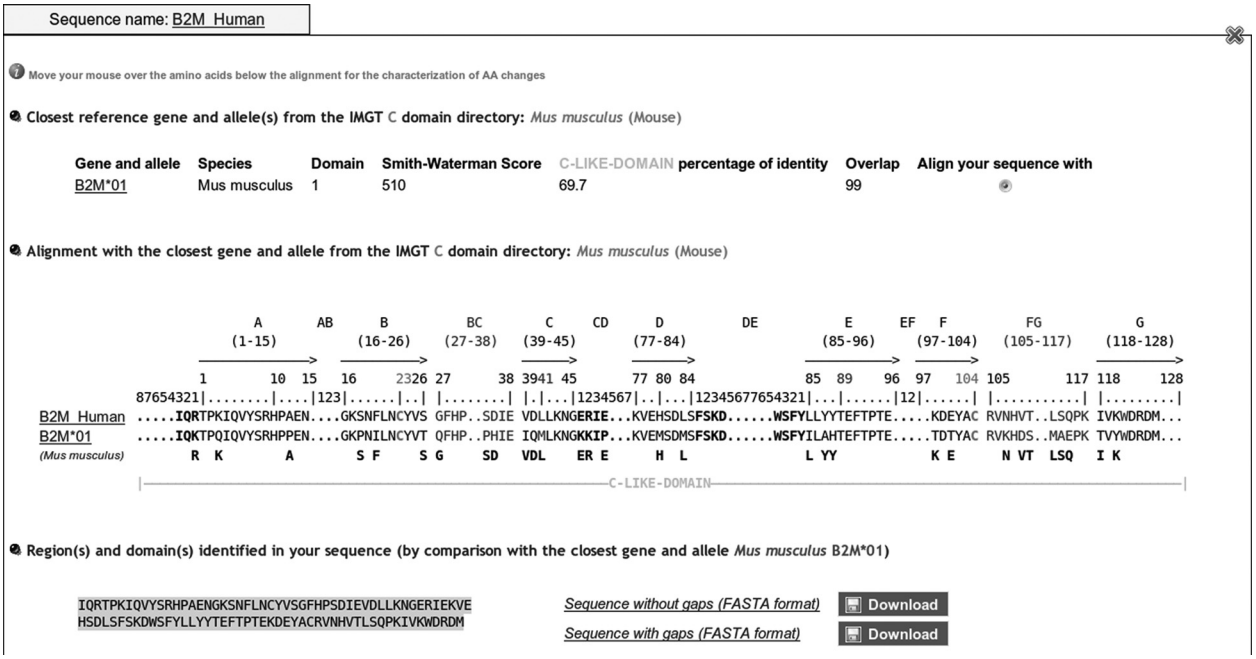


FIGURE 6. IMGT/DomainGapAlign *top* of the results page for a C type domain. In this example, the user sequence is the *Homo sapiens* β-2-microglobulin (B2M) domain. The B2M domain (*Homo sapiens*) is identified as having 69.7% identity with the *Mus musculus* B2M.

percentage of identity and Overlap score). Note that if several closest alleles are identified (same Smith and Waterman score) or if several alignments to display are chosen, the user can select the display of each corresponding alignment.

- The Alignment(s) with the domain of the closest gene and allele from the IMGT domain directory.
- The C-DOMAIN or C-LIKE-DOMAIN identified in your sequence (by comparison with the closest genes and alleles) colored according to IMGT color menu.
- Sequence with or without gaps in FASTA format (HTML page or downloadable).

ii. Then, the Results summary (by comparison with the closest gene and allele) (Fig. 7):

- Sequence name, C-DOMAIN, or C-LIKE-DOMAIN identity percentage, Total number of AA changes.
- The characteristics of the amino acid changes in strands (A, B, C, CD, D, E, F, G), turns (AB, DE, EF) and loops (BC, FG) (Lefranc et al. 2005b; Lefranc 2011e) according to the IMGT physicochemical classes (Pommié et al. 2004).

iii. If selected in Step 6, IMGT Colliers de Perles on one and two layers, and with or without amino acid changes (Fig. 8).

G Domain Analysis Using IMGT/DomainGapAlign

11. Follow Steps 1–8 above, selecting "G" in Step 4.

The results page for a G domain is similar for a G-DOMAIN or a G-LIKE-DOMAIN.

Results summary (by comparison with the closest gene and allele B2M*01)

Sequence name	C-LIKE-DOMAIN identity percentage	Total number of AA changes in C-LIKE-DOMAIN
B2M_Human	69.7%	30

Strands	Number of different AA	AA changes
A strand (1.9-1.1,1-15)	3	K1.1>R (+ + +) very similar Q3>K (+ - -) dissimilar P12>A (- - -) very dissimilar
B strand (16-26)	3	P18>S (+ - -) dissimilar I20>F (+ - -) dissimilar T26>S (+ - +) similar
C strand (39-45)	3	I39>V (+ - +) similar Q40>D (+ - -) dissimilar M41>L (+ + -) similar
CD strand (45.1-45.7)	3	K45.1>E (+ - -) dissimilar K45.2>R (+ + +) very similar P45.4>E (- - -) very dissimilar
D strand (77-84)	2	M80>H (- - -) very dissimilar M83>L (+ + -) similar
E strand (85-96)	3	I85>L (+ + +) very similar A87>Y (- - -) very dissimilar H88>Y (+ - -) dissimilar
F strand (97-104)	2	T99>K (- - -) very dissimilar T101>E (- - -) very dissimilar
G strand (118-128)	2	T118>I (- - -) very dissimilar Y120>K (- - -) very dissimilar

Turns	Number of different AA	AA changes
AB turn (15.1-15.3)	0	-
DE turn (84.1-84.7, 85.7-85.1)	0	-
EF turn (96.1-96.2)	0	-

Loops	Number of different AA	AA changes
BC loop (27-36)	3	Q27>G (- - -) very dissimilar P33>S (+ - -) dissimilar H34>D (- - -) very dissimilar
FG loop (105-117,111.1-111.6,112.1-112.6)	6	K107>N (+ - -) dissimilar D109>V (- - -) very dissimilar S110>T (+ - +) similar M113>L (+ + -) similar A114>S (- + -) dissimilar E115>Q (+ + -) similar

FIGURE 7. IMGT/DomainGapAlign Results summary for a C-LIKE-DOMAIN.

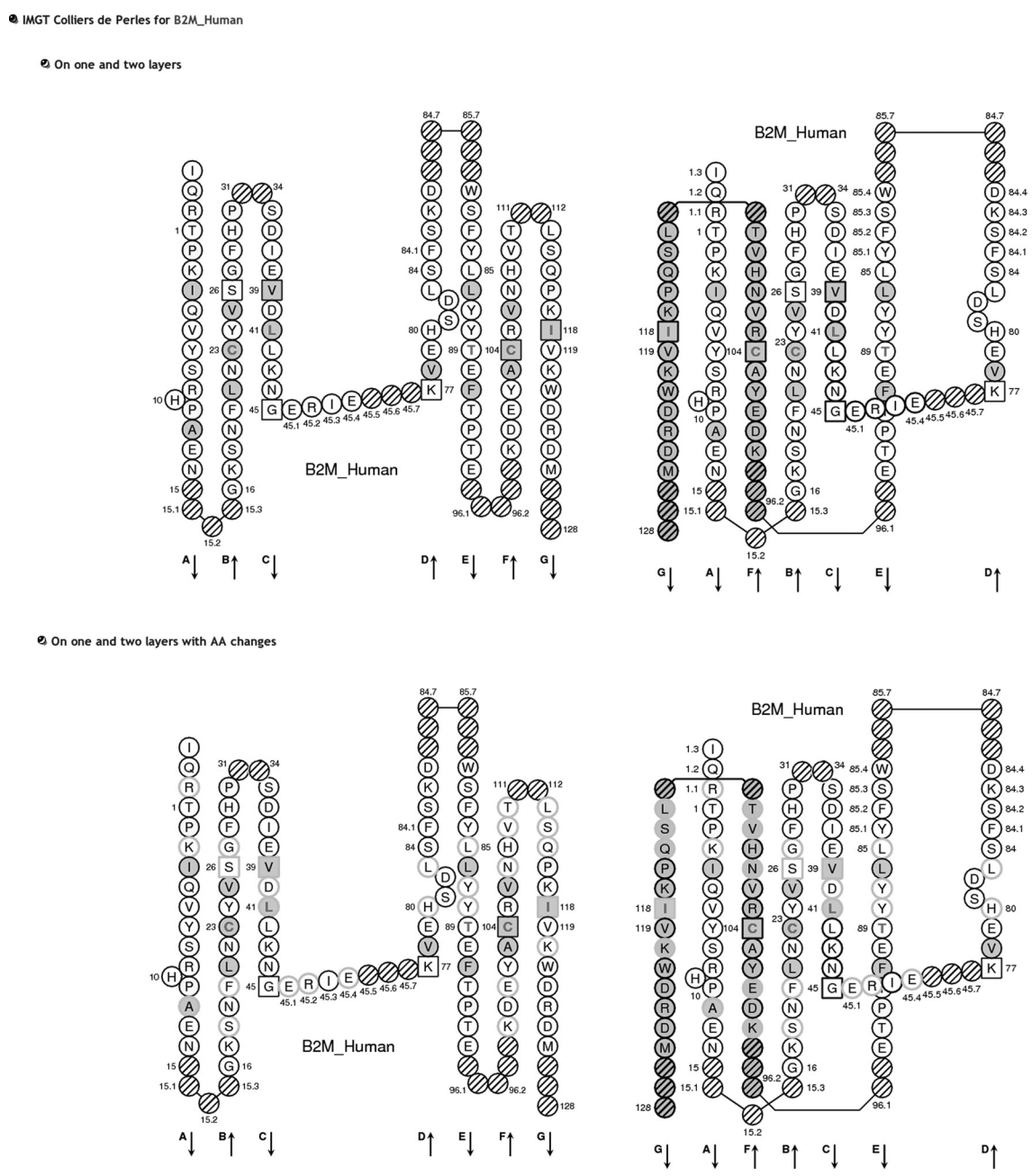


FIGURE 8. IMGT/DomainGapAlign IMGT Colliers de Perles for a C-LIKE-DOMAIN. IMGT Colliers de Perles on one and two layers are on the *left* and *right* side of the figure, respectively. A display of amino acid changes (pink border) is shown in the *bottom* section of the figure. Positions in blue mean that the amino acid of the user sequence at these positions is hydrophobic (hydropathy index with positive value) or is a tryptophan (W), like in 50% or more of analyzed C domains. Positions with red and bold letters are by analogy with a V domain. In a C domain they correspond to the two conserved cysteines, 1st-CYS 23 and 2nd-CYS 104, and to the usually hydrophobic positions 41 and 89 (here 89 is occupied by tyrosine, a neutral amino acid) (Pommié et al. 2004), whereas position 118 is not conserved. Anchor positions are in squares. Hatched positions correspond to gaps according to the IMGT unique numbering for C domain (Lefranc et al. 2005b). Prolines are shown in yellow. Arrows indicate the β strands and their direction.

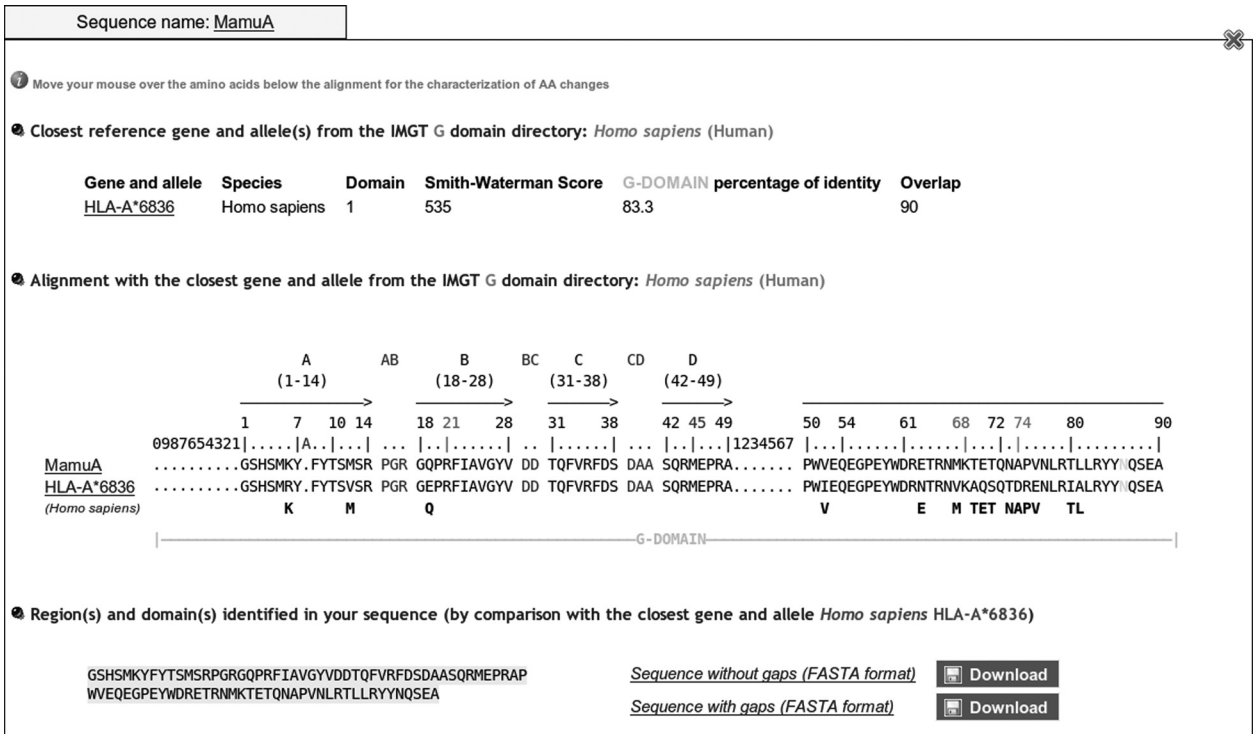


FIGURE 9. IMGT/DomainGapAlign *top* of the results page for a G domain, G-ALPHA1. G-ALPHA1 is the first domain of an MH1 chain. In this example, the user sequence is the G-ALPHA1 of MamuA (Macmul MH1-A) of the rhesus monkey *Macaca mulatta* aligned with the G-ALPHA1 of the closest human HLA-A allele.

