Biocuration of IMGT/2Dstructure-DB and IMGT/3Dstructure-DB: IMGT unique numbering and IMGT Colliers de Perles bridging the gap between sequences and structures in IMGT®

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

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