The results page shows:

i. At the top (Fig. 9):

- The sequence name (as provided by the user).
- The Closest reference gene and allele(s) from the IMGT domain directory (gene and allele name, Species, Domain number, percentage of identity, G-DOMAIN, or G-LIKE-DOMAIN percentage of identity and Overlap score). Note that if several closest alleles are identified (same Smith and Waterman score) or if several alignments to display are chosen, the user can select the display of each corresponding alignment.
- The Alignment(s) with the domain of the closest gene and allele from the IMGT domain directory.
- The G-DOMAIN or G-LIKE-DOMAIN identified in your sequence (by comparison with the closest genes and alleles) colored according to IMGT color menu.
- Sequence with or without gaps in FASTA format (HTML page or downloadable).

ii. Then, the Results summary (by comparison with the closest gene and allele) (Fig. 10):

- Sequence name, G-DOMAIN or G-LIKE-DOMAIN identity percentage, Total number of AA changes.
- The characteristics of the amino acid changes in strands (A, B, C, D), turns (AB, BC, CD), and in the helix (Lefranc et al. 2005c; Lefranc 2011e), according to the IMGT physico-chemical classes (Pommié et al. 2004).

iii. If selected in Step 6, the IMGT Collier de Perles (Fig. 11).

Results summary (by comparison with the closest gene and allele HLA-A*6836)

Sequence name	G-DOMAIN identity percentage	Total number of AA changes in G-DOMAIN
MamuA	83.3%	15

Turns	Number of different AA	AA changes
AB turn (15-17)	0	-
BC turn (29-30)	0	-
CD turn (39-41)	0	-

Strands	Number of different AA	AA changes
A strand (1.9-1.1,1-14)	2	R6>K (+ + +) very similar V13>M (+ - -) dissimilar
B strand (18-28)	1	E20>Q (+ + -) similar
C strand (31-38)	0	-
D strand (42-49,49.1-49.5)	0	-

Helix	Number of different AA	AA changes
Helix (50-92)	12	I53>V (+ - +) similar N62>E (+ - -) dissimilar V66>M (+ - -) dissimilar A68>T (- - -) very dissimilar Q69>E (+ + -) similar S70>T (+ - +) similar T72>N (- + -) dissimilar D72.>A (- - -) very dissimilar R73>P (- - -) very dissimilar E74>V (- + -) dissimilar I78>T (- - -) very dissimilar A79>L (+ - +) similar

FIGURE 10. IMGT/DomainGapAlign Results summary for a G domain, G-ALPHA1.

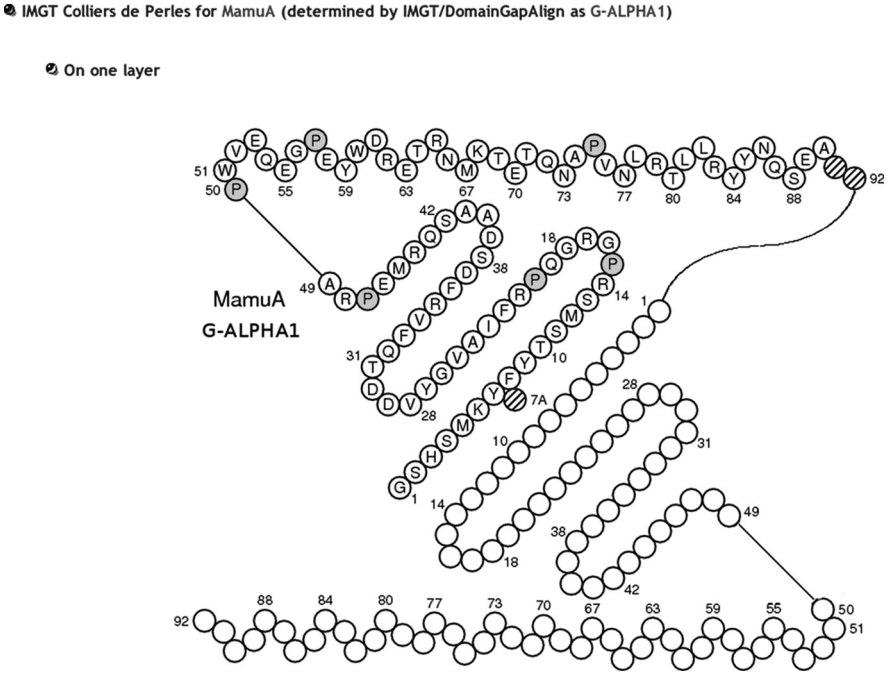


FIGURE 11. IMGT/DomainGapAlign IMGT Collier de Perles for a G domain, G-ALPHA1. A G-ALPHA1 is automatically located in the first half of the figure (for an MH2, the domain automatically displayed in the first half of the figure is G-ALPHA). Hatched positions correspond to gaps according to the IMGT unique numbering for G domain (Lefranc et al. 2005c). Prolines are shown in yellow.

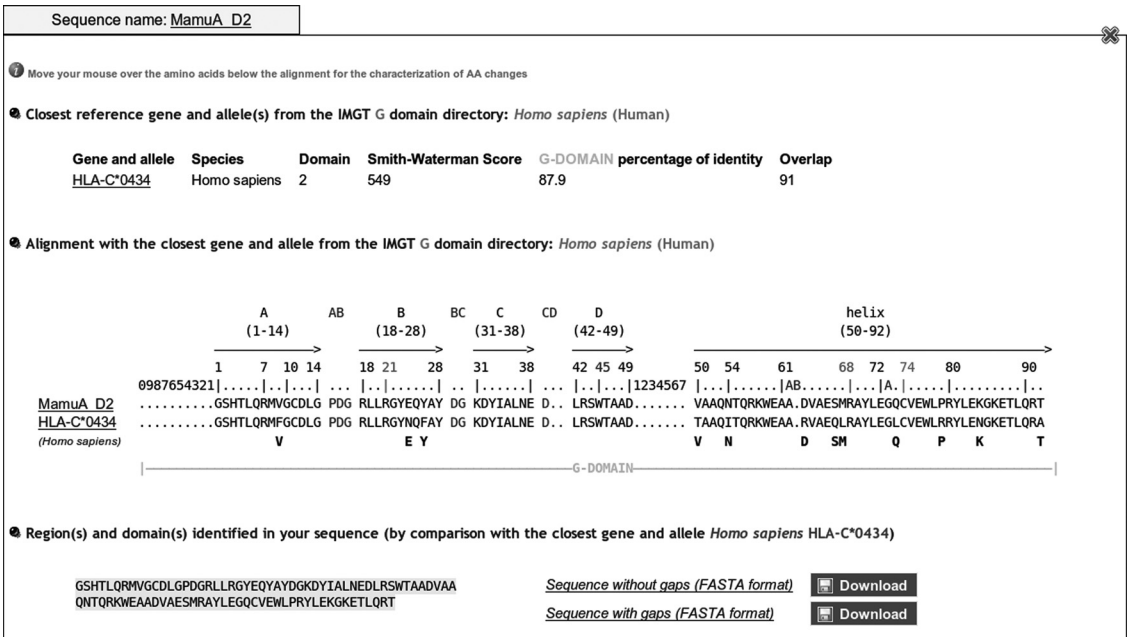


FIGURE 12. IMGT/DomainGapAlign *top* of the results page for a G domain, G-ALPHA2. G-ALPHA2 is the second domain of an MH1 chain. In this example, the user sequence is the G-ALPHA2 of MamuA (Macmul MH1-A) of the rhesus monkey *Macaca mulatta* aligned with the G-ALPHA2 of the closest human HLA-C allele.

Sequence analysis results are similar regardless of the D domain description (G-ALPHA1, G-ALPHA2, G-ALPHA, G-BETA, G-ALPHA1-LIKE, G-ALPHA2-LIKE). Thus, Figure 12 and Figure 13 (G-ALPHA2) are similar to Figure 9 and Figure 10 (G-ALPHA1), respectively. Depending on the G domain description determined by IMGT/DomainGapAlign, the IMGT Collier de Perles is automatically located in the top part for G-ALPHA1 (Fig. 11), G-ALPHA and G-ALPHA1-LIKE (data not shown), or in the bottom part for G-ALPHA2 (Fig. 14), G-BETA, and G-ALPHA2-LIKE (data not shown).

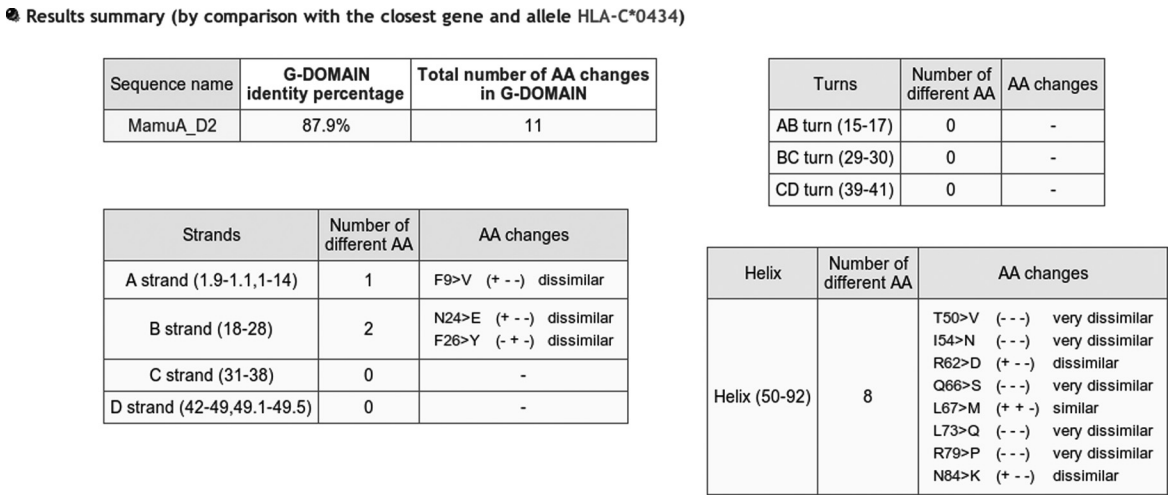


FIGURE 13. IMGT/DomainGapAlign Results summary for a G domain, G-ALPHA2.

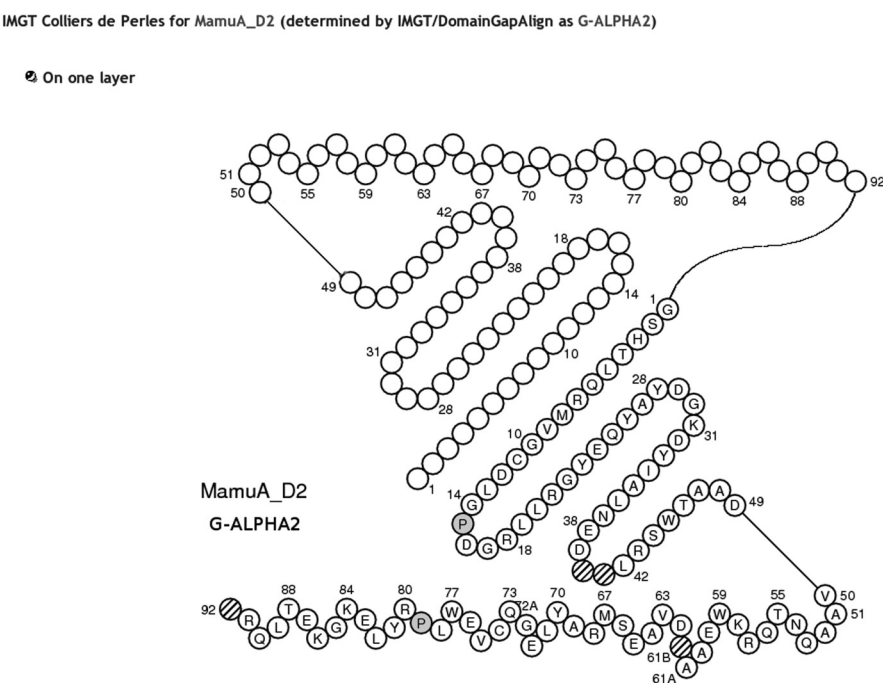


FIGURE 14. IMGT/DomainGapAlign IMGT Collier de Perles for a G domain, G-ALPHA2. A G-ALPHA2 is automatically located in the second half of the figure (for an MH2, the domain automatically displayed in the second half of the figure is G-BETA). Hatched positions correspond to gaps according to the IMGT unique numbering for G domain (Lefranc et al. 2005c). Prolines are shown in yellow.

DISCUSSION

Coupled with the IMGT/Collier de Perles tool, IMGT/DomainGapAlign provides invaluable help for antibody engineering and humanization design based on CDR grafting, as it precisely defines the standardized FR-IMGT and CDR-IMGT (Lefranc 2009; Ehrenmann et al. 2010a). IMGT/DomainGapAlign facilitates the identification of potential immunogenic residues at given positions in chimeric or humanized antibodies, including those of the constant domains (Magdelaine-Beuzelin et al. 2007). The low potential immunogenicity of nonhuman primate antibodies and their potential use as therapeutics in humans has been shown by this approach. This is particularly important for nonhuman primate antibodies neutralizing *Bacillus anthracis* and the anthrax lethal toxin or the ricin toxin that could be used as therapeutics in ricin intoxication and against bioweapons (Pelat et al. 2008, 2009).

ACKNOWLEDGMENTS

We thank Christophe le Roy for checking the protocol. We are grateful to Gérard Lefranc and to the IMGT team for their motivation and expertise.

REFERENCES

Duroux P, Kaas Q, Brochet X, Lane J, Ginestoux C, Lefranc M-P, Giudicelli V. 2008. IMGT-Kaleidoscope, the Formal IMGT-ONTOLOGY paradigm. *Biochimie* **90:** 570–583.

Ehrenmann F, Lefranc M-P. 2011. IMGT/3Dstructure-DB: Querying the IMGT database for 3D structures in immunology and immunoinformatics (IG or antibodies, TR, MH, RPI, and FPIA). *Cold Spring Harb Protoc* doi: 10.1101/pdb.prot5637.

Ehrenmann F, Duroux P, Giudicelli V, Lefranc M-P. 2010a. Standardized sequence and structure analysis of antibody using IMGT. In *Antibody engineering* 2nd ed. (ed. R Kontermann, S Dübel), Vol 2, pp. 11–31. Springer-Verlag, Berlin/Heidelberg, Germany.

Ehrenmann F, Kaas Q, Lefranc M-P. 2010b. IMGT/3Dstructure-DB and IMGT/DomainGapAlign: A database and a tool for immunoglobulins or antibodies, T cell receptors, MHC, IgSF and MhcSF. *Nucleic Acids Res* **38:** D301–D307.

Ehrenmann F, Giudicelli V, Brochet X, Lefranc M-P. 2011. IMGT/Collier de Perles : IMGT standardized representation of domains (IG, TR, and IgSF variable and constant domains, MH and MhSF groove domains). *Cold Spring Harb Protoc* doi: 10.1101/pdb.prot5635.

Giudicelli V, Lefranc M-P. 1999. Ontology for immunogenetics: IMGT-ONTOLOGY. *Bioinformatics* **15:** 1047–1054.

Giudicelli V, Lefranc M-P. 2011. IMGT/JunctionAnalysis: IMGT standardized analysis of the V-J and V-D-J junctions of the rearranged immunoglobulins (IG) and T cell receptors (TR). *Cold Spring Harb Protoc* doi: 10.1101/pdb.prot5634.

Giudicelli V, Chaume D, Lefranc M-P. 2005. IMGT/GENE-DB: A comprehensive database for human and mouse immunoglobulin and T cell receptor genes. *Nucl Acids Res* **33:** D256–D261.

Giudicelli V, Brochet X, Lefranc M-P. 2011. IMGT/V-QUEST: IMGT standardized analysis of the immunoglobulin (IG) and T cell receptor (TR) nucleotide sequences. *Cold Spring Harb Protoc* doi: 10.1101/pdb.prot5633.

Lefranc M-P. 1997. Unique database numbering system for immunogenetic analysis. *Immunol Today* **18:** 509. doi: 10.1016/S0167-5699(97)01163-8.

Lefranc M-P. 1999. The IMGT unique numbering for Immunoglobulins, T cell receptors and Ig-like domains. *The Immunologist* **7:** 132–136.

Lefranc M-P. 2009. Antibody database and tools: The IMGT experience. In *Therapeutic monoclonal antibodies: from Bench to Clinic* (ed. A Zhiqiang), pp. 91–114. John Wiley Sons, Inc, Hoboken, NJ.

Lefranc M-P. 2011a. IMGT, the International ImMunoGeneTics Information System. *Cold Spring Harb Protoc* doi: 10.1101/pdb.top115.

Lefranc M-P. 2011b. From IMGT-ONTOLOGY IDENTIFICATION axiom to IMGT standardized keywords: For immunoglobulins (IG), T cell receptors (TR), and conventional genes. *Cold Spring Harb Protoc* doi: 10.1101/pdb.ip82.

Lefranc M-P. 2011c. From IMGT-ONTOLOGY DESCRIPTION axiom to IMGT standardized labels: For immunoglobulin (IG) and T cell receptor (TR) sequences and structures. *Cold Spring Harb Protoc* doi: 10.1101/pdb.ip83.

Lefranc M-P. 2011d. From IMGT-ONTOLOGY CLASSIFICATION axiom to IMGT standardized gene and allele nomenclature: For immunoglobulins (IG) and T cell receptors (TR). *Cold Spring Harb Protoc* doi: 10.1101/pdb.ip84.

Lefranc M-P. 2011e. IMGT unique numbering for the variable (V), constant (C), and groove (G) domains of IG, TR, MH, IgSF, and MhSF. *Cold Spring Harb Protoc* doi: 10.1101/pdb.ip85.

Lefranc M-P. 2011f. IMGT Collier de Perles for the Variable (V), Constant (C), and Groove (G) Domains of IG, TR, MH, IgSF, and MhSF. *Cold Spring Harb Protoc* doi: 10.1101/pdb.ip86.

Lefranc M-P, Pommié C, Ruiz M, Giudicelli V, Foulquier E, Truong L, Thouvenin-Contet V, Lefranc G. 2003. IMGT unique numbering for immunoglobulin and T cell receptor variable domains and Ig superfamily V-like domains. *Dev Comp Immunol* **27:** 55–77.

Lefranc M-P, Giudicelli V, Ginestoux C, Bosc N, Folch G, Guiraudou D, Jabado-Michaloud J, Magris S, Scaviner D, Thouvenin V, et al. 2004. IMGT-ONTOLOGY for Immunogenetics and Immunoinformatics. *In Silico Biol*, Epub 2003 Nov 22; **4:** 17–29.

Lefranc M-P, Clément O, Kaas Q, Duprat E, Chastellan P, Coelho I, Combres K, Ginestoux C, Giudicelli V, Chaume D, et al. 2005a. IMGT-Choreography for Immunogenetics and Immunoinformatics. *In Silico Biol*, Epub 2004, 5, 0006, Dec 24; **5:** 45–60.

Lefranc M-P, Pommié C, Kaas Q, Duprat E, Bosc N, Guiraudou D, Jean C, Ruiz M, Da Piedade I, Rouard M, et al. 2005b. IMGT unique numbering for immunoglobulin and T cell receptor constant domains and Ig superfamily C-like domains. *Dev Comp Immunol* **29:** 185–203.

Lefranc M-P, Duprat E, Kaas Q, Tranne M, Thiriot A, Lefranc G. 2005c. IMGT unique numbering for MHC groove G-DOMAIN and MHC superfamily (MhcSF) G-LIKE-DOMAIN. *Dev Comp Immunol* **29:** 917–938.

Lefranc M-P, Giudicelli V, Regnier L, Duroux P. 2008. IMGT, a system and an ontology that bridge biological and computational spheres in bioinformatics. *Brief Bioinform* **9:** 263–275.

Lefranc M-P, Giudicelli V, Ginestoux C, Jabado-Michaloud J, Folch G, Bellahcene F, Wu Y, Gemrot E, Brochet X, Lane J, et al. 2009. IMGT, the international ImMunoGeneTics information system. *Nucl Acids Res* **37:** D1006–D1012.

Magdelaine-Beuzelin C, Kaas Q, Wehbi V, Ohresser M, Jefferis R, Lefranc M-P, Watier H. 2007. Structure-function relationships of the variable domains of monoclonal antibodies approved for cancer treatment. *Crit Rev Oncol Hematol* **64:** 210–225.

Pelat T, Bedouelle H, Rees AR, Crennell SJ, Lefranc M-P, Thullier P. 2008. Germline humanization of a non-human Primate antibody that neutralizes the anthrax toxin, by in vitro and in silico engineering. *J Mol Biol* **384:** 1400–1407.

Pelat T, Hust M, Hale M, Lefranc M-P, Dübel S, Thullier P. 2009. Isolation of a human-like antibody fragment (scFv) that neutralizes ricin biological activity. *BMC Biotechnol* **9:** 60. doi: 10.1186/1472-6750-9-60.

Pommié C, Levadoux S, Sabatier R, Lefranc M-P. 2004. IMGT standardized criteria for statistical analysis of immunoglobulin V-REGION amino acid properties. *J Mol Recognit* **17:** 17–32.

Protocol

IMGT/3Dstructure-DB: Querying the IMGT Database for 3D Structures in Immunology and Immunoinformatics (IG or Antibodies, TR, MH, RPI, and FPIA)

François Ehrenmann and Marie-Paule Lefranc[1]

IMGT, the international ImMunoGeneTics information system, Laboratoire d'ImmunoGénétique Moléculaire LIGM, Université Montpellier 2, Institut de Génétique Humaine IGH, UPR CNRS 1142, 34396 Montpellier cedex 5, France

INTRODUCTION

IMGT/3Dstructure-DB is the three-dimensional (3D) structure database of IMGT, the international ImMunoGenetics information system that is acknowledged as the global reference in immunogenetics and immunoinformatics. IMGT/3Dstructure-DB contains 3D structures of (1) antigen receptors that comprise immunoglobulins (IG) or antibodies and T cell receptors (TR), (2) major histocompatibility (MH) proteins of class I (MH1) and class II (MH2), (3) peptide/MH (pMH) complexes (pMH1, pMH2), (4) antigen receptor/antigen complexes (IG/Ag, TR/pMH), and (5) related proteins of the immune system (RPI) belonging to the immunoglobulin and MH superfamilies (IgSF and MhSF, respectively) and found in complexes with IG, TR, or MH. IMGT/3Dstructure-DB data are annotated according to the IMGT criteria, using IMGT/DomainGapAlign, and based on the IMGT-ONTOLOGY axioms and concepts. The use of IMGT/3Dstructure-DB is described here. It provides IMGT gene and allele identification (CLASSIFICATION), region and domain delimitations (DESCRIPTION), and amino acid (AA) positions according to the IMGT unique numbering (NUMEROTATION). IMGT/3Dstructure-DB provides the closest genes and alleles that are expressed in the AA sequences of the 3D structures, by aligning these sequences with the IMGT domain reference directory.

RELATED INFORMATION

IMGT/3Dstructure-DB (Kaas et al. 2004; Ehrenmann et al. 2010a) is the 3D structure database of IMGT, http://www.imgt.org (Lefranc et al. 2009). Standardization and IMGT Scientific chart rules are based on the IMGT-ONTOLOGY concepts of identification, classification, description (Giudicelli and Lefranc 1999; Lefranc et al. 2004, 2005a, 2008) and numerotation (Lefranc 1997, 1999; Lefranc et al. 2003, 2005b,c) generated from the axioms of the Formal IMGT-ONTOLOGY or IMGT-Kaleidoscope (Duroux et al. 2008). The concepts of identification led to the IMGT standardized keywords, the concepts of classification to the IMGT standardized gene and allele names, the concepts of description to the IMGT standardized labels, and the concepts of numerotation to the IMGT unique numbering.

A detailed description of IMGT is provided in **IMGT, the International ImMunoGeneTics Information System** (Lefranc 2011a). Information is also available in **From IMGT-ONTOLOGY IDENTIFICATION Axiom to IMGT Standardized Keywords: For Immunoglobulins (IG), T Cell Receptors (TR), and Conventional Genes** (Lefranc 2011b), **From IMGT-ONTOLOGY DESCRIPTION Axiom to IMGT Standardized Labels: For Immunoglobulin (IG) and T Cell Receptor (TR) Sequences and Structures** (Lefranc 2011c), **From IMGT-ONTOLOGY CLASSIFICATION Axiom to IMGT Standardized Gene and Allele Nomenclature: For Immunoglobulins (IG) and T Cell Receptors (TR)** (Lefranc 2011d), **IMGT Unique Numbering for the Variable (V), Constant (C), and Groove (G) Domains of IG, TR, MH, IgSF, and MhSF** (Lefranc 2011e), and **IMGT Collier de Perles for the Variable (V), Constant (C), and Groove (G) Domains of IG, TR, MH, IgSF, and MhSF** (Lefranc 2011f).

In addition, protocols are available for **IMGT/V-QUEST: IMGT Standardized Analysis of the Immunoglobulin (IG) and T Cell Receptor (TR) Nucleotide Sequences** (Giudicelli et al. 2011), **IMGT/JunctionAnalysis: IMGT Standardized Analysis of the V-J and V-D-J Junctions of the Rearranged**

[1]Corresponding author (Marie-Paule.Lefranc@igh.cnrs.fr).
Cite as: Cold Spring Harb Protoc; 2011; doi:10.1101/pdb.prot5637

Immunoglobulins (IG) and T Cell Receptors (TR) (Giudicelli and Lefranc 2011), **IMGT/Collier de Perles: IMGT Standardized Representation of Domains (IG, TR, and IgSF Variable and Constant Domains, MH and MhSF Groove Domains)** (Ehrenmann et al. 2011), and **IMGT/DomainGapAlign: IMGT Standardized Analysis of Amino Acid Sequences of Variable, Constant, and Groove Domains (IG, TR, MH, IgSF, MhSF)** (Ehrenmann and Lefranc 2011).

MATERIALS

It is essential that you consult the appropriate Material Safety Data Sheets and your institution's Environmental Health and Safety Office for proper handling of equipment and hazardous materials used in this protocol.

Equipment

Computer (Internet-connected).

METHOD

Querying the IMGT/3Dstructure-DB Database

1. Using any modern web browser, go to the IMGT Home page http://www.imgt.org.
2. Click the "IMGT/3Dstructure-DB" link in the IMGT databases section of the IMGT Home page.

   This will take you to the IMGT/3Dstructure-DB database web application (Fig. 1).

   Six sections are available in the IMGT/3Dstructure-DB query page: "Search by Entry code or Molecule name (receptor or ligand)," "Search for complexes," "Search by IMGT entry type using IMGT-ONTOLOGY concepts," "Search by Resolution, Release date or Experimental method," "Search by bibliographical references," and "Chain alignment." These sections are independent of each other, and thus, each user query only concerns one of these six sections. Choose

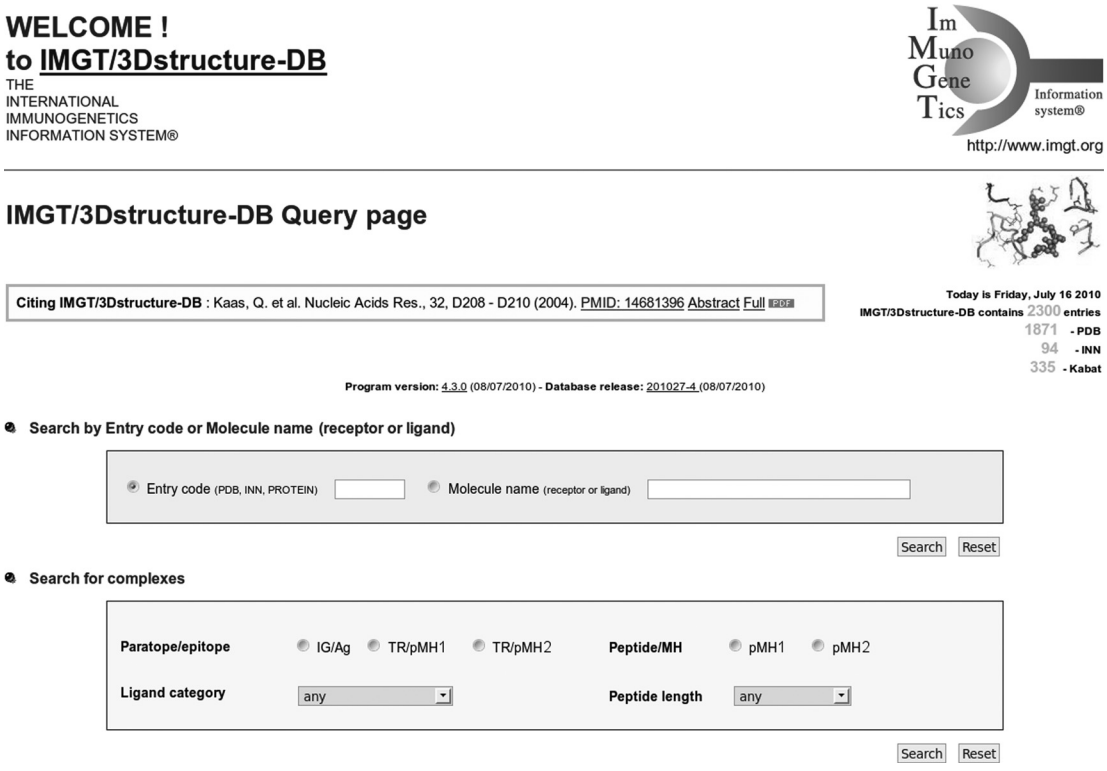


FIGURE 1. IMGT/3Dstructure-DB Welcome and Query page.

selection A (Steps 3–5), B (Steps 6–9), C (Steps 10–13), D (Steps 14–17), or E (Steps 18–21) and then continue to Step 24, or selection F (Steps 22–23).

Selection A: Search by Entry Code or Molecule Name

3. Locate the "Entry code" or the "Molecule name (receptor or ligand)" text areas.

4. Enter an entry code (PDB, INN, PROTEIN) in "Entry code" text area or a Molecule name (receptor or ligand) in "Molecule name" text area.

   The IMGT/3Dstructure-DB entry code (ID) is identical to either the PDB code, INN code, or PROTEIN code. PDB code is an alphanumeric code (ex: 1ao7), INN code is a numeric code (ex: 7637), while PROTEIN code is an alphanumeric code (ex : p00054). An AutoComplete behavior is available for this field. AutoComplete works in the following manner: as you type in the beginning of a PDB code, previous entries are pulled from a storage area and you may elect to simply select one of these entries. By selecting an entry, you no longer have to input the string because AutoComplete finishes entering the string for you.

   An IMGT molecule name is a name retrieved from the literature, a name modified by IMGT, or a name added by IMGT. The search by IMGT molecule name is case insensitive.

5. Proceed to Step 24.

Selection B: Search for Complexes

6. Locate the "Paratope/epitope" or "Peptide/MH" radio buttons, and select complexes you want to retrieve.

   Search by Paratope/epitope can retrieve entries containing the following complexes:

   IG/Ag: Immunoglobulin/antigen
   TR/pMH1: T cell Receptor/peptide/MH1
   TR/pMH2: T cell Receptor/peptide/MH2

   Search by Peptide/MH can retrieve entries containing the following complexes:

   pMH1: peptide/MH1
   pMH2 peptide/MH2

7. Locate the "Ligand category" radio button, and select the ligand category for which you want to retrieve the corresponding complexes.

   Search by Ligand category can retrieve entries with ligands that can belong to the following categories: Chemical compound, DNA, Nucleotide, Peptide, Protein, or RNA.

8. Select a peptide length if needed.

   Selecting a peptide length automatically enters the "Peptide" choice in the "Ligand category" drop-down list.

9. Proceed to Step 24.

Selection C: Search by IMGT Entry Type Using IMGT-ONTOLOGY Concepts

10. Locate the "IMGT entry type" radio button, select an entry type (PDB, INN, Kabat, or any) (Fig. 2).

11. Locate the three sections "IDENTIFICATION," "DESCRIPTION," and "CLASSIFICATION."

    Selecting an "IMGT receptor type" (e.g., "IG") will update the "IMGT receptor description" list, so that it only contains IMGT receptor description labels for the "IG." In the same way, selecting an "IMGT group" (e.g., "TRAV") will update the "IMGT subgroup" list, so that it only contains IMGT subgroup(s) from the IMGT "TRAV" group.

12. Select the criteria as needed:

    i. In "IDENTIFICATION:"

       Species
       IMGT receptor type (IG, TR, MH, RPI, FPIA, or any)

       Options menu (MH1, MH2, or RPI-MH1Like) allows you to refine your research.

    ii. In "DESCRIPTION:"

       IMGT receptor description: IG, TR, MH, or RPI receptors are described with standardized IMGT labels.

       Options menu (FV, SCFV, FAB, or FC) allows you to refine your research.

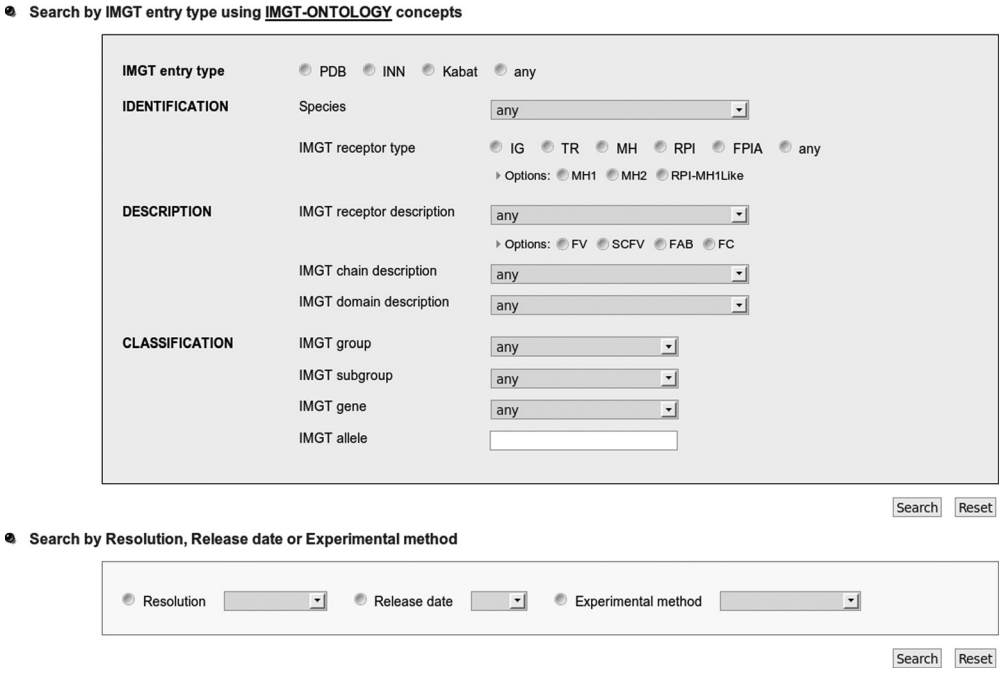


FIGURE 2. IMGT/3Dstructure-DB Query page: Search by IMGT entry type using IMGT-ONTOLOGY concepts.

IMGT chain description: IMGT IG, TR, MH, or RPI chains are described with standardized IMGT labels (Lefranc 2010c).

iii. In "CLASSIFICATION:"

IMGT group
IMGT subgroup
IMGT gene
IMGT allele

Human IG (Lefranc and Lefranc 2001a) and TR (Lefranc and Lefranc 2001b) genes and alleles are from IMGT/GENE-DB (Giudicelli et al. 2005) with translation in IMGT/DomainDisplay (Ehrenmann et al. 2010b). Genes and alleles from other species and/or other proteins are added when the genome of these species is annotated and nomenclature can be set up (Lefranc 2010d), and/or when 3D structures are available in IMGT/3Dstructure-DB.

13. Proceed to Step 24.

Selection D: Search by Resolution, Release Date, or Experimental Method

14. Locate the "Resolution" radio button (Fig. 2), and select resolution for which you want to retrieve structures (only for PDB entries).

Six value ranges are available: 0–1.00 Å, 1.01–2.00 Å, 2.01–3.00 Å, 3.01–4.00 Å, 4.01–5.00 Å, and 5.01–6.00 Å.

15. Locate the "Release date" radio button, and select PDB release date for which you want to retrieve structures (only for PDB entries).

16. Locate the "Experimental method" radio button, and select PDB experimental method for which you want to retrieve structures (only for PDB entries).

Users can choose among nine experimental methods: Electron microscopy; NMR; NMR, 18 structures; NMR, 20 structures; Solution scattering; Solution scattering, theoretical model; Theoretical model; X-ray diffraction; or X-ray diffraction, single crystal.

17. Proceed to Step 24.

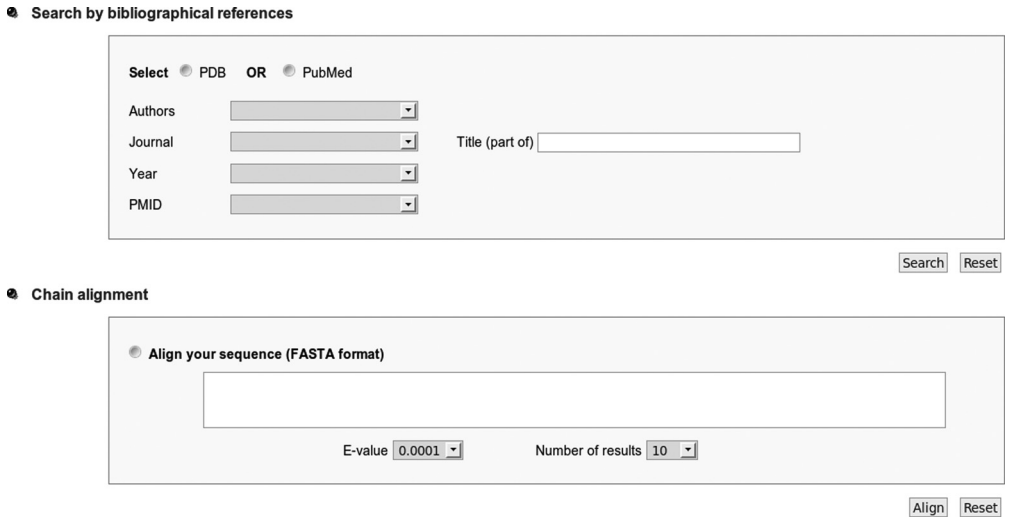


FIGURE 3. IMGT/3Dstructure-DB Query page: Search by bibliographical references, and Chain alignment.

Selection E: Search by Bibliographical References

18. Locate the section "Search by bibliographical references" (Fig. 3) and select the type of reference: PDB or PubMed.
19. Then, locate the drop-down lists below and select an author and/or a journal and/or a year and/or a PMID (for PubMed references).

    Again, this kind of search uses linked select elements. For example, selecting an author will update the "Journal" drop-down list to only contain Journal(s) from the author selected.
20. The research can also be performed on the title of publications (part of).
21. Proceed to Step 24.

Selection F: Chain Alignment

22. Locate the "Align your sequence" radio button (Fig. 3), click on it, and paste your sequence in FASTA format in the text area.

    Your own sequence will be aligned using FASTA (Pearson and Lipman 1988) against the IMGT/3Dstructure-DB sequences with the "FASTA chain alignment" tool.

    The E-value can be modified and users can choose the numbers of results to display. For instance, if there are no results in "Chain Alignment results," you may try changing the E-value from 0.0001 to 0.001. Results can also be obtained with peptides of 11 AA (or even less) with an E-value of 0.5.
23. Click on Align. This will give you the Chain alignment results.

IMGT/3Dstructure-DB Visualization Choice for the Results

24. Locate the "Display results" section at the bottom of the IMGT/3Dstructure-DB Query page (Fig. 4).

    There are three possible results visualization (select by clicking on a radio button):

    - *"Overview" (selected by default)*
    - *"Domain type sequences"*
    - *"FR-IMGT/CDR-IMGT sequences and lengths" (only for V domains)*
25. Select, if needed, the type of display and the corresponding views.
26. Click on Search.

FIGURE 4. IMGT/3Dstructure-DB Query page: Display results section (for choice of the results visualization).

IMGT/3Dstructure-DB Display Results

27. You will get one of the possible visualizations: "Overview" (Fig. 5) (by default or after selection of some views), "Domain type sequences" (Fig. 6), or "FR-IMGT and/or CDR-IMGT sequences" (Fig. 7), depending on your selection.

    At the top of the results page, the visualization choice and the query are recalled, and the number of results is shown. For results that correspond to a list of chains (and not to a list of entries), it is recalled, between parentheses, that the number of results corresponds to the "number of chains containing the selected domain type." The list of results is sorted by ID (entry or chain).

THANK YOU

for using IMGT/3Dstructure-DB

THE
INTERNATIONAL
IMMUNOGENETICS
INFORMATION SYSTEM®

Im Muno Gene Tics Information system®

http://www.imgt.org

Overview

Your query: Molecule name (receptor or ligand) HIV

Number of results: 111

Click on IMGT entry ID (2nd column) for entry card

	IMGT entry ID	IMGT molecule name	IMGT entry type	IMGT receptor description	Species	Ligand(s)
1	1acy	59.1 mAb, anti-gp120 [HIV-1]	PDB	FAB-GAMMA-1_KAPPA	*Mus musculus*	Envelope glycoprotein gp120 [Human immunodeficiency virus 1, HIV-1]
2	1afv	25.3 mAb, anti-p24 [HIV-1]	PDB	FAB-GAMMA-1_KAPPA	*Mus musculus*	Capsid protein p24 (residues: 1-151) [Human immunodeficiency virus 1, HIV-1]
3	1cfn	Cb41 mAb, anti-p24 [HIV-1]	PDB	FAB-GAMMA-2C_KAPPA	*Mus musculus*	Epitope-related peptide p24 [Human immunodeficiency virus 1, HIV-1]
4	1cfs	Cb41 mAb, anti-p24 [HIV-1]	PDB	FAB-GAMMA-2C_KAPPA	*Mus musculus*	Unrelated peptide p24 [Human immunodeficiency virus 1, HIV-1]
5	1cft	Cb41 mAb, anti-p24 [HIV-1]	PDB	FAB-GAMMA-2C_KAPPA	*Mus musculus*	Unrelated D-peptide p24 [Human immunodeficiency virus 1, HIV-1]
6	1e6j	13B5 mAb, anti-p24 [HIV-1]	PDB	FAB-GAMMA-1_KAPPA	*Mus musculus*	Capsid protein p24 (residues: 142-351) [Human immunodeficiency virus 1, HIV-1]
7	1f58	58.2 mAb, anti-gp120 [HIV-1]	PDB	FAB-GAMMA-1_KAPPA	*Mus musculus*	Envelope glycoprotein gp120 [Human immunodeficiency virus 1, HIV-1]
8	1g9m	17B mAb, anti-gp120 [HIV-1] CD4	PDB	FAB-GAMMA-1_KAPPA 1V-1C-LIKE	*Homo sapiens*	Envelope glycoprotein gp120 [Human immunodeficiency virus 1, HIV-1] 2-Propanol, Isopropanol
9	1g9n	17B mAb, anti-gp120 [HIV-1] CD4	PDB	FAB-GAMMA-1_KAPPA 1V-1C-LIKE	*Homo sapiens*	Envelope glycoprotein gp120 [Human immunodeficiency virus 1, HIV-1]
10	1gc1	17B mAb, anti-gp120 [HIV-1] CD4	PDB	FAB-GAMMA-1_KAPPA 1V-1C-LIKE	*Homo sapiens*	Envelope glycoprotein gp120 [Human immunodeficiency virus 1, HIV-1]
11	1ggi	50.1 mAb, anti-gp120 [HIV-1]	PDB	FAB-GAMMA-2A_KAPPA	*Mus musculus*	Envelope glycoprotein gp120 [Human immunodeficiency virus 1, HIV-1]
12	1hh6	Cb41 mAb, anti-p24 [HIV-1]	PDB	FAB-GAMMA-2C_KAPPA	*Mus musculus*	Epitope-related peptide p24 [Human immunodeficiency virus 1, HIV-1]
13	1hh9	Cb41 mAb, anti-p24 [HIV-1]	PDB	FAB-GAMMA-2C_KAPPA	*Mus musculus*	Epitope-related peptide p24 [Human immunodeficiency virus 1, HIV-1]
14	1hi6	Cb41 mAb, anti-p24 [HIV-1]	PDB	FAB-GAMMA-2C_KAPPA	*Mus musculus*	Epitope-related peptide p24 [Human immunodeficiency virus 1, HIV-1]
15	1hys	Fab-28 mAb anti-reverse transcriptase [HIV-1]	PDB	FAB-GAMMA-1_KAPPA	*Mus musculus*	Polypurine tract (PPT) RNA: DNA [Human immunodeficiency virus 1, HIV-1] Reverse transcriptase (RT) EC: 2.7.7.49 [Human immunodeficiency virus 1, HIV-1]
16	1hzh	b12 neutralizing mAb, anti-gp120 [HIV-1]	PDB	IG-GAMMA-1_KAPPA	*Homo sapiens*	
17	1iif	Cb41 mAb, anti-p24 [HIV-1]	PDB	FV-HEAVY_KAPPA	*Mus musculus*	Derived peptide p24 [Human immunodeficiency virus 1, HIV-1]
18	1j5o	Fab28 mAb, anti-reverse transcriptase [HIV-1]	PDB	FAB-GAMMA-1_KAPPA	*Mus musculus*	Polydeoxyribonucleotide [Synthetic] Reverse transcriptase (RT), M184>I, EC:2.7.7.49 [Human immunodeficiency virus 1, HIV-1]
19	1jp5	1696 mAb, anti-protease [HIV-1]	PDB	SCFV-HEAVY-KAPPA	*Mus musculus*	N-terminus of protease [Human immunodeficiency virus 1, HIV-1]

FIGURE 5. IMGT/3Dstructure-DB "Overview" display of results. The "Overview" results provide IMGT entry ID, IMGT molecule name, IMGT entry type, IMGT receptor description, Species, Ligand(s), Experimental technique, Resolution, and release date (only for PDB).

THANK YOU

for using IMGT/3Dstructure-DB

THE
INTERNATIONAL
IMMUNOGENETICS
INFORMATION SYSTEM®

IMGT Information system®

http://www.imgt.org

V domain sequences

Your query: Species Homo sapiens and Entry type PDB and Receptor type IG

Number of results: 939
(number of chains containing the selected domain type).

Click on IMGT entry ID (2nd column) for entry card

	IMGT chain ID	Species	V type domain
			FR1-IMGT (1-26) A (1-15) B (16-26) CDR1-IMGT (27-38) BC (27-38) FR2-IMGT (39-55) C (39-46) C' (47-55) CDR2-IMGT (56-65) C'C" (56-65) C" (66-74) FR3-IMGT (66-104) D (75-84) E (85-96) F (97-104) CDR3-IMGT (105-117) FG (105-117) FR4-IMGT (118-128) G (118-128)
			1 10 15 16 23 26 27 38 39 41 46 47 55 56 65 66 74 75 80 84 85 89 96 97 104 105 111 112 117 118 128
1	1a8j_H	*Homo sapiens*	.SALTQPPS.ASGSL GQSVTISCTGT SSDVG...GYNY VSWYQQHA GKAPKVIIY EV.....N KRPSGVP.D RFSGSK..SG NTASLTVSGLQA EDEADYYC SSYEG.....SDNFV FGTGTKVTVL....
2	1a8j_L	*Homo sapiens*	.SALTQPPS.ASGSL GQSVTISCTGT SSDVG...GYNY VSWYQQHA GKAPKVIIY EV.....N KRPSGVP.D RFSGSK..SG NTASLTVSGLQA EDEADYYC SSYEG.....SDNFV FGTGTKVTVL....
3	1adq_H	*Homo sapiens*	EVQLVESGG.GLVQP GRSLRLSCVTS GFTF....DDYA MHWVRQSP GKGLEWVSG ISWN..TGTI IYADSVK.G RFIISRDNAK NSLYLQMNSLRV EDTALYYC AKTRSYVV.....AAEYYFHY WGQGILVTVSS...
4	1adq_L	*Homo sapiens*	.YVLTQPPS.VSVAP GQTARITCGGN NIG.....SKS VHWYQQKP GQAPVLVVY DD.....S DRPPGIP.E RFSGSN..SG NTATLTISRVEA GDEADYYC QVWDSS.....SDHAV FGGGTKLTVL....
5	1aqk_H	*Homo sapiens*	.VQLVESGG.GVVQP GRSLRLSCAAS GFTF....NNYA IHWVRQAP GKGLEWVAF ISYD..GSKN YYADSVK.G RFTISRDNSK NTLFLQMNSLRP EDTAIYYC ARVLFQQL.....VLYAPFDI WGQGTMVTVSS...
6	1aqk_L	*Homo sapiens*	.NVLTQPPS.VSGAP GQRVTISCTGS NSNIG...AGFT VHWYQHLP GTAPKLLIF AN.....T NRPSGVP.D RFSGSK..SG TSASLAITGLQA EDEADYYC QSYDS.....SLSAR FGGGTRLTVL....
7	1ar2_-	*Homo sapiens*	DIQMTQSPSSLSASV GDRVTITVQAS QDI.....IKH LNWYQQTP GKAPKLLIY EA.....S NLQAGVP.S RFSGSG..SG TDYTFTISSLQP EDIATYYC QQYQS.....LPYT FGQGTKLQI.....
8	1b0w_A	*Homo sapiens*	DIQMTQSPSSLSASV GDRVTITCQAS QDI.....SDY LIWYQQKL GKAPNLLIY DA.....S TLETGVP.S RFSGSG..SG TEYTFTISSLQP EDIATYYC QQYDD.....LPYT FGQGTKVEIK....
9	1b0w_B	*Homo sapiens*	DIQMTQSPSSLSASV GDRVTITCQAS QDI.....SDY LIWYQQKL GKAPNLLIY DA.....S TLETGVP.S RFSGSG..SG TEYTFTISSLQP EDIATYYC QQYDD.....LPYT FGQGTKVEIK....
10	1b0w_C	*Homo sapiens*	DIQMTQSPSSLSASV GDRVTITCQAS QDI.....SDY LIWYQQKL GKAPNLLIY DA.....S TLETGVP.S RFSGSG..SG TEYTFTISSLQP EDIATYYC QQYDD.....LPYT FGQGTKVEIK....
11	1b6d_A	*Homo sapiens*	DIQMTQSPSSLSASV GDRVTITCQAS QDI.....SSY LNWYQQKP GKAPKLLIH AA.....S SLETGVP.S RFSGSG..SG TDFSFTISSLQP EDLATYYC QQYDS.....LPLT FGGGTKVEIK....
12	1b6d_B	*Homo sapiens*	DIQMTQSPSSLSASV GDRVTITCQAS QDI.....SSY LNWYQQKP GKAPKLLIH AA.....S SLETGVP.S RFSGSG..SG TDFSFTISSLQP EDLATYYC QQYDS.....LPLT FGGGTKVEIK....
13	1bjm_A	*Homo sapiens*	.SVLTQPPS.ASGTP GQRVTISCSGS SSNI....GENS VTWYQHLS GTAPKLLIY ED.....N SRASGVS.D RFSASK..SG TSASLAISGLQP EDETDYYC AAWDDS.....LDVAV FGTGTKVTVL....
14	1bjm_B	*Homo sapiens*	.SVLTQPPS.ASGTP GQRVTISCSGS SSNI....GENS VTWYQHLS GTAPKLLIY ED.....N SRASGVS.D RFSASK..SG TSASLAISGLQP EDETDYYC AAWDDS.....LDVAV FGTGTKVTVL....
15	1bre_A	*Homo sapiens*	DIQMTQSPSSLSASV GDRVTITCQAS QDI.....SDY LIWYQQKL GKAPNLLIY DA.....S TLETGVP.S RFSGSG..SG TEYTFTISSLQP EDIATYYC QQYDD.....LPYT FGQGTKVEIK....

FIGURE 6. IMGT/3Dstructure-DB "Domain type sequences" display of results. This visualization provides the sequences of the domains of the IMGT/3Dstructure-DB chains that satisfy the query.

28. Click on the IMGT entry ID or IMGT chain ID to get access to the individual IMGT 3Dstructure-DB cards.

IMGT/3Dstructure-DB card-Header

29. Locate the IMGT/3Dstructure-DB card header. The header comprises the following.

i. The summary table of the entry:

IMGT molecule name
IMGT receptor type
IMGT receptor description
Ligands
Species
Crystallographic complex (CC)
Chain ID

ii. Below the summary table:

Experimental technique
Resolution (in angstroms)
PDB release date

iii. Seven or eight (if Paratope and epitope are present) tabs that give access to:

Chain details (open by default; see Step 30).
Contact analysis (if you click on the tab, proceed to Step 33)
Paratope and epitope

THANK YOU

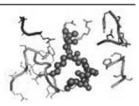
for using IMGT/3Dstructure-DB

THE
INTERNATIONAL
IMMUNOGENETICS
INFORMATION SYSTEM®

Im Muno Gene Tics Information system®

http://www.imgt.org

FR-IMGT and/or CDR-IMGT sequences



Your query: Species Homo sapiens and Entry type PDB and Receptor type IG and IMGT group IGHV

Number of results: 415
(number of chains containing the selected domain type).

Click on IMGT entry ID (2nd column) for entry card

	IMGT chain ID	Species	CDR3-IMGT sequence	CDR3-IMGT length
1	1adq_H	*Homo sapiens*	AKTRSYVVAAEYYFHY	16
2	1aqk_H	*Homo sapiens*	ARVLFQQLVLYAPFDI	16
3	1dee_B	*Homo sapiens*	AKVKFYDPTAPNDY	14
4	1dee_D	*Homo sapiens*	AKVKFYDPTAPNDY	14
5	1dee_F	*Homo sapiens*	AKVKFYDPTAPNDY	14
6	1dfb_H	*Homo sapiens*	VKGRDYYDSGGYFTVAFDI	19
7	1dn0_B	*Homo sapiens*	ARPPHDTSGHYWNY	14
8	1dn0_D	*Homo sapiens*	ARPPHDTSGHYWNY	14
9	1dql_H	*Homo sapiens*	ARGNPPYSSGWGGGDY	16
10	1dx3_H	*Homo sapiens*	ARFAIKGDY	9
11	1fl6_B	*Homo sapiens*	ARDGSYAMDY	10
12	1fl6_H	*Homo sapiens*	ARDGSYAMDY	10
13	1g9m_H	*Homo sapiens*	AGVYEGEADEGEYRNNGFLKH	21
14	1g9n_H	*Homo sapiens*	AGVYEGEADEGEYRNNGFLKH	21
15	1gc1_H	*Homo sapiens*	AGVYEGEADEGEYDNNGFLKH	21
16	1hez_B	*Homo sapiens*	AKVKFYDPTAPNDY	14
17	1hez_D	*Homo sapiens*	AKVKFYDPTAPNDY	14
18	1hou_H	*Homo sapiens*	AKGYGMDV	8
19	1hzh_H	*Homo sapiens*	ARVGPYSWDDSPQDNYYMDV	20
20	1hzh_K	*Homo sapiens*	ARVGPYSWDDSPQDNYYMDV	20
21	1i9r_H	*Homo sapiens*	TRSDGRNDMDS	11
22	1i9r_K	*Homo sapiens*	TRSDGRNDMDS	11
23	1i9r_X	*Homo sapiens*	TRSDGRNDMDS	11
24	1iga_A	*Homo sapiens*	CARDPYGGGKSEFDY	15
25	1iga_B	*Homo sapiens*	CARDPYGGGKSEFDY	15

FIGURE 7. IMGT/3Dstructure-DB "FR-IMGT and/or CDR-IMGT sequences" display of results. This visualization provides the sequences and lengths of the FR1-IMGT, FR2-IMGT, FR3-IMGT, FR4-IMGT, CDR1-IMGT, CDR2-IMGT, or CDR3-IMGT of the V domains of the IMGT/3Dstructure-DB chains that satisfy the query.

3D visualization (JMOL or QuickPDF; only available for PDB entries)
Renumbered IMGT file (access to the renumbered coordinate file according to the IMGT unique numbering; the "download" link directly downloads the renumbered IMGT file [in gzip format])
INN definitions and properties (only available f or INN entries)
IMGT numbering comparison (comparison of the residue numbering between the IMGT file and the PDB file)
Reference and links
Printable card

IMGT/3Dstructure-DB Card – Chain Details

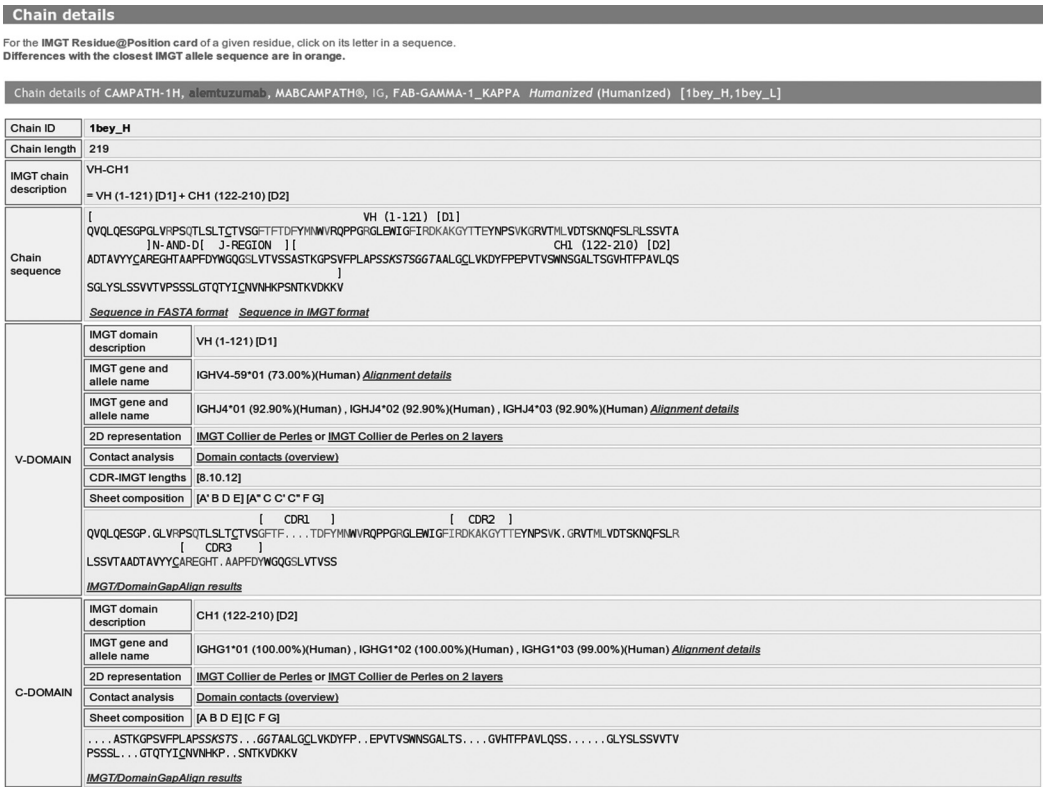
30. Locate in the "IMGT/3Dstructure-DB card," the section "Chain details" (open by default) (Fig. 8) that comprises information on the chain itself, then per domain.

The information for each chain includes:

- *Chain ID (for example 1bey_H)*
- *Chain length in AA (for example 219)*
- *IMGT chain description with the delimitations of the different domains (for example VH-CH1 = VH(1–121)[D1] + CH1(122–210)[CH2])*
- *Chain sequence with delimitations of the regions and domains, highlighting of amino acids that are different from the closest genes and alleles.*

The information for each V domain, as an example, includes:

- *IMGT domain description with the delimitations of the domain (for example VH(1–121)[D1])*



Chain details

For the **IMGT Residue@Position card** of a given residue, click on its letter in a sequence.
Differences with the closest IMGT allele sequence are in orange.

Chain details of CAMPATH-1H, alemtuzumab, MABCAMPATH®, IG, FAB-GAMMA-1_KAPPA *Humanized* (Humanized) [1bey_H, 1bey_L]

Chain ID	**1bey_H**
Chain length	219
IMGT chain description	VH-CH1 = VH (1-121) [D1] + CH1 (122-210) [D2]
Chain sequence	[VH (1-121) [D1] QVQLQESGPGLVRPSQTLSLTCTVSGFTFTDFYMNWVRQPPGRGLEWIGFIRDKAKGYTTEYNPSVKGRVTMLVDTSKNQFSLRLSSVTA]N-AND-D[J-REGION][CH1 (122-210) [D2] ADTAVYYCAREGHTAAPFDYWGQGSLVTVSSASTKGPSVFPLAPSSKSTSGGTAALGCLVKDYFPEPVTVSWNSGALTSGVHTFPAVLQS] SGLYSLSSVVTVPSSSLGTQTYICNVNHKPSNTKVDKKV *Sequence in FASTA format* *Sequence in IMGT format*
V-DOMAIN	IMGT domain description: VH (1-121) [D1]
	IMGT gene and allele name: IGHV4-59*01 (73.00%)(Human) *Alignment details*
	IMGT gene and allele name: IGHJ4*01 (92.90%)(Human) , IGHJ4*02 (92.90%)(Human) , IGHJ4*03 (92.90%)(Human) *Alignment details*
	2D representation: IMGT Collier de Perles or IMGT Collier de Perles on 2 layers
	Contact analysis: Domain contacts (overview)
	CDR-IMGT lengths: [8.10.12]
	Sheet composition: [A' B D E] [A" C C' C" F G]
	[CDR1] [CDR2] QVQLQESGP.GLVRPSQTLSLTCTVSGFTF....TDFYMNWVRQPPGRGLEWIGFIRDKAKGYTTEYNPSVK.GRVTMLVDTSKNQFSLR [CDR3] LSSVTAADTAVYYCAREGHT.AAPFDYWGQGSLVTVSS *IMGT/DomainGapAlign results*
C-DOMAIN	IMGT domain description: CH1 (122-210) [D2]
	IMGT gene and allele name: IGHG1*01 (100.00%)(Human) , IGHG1*02 (100.00%)(Human) , IGHG1*03 (99.00%)(Human) *Alignment details*
	2D representation: IMGT Collier de Perles or IMGT Collier de Perles on 2 layers
	Contact analysis: Domain contacts (overview)
	Sheet composition: [A B D E] [C F G]
	....ASTKGPSVFPLAPSSKSTS...GGTAALGCLVKDYFP..EPVTVSWNSGALTS....GVHTFPAVLQSS......GLYSLSSVVTV PSSSL...GTQTYICNVNHKP..SNTKVDKKV *IMGT/DomainGapAlign results*

FIGURE 8. IMGT/3Dstructure-DB card. "IMGT/3Dstructure-DB card" is available for each entry of the database. "Chain details" section for the VH-CH1 chain (1bey_H) of the alemtuzumab Fab is shown. Chains and domains are described with standardized IMGT labels. Similar result displays are provided for IMGT/2Dstructure-DB cards. However, in those cases information on experimental structural data (hydrogen bonds in IMGT Collier de Perles on two layers, Domain contacts) are only available in the corresponding IMGT/3Dstructure-DB cards, if the antibodies have been crystallized.

- *IMGT gene and allele name with the percentage of identity for the V (for example IGHV4-59*01 [73.00%] [Human])*
- *IMGT gene and allele name with the percentage of identity for the J (for example IGHJ4*01 [92.90%][Human] as well as other alleles giving the same percentage of identity)*
- *For V domain: CDR-IMGT lengths (for example [8.10.12]), Sheet composition (for example [A'BDE] [A"CC'C"FG]), the domain AA sequence with CDR-IMGT delimitations and highlighting of AA that are different from the closest V and J genes and alleles*
- *2D representation (Step 31)*
- *Contact analysis (Step 32)*

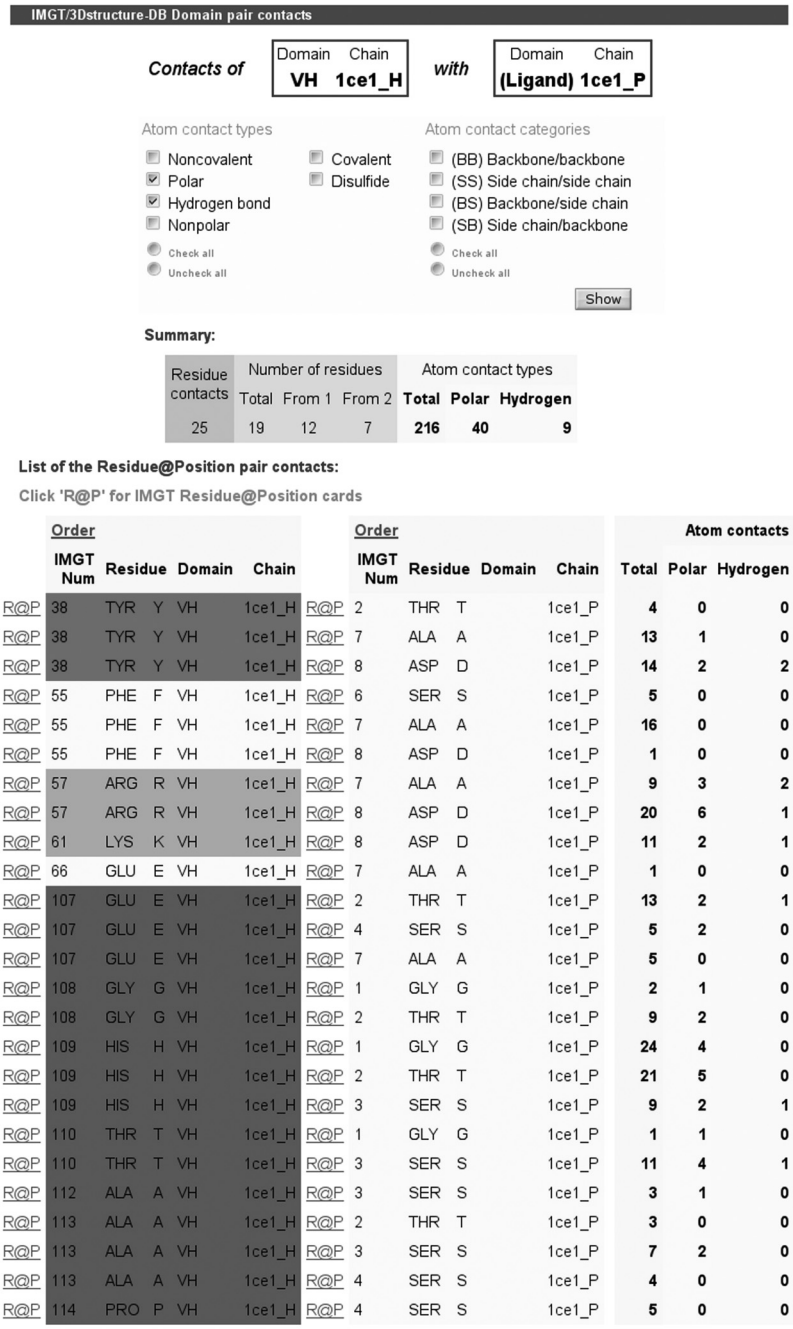
31. For more information, click on:

- "Sequence in FASTA format" or "Sequence in IMGT format," to access the chain sequence in FASTA or IMGT format.
- "Alignment details" to access the alignment with the closest identified gene and allele name for each domain.
- "IMGT Collier de Perles" (on one layer or two layers for V and C) to get the IMGT Collier de Perles for each domain **(IMGT/Collier de Perles: IMGT Standardized Representation of Domains (IG, TR, and IgSF Variable and Constant Domains, MH and MhSF Groove Domains))** (Ehrenmann et al. 2011).
- "IMGT/DomainGapAlign results" to get the IMGT/DomainGapAlign results for each domain **(IMGT/DomainGapAlign: IMGT Standardized Analysis of Amino Acid Sequences of**

Variable, Constant, and Groove Domains (IG, TR, MH, IgSF, MhSF)) (Ehrenmann and Lefranc 2011).

IMGT/3Dstructure-DB Card-Contact Analysis

32. In the IMGT/3Dstructure-DB card, click on "Domain contacts (overview)."
This gives access to "IMGT/3Dstructure-DB Domain pair contacts (overview)."



IMGT/3Dstructure-DB Domain pair contacts

Contacts of Domain VH Chain 1ce1_H *with* Domain (Ligand) Chain 1ce1_P

Atom contact types: Noncovalent, Polar, Hydrogen bond, Nonpolar, Covalent, Disulfide; Check all / Uncheck all

Atom contact categories: (BB) Backbone/backbone, (SS) Side chain/side chain, (BS) Backbone/side chain, (SB) Side chain/backbone; Check all / Uncheck all

Show

Summary:

Residue contacts	Number of residues			Atom contact types		
	Total	From 1	From 2	Total	Polar	Hydrogen
25	19	12	7	216	40	9

List of the Residue@Position pair contacts:

Click 'R@P' for IMGT Residue@Position cards

	Order						Order					Atom contacts		
	IMGT Num	Residue		Domain	Chain		IMGT Num	Residue		Domain	Chain	Total	Polar	Hydrogen
R@P	38	TYR	Y	VH	1ce1_H	R@P	2	THR	T		1ce1_P	4	0	0
R@P	38	TYR	Y	VH	1ce1_H	R@P	7	ALA	A		1ce1_P	13	1	0
R@P	38	TYR	Y	VH	1ce1_H	R@P	8	ASP	D		1ce1_P	14	2	2
R@P	55	PHE	F	VH	1ce1_H	R@P	6	SER	S		1ce1_P	5	0	0
R@P	55	PHE	F	VH	1ce1_H	R@P	7	ALA	A		1ce1_P	16	0	0
R@P	55	PHE	F	VH	1ce1_H	R@P	8	ASP	D		1ce1_P	1	0	0
R@P	57	ARG	R	VH	1ce1_H	R@P	7	ALA	A		1ce1_P	9	3	2
R@P	57	ARG	R	VH	1ce1_H	R@P	8	ASP	D		1ce1_P	20	6	1
R@P	61	LYS	K	VH	1ce1_H	R@P	8	ASP	D		1ce1_P	11	2	1
R@P	66	GLU	E	VH	1ce1_H	R@P	7	ALA	A		1ce1_P	1	0	0
R@P	107	GLU	E	VH	1ce1_H	R@P	2	THR	T		1ce1_P	13	2	1
R@P	107	GLU	E	VH	1ce1_H	R@P	4	SER	S		1ce1_P	5	2	0
R@P	107	GLU	E	VH	1ce1_H	R@P	7	ALA	A		1ce1_P	5	0	0
R@P	108	GLY	G	VH	1ce1_H	R@P	1	GLY	G		1ce1_P	2	1	0
R@P	108	GLY	G	VH	1ce1_H	R@P	2	THR	T		1ce1_P	9	2	0
R@P	109	HIS	H	VH	1ce1_H	R@P	1	GLY	G		1ce1_P	24	4	0
R@P	109	HIS	H	VH	1ce1_H	R@P	2	THR	T		1ce1_P	21	5	0
R@P	109	HIS	H	VH	1ce1_H	R@P	3	SER	S		1ce1_P	9	2	1
R@P	110	THR	T	VH	1ce1_H	R@P	1	GLY	G		1ce1_P	1	1	0
R@P	110	THR	T	VH	1ce1_H	R@P	3	SER	S		1ce1_P	11	4	1
R@P	112	ALA	A	VH	1ce1_H	R@P	3	SER	S		1ce1_P	3	1	0
R@P	113	ALA	A	VH	1ce1_H	R@P	2	THR	T		1ce1_P	3	0	0
R@P	113	ALA	A	VH	1ce1_H	R@P	3	SER	S		1ce1_P	7	2	0
R@P	113	ALA	A	VH	1ce1_H	R@P	4	SER	S		1ce1_P	4	0	0
R@P	114	PRO	P	VH	1ce1_H	R@P	4	SER	S		1ce1_P	5	0	0

FIGURE 9. IMGT/3Dstructure-DB Domain pair contacts between the VH domain of alemtuzumab (1ce1_H) and the CD52 peptide ligand (1ce1_P). "Polar" and "Hydrogen bonds" were selected prior to display, in "Atom contact types." Amino acids belonging to the CDR1-IMGT, CDR2-IMGT, and CDR3-IMGT are colored in red, orange, and purple, respectively. Positions 55 and 66 are anchor positions. Clicking on R@P gives access to the IMGT Residue@Position cards.

33. Locate "Atom contact types" (noncovalent, polar, hydrogen bond, etc.) and "Atom contact categories" ((BB) Backbone/backbone, (SS) Side chain/side chain, etc.) and select those you want to analyze.

34. Select the domain pair for which you want to analyze contacts by clicking on the corresponding "DomPair" (for instance between VH and ligand).

    This gives access to the "IMGT/3Dstructure-DB Domain pair contacts" (e.g., VH domain of 1ce1_H with the ligand 1ce1_P) (Fig. 9).

35. Select the "IMGT Residue@Position card" that provides the contacts of a residue at a given position in a domain by clicking on the corresponding "R@P."

    This gives access to the IMGT Residue@Position card.

    A Residue@Position, for example: 28 – PHE (F) – VH – 1ce1_H, is defined by the IMGT position numbering, the residue name, the IMGT domain description, and the IMGT chain ID. The IMGT Residue@Position card provides, for a given residue at a given position, structural information, and contacts classified by atom contact types (noncovalent, polar, hydrogen bond, etc.) and atom contact categories ([BB] Backbone/backbone, [SS] Side chain/side chain, etc).

    The IMGT Residue@Position cards can also be accessed directly from the amino acid sequences of the IMGT/3Dstructure-DB card (Step 29 in this article; Kaas et al. 2004) or from the IMGT Colliers de Perles (Ruiz and Lefranc 2002); ***IMGT/Collier de Perles: IMGT Standardized Representation of Domains (IG, TR, and IgSF Variable and Constant Domains, MH and MhSF Groove Domains)*** *(Ehrenmann et al. 2011), by clicking on one AA.*

DISCUSSION

IMGT/3Dstructure-DB data are particularly useful in antibody engineering and humanization design (Lefranc 2009; Ehrenmann et al. 2010a). Indeed, they allow one to precisely analyze and to easily compare amino acid sequences of the framework regions (FR-IMGT) and complementarity determining regions (CDR-IMGT), between the nonhuman (mouse, rat…) and the closest human V domains. A recent analysis performed on humanized antibodies used in oncology underlines the importance of a correct delimitation of the CDR to be grafted (Magdelaine-Beuzelin et al. 2007). IMGT3Dstructure-DB facilitates the identification of potential immunogenic residues at given positions in chimeric or humanized antibodies, including those of the constant domains (Jefferis and Lefranc 2009). Moreover, the low potential immunogenicity of nonhuman primate antibodies and their potential use as therapeutics in humans has been shown by this approach. This is particularly important for nonhuman primate antibodies neutralizing *Bacillus anthracis* and the anthrax lethal toxin or the ricin toxin that could be used as therapeutics in ricin intoxication and against bioweapons (Pelat et al. 2008, 2009). These therapeutic applications emphasize the importance of the IMGT/3Dstructure-DB standardized approach that bridges the gap between sequences and 3D structures, whatever the species (Kaas and Lefranc 2007; Lefranc et al. 2008).

ACKNOWLEDGMENTS

We thank Christophe le Roy for checking the protocol. We are grateful to Gérard Lefranc and to the IMGT team for their motivation and expertise.

REFERENCES

Duroux P, Kaas Q, Brochet X, Lane J, Ginestoux C, Lefranc M-P, Giudicelli V. 2008. IMGT-Kaleidoscope, the Formal IMGT-ONTOLOGY paradigm. *Biochimie* **90:** 570–583.

Ehrenmann F, Lefranc M-P. 2011. IMGT/DomainGapAlign: IMGT standardized analysis of amino acid sequences of variable, constant and groove domains (IG, TR, MH, IgSF, MhSF). *Cold Spring Harb Protoc* doi: 10.1101/pdb.prot5636.

Ehrenmann F, Duroux P, Giudicelli V, Lefranc M-P. 2010a. Standardized sequence and structure analysis of antibody using IMGT. In *Antibody engineering* 2nd ed. (ed. R Kontermann, S Dübel), Vol 2, pp. 11–31. Springer-Verlag, Berlin/Heidelberg, Germany.

Ehrenmann F, Kaas Q, Lefranc M-P. 2010b. IMGT/3Dstructure-DB and IMGT/DomainGapAlign: A database and a tool for immunoglobulins or antibodies, T cell receptors, MHC, IgSF and MhcSF. *Nucleic Acids Res* **38:** D301–D307.

Ehrenmann F, Giudicelli V, Brochet X, Lefranc M-P. 2011. IMGT/Collier de Perles : IMGT standardized representation of domains (IG, TR, and IgSF variable and constant domains, MH and MhSF groove domains). *Cold Spring Harb Protoc* doi: 10.1101/pdb.prot5635.

Giudicelli V, Lefranc M-P. 1999. Ontology for immunogenetics: IMGT-ONTOLOGY. *Bioinformatics* **15:** 1047–1054.

Giudicelli V, Lefranc M-P. 2011. IMGT/JunctionAnalysis: IMGT standardized analysis of the V-J and V-D-J junctions of the rearranged immunoglobulins (IG) and T cell receptors (TR). *Cold Spring Harb Protoc* doi: 10.1101/pdb.prot5634.

Giudicelli V, Chaume D, Lefranc M-P. 2005. IMGT/GENE-DB: A comprehensive database for human and mouse immunoglobulin and T cell receptor genes. *Nucleic Acids Res* **33:** D256–D261.

Giudicelli V, Brochet X, Lefranc M-P. 2011. IMGT/V-QUEST: IMGT standardized analysis of the immunoglobulin (IG) and T cell receptor

(TR) nucleotide sequences. *Cold Spring Harb Protoc* doi: 10.1101/pdb.prot5633.

Jefferis R, Lefranc M-P. 2009. Human immunoglobulin allotypes: Possible implications for immunogenicity. *MAbs* **1:** 332–338.

Kaas Q, Lefranc M-P. 2007. IMGT Colliers de Perles: Standardized sequence-structure representations of the IgSF and MhcSF superfamily domains. *Current Bioinformatics* **2:** 21–30.

Kaas Q, Ruiz M, Lefranc M-P. 2004. IMGT/3Dstructure-DB and IMGT/StructuralQuery, a database and a tool for immunoglobulin, T cell receptor and MHC structural data. *Nucl Acids Res* **32:** D208–D210.

Kaas Q, Ehrenmann F, Lefranc M-P. 2007. IG, TR and IgSf, MHC and MhcSF: What do we learn from the IMGT Colliers de Perles? *Brief. Funct Genomic Proteomic* **6:** 253–264.

Lefranc M-P. 1997. Unique database numbering system for immunogenetic analysis. *Immunol Today* **18:** 509. doi:10.1016/S0167-5699(97)01163-8.

Lefranc M-P. 1999. The IMGT unique numbering for Immunoglobulins, T cell receptors and Ig-like domains. *The Immunologist* **7:** 132–136.

Lefranc M-P. 2009. Antibody database and tools: The IMGT experience. In *Therapeutic monoclonal antibodies: from Bench to Clinic* (ed. A Zhiqiang), pp. 91–114. John Wiley Sons, Inc, Hoboken, NJ.

Lefranc M-P. 2011a. IMGT, the International ImMunoGeneTics Information System. *Cold Spring Harb Protoc* doi: 10.1101/pdb.top115.

Lefranc M-P. 2011b. From IMGT-ONTOLOGY IDENTIFICATION axiom to IMGT standardized keywords: For immunoglobulins (IG), T cell receptors (TR), and conventional genes. *Cold Spring Harb Protoc* doi: 10.1101/pdb.ip82.

Lefranc M-P. 2011c. From IMGT-ONTOLOGY DESCRIPTION axiom to IMGT standardized labels: For immunoglobulin (IG) and T cell receptor (TR) sequences and structures. *Cold Spring Harb Protoc* doi: 10.1101/pdb.ip83.

Lefranc M-P. 2011d. From IMGT-ONTOLOGY CLASSIFICATION axiom to IMGT standardized gene and allele nomenclature: For immunoglobulins (IG) and T cell receptors (TR). *Cold Spring Harb Protoc* doi: 10.1101/pdb.ip84.

Lefranc M-P. 2011e. IMGT unique numbering for the variable (V), constant (C), and groove (G) domains of IG, TR, MH, IgSF, and MhSF. *Cold Spring Harb Protoc* doi: 10.1101/pdb.ip85.

Lefranc M-P. 2011f. IMGT Collier de Perles for the Variable (V), Constant (C), and Groove (G) Domains of IG, TR, MH, IgSF, and MhSF. *Cold Spring Harb Protoc* doi: 10.1101/pdb.ip86.

Lefranc M-P, Lefranc G. 2001a. *The Immunoglobulin FactsBook*, 1–458. Academic Press, London, UK.

Lefranc M-P, Lefranc G. 2001b. *The T cell receptor FactsBook*, 1–398. Academic Press, London, UK.

Lefranc M-P, Pommié C, Ruiz M, Giudicelli V, Foulquier E, Truong L, Thouvenin-Contet V, Lefranc G. 2003. IMGT unique numbering for immunoglobulin and T cell receptor variable domains and Ig superfamily V-like domains. *Dev Comp Immunol* **27:** 55–77.

Lefranc M-P, Giudicelli V, Ginestoux C, Bosc N, Folch G, Guiraudou D, Jabado-Michaloud J, Magris S, Scaviner D, Thouvenin V, et al. 2004. IMGT-ONTOLOGY for Immunogenetics and Immunoinformatics. *In Silico Biol* **4:** 17–29.

Lefranc M-P, Clément O, Kaas Q, Duprat E, Chastellan P, Coelho I, Combres K, Ginestoux C, Giudicelli V, Chaume D, et al. 2005a. IMGT-Choreography for Immunogenetics and Immunoinformatics. *In Silico Biol* **5:** 45–60.

Lefranc M-P, Pommié C, Kaas Q, Duprat E, Bosc N, Guiraudou D, Jean C, Ruiz M, Da Piedade I, Rouard M, et al. 2005b. IMGT unique numbering for immunoglobulin and T cell receptor constant domains and Ig superfamily C-like domains. *Dev Comp Immunol* **29:** 185–203.

Lefranc M-P, Duprat E, Kaas Q, Tranne M, Thiriot A, Lefranc G. 2005c. IMGT unique numbering for MHC groove G-DOMAIN and MHC superfamily (MhcSF) G-LIKE-DOMAIN. *Dev Comp Immunol* **29:** 917–938.

Lefranc M-P, Giudicelli V, Regnier L, Duroux P. 2008. IMGT, a system and an ontology that bridge biological and computational spheres in bioinformatics. *Brief Bioinform* **9:** 263–275.

Lefranc M-P, Giudicelli V, Ginestoux C, Jabado-Michaloud J, Folch G, Bellahcene F, Wu Y, Gemrot E, Brochet X, Lane J, et al. 2009. IMGT, the international ImMunoGeneTics information system. *Nucl Acids Res* **37:** D1006–D1012.

Magdelaine-Beuzelin C, Kaas Q, Wehbi V, Ohresser M, Jefferis R, Lefranc M-P, Watier H. 2007. Structure-function relationships of the variable domains of monoclonal antibodies approved for cancer treatment. *Crit Rev Oncol Hematol* **64:** 210–225.

Pearson WR, Lipman DJ. 1988. Improved tools for biological sequence comparison. *Proc Natl Acad Sci* **85:** 2444–2448.

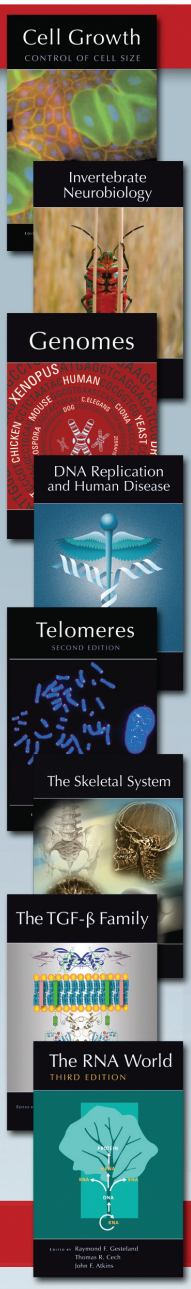
Pelat T, Bedouelle H, Rees AR, Crennell SJ, Lefranc M-P, Thullier P. 2008. Germline humanization of a non-human Primate antibody that neutralizes the anthrax toxin, by in vitro and in silico engineering. *J Mol Biol* **384:** 1400–1407.

Pelat T, Hust M, Hale M, Lefranc M-P, Dübel S, Thullier P. 2009. Isolation of a human-like antibody fragment (scFv) that neutralizes ricin biological activity. *BMC Biotechnology* **9:** 60.

Ruiz M, Lefranc M-P. 2002. IMGT gene identification and Colliers de Perles of human immunoglobulins with known 3D structures. *Immunogenetics* **53:** 857–883.

Cold Spring Harbor Monograph Archive

Now Online!



First Web Availability of the Renowned Book Series

The CSH Monograph Archive offers the complete collection of scholarly monographs published by Cold Spring Harbor Laboratory Press from 1970 to 2009. The archive's 59 full-text volumes provide the life science community with definitive reviews of progress in areas of molecular, cell, and developmental biology, genetics, evolutionary biology, neuroscience, cancer biology, and molecular pathology. Each text is written and commissioned by foremost researchers in their particular discipline.

This new online archive is an unmatched resource for its breadth of coverage in key topics and provides an in-depth account of developments as they occurred in numerous fields. It is available online as a complete collection for one-time purchase (with perpetual access) or on pay-per-view basis by book chapter.

- **The complete text of all Cold Spring Harbor Laboratory Press monographs from 1970 to 2008 (59 titles)**
- **Fully searchable and browsable by author, title, or subject**
- **One-time purchase provides perpetual access**
- **Chapters offer HTML abstracts and full-text pdfs with full-color illustrations**

Includes these notable titles:

- *The RNA World*
- *Translational Control*
- *Stem Cell Biology*
- *DNA Replication and Human Disease*
- *The Molecular Biology of Tumour Viruses*
- *The Bacteriophage Lambda*
- *Epigenetic Mechanisms of Gene Regulation*
- *Prion Biology and Diseases*
- *The Nematode*
- *Oncogenes and the Molecular Origins of Cancer*
- *Transcriptional Regulation*
- *Telomeres*
- *The Development of Human Gene Therapy*
- *The Dog and Its Genome*
- *Molecular Biology of Aging*
- *Adult Neurogenesis*

See website for complete list.

www.cshmonographs.org

To order or request additional information, please visit our website or:

Call: 1-800-843-4388 (Continental US and Canada) **516-422-4100** (All other locations)
FAX: 516-422-4097 **E-mail: cshpress@cshl.edu**
Write:Cold Spring Harbor Laboratory Press, 500 Sunnyside Blvd., Woodbury, NY 11797-2924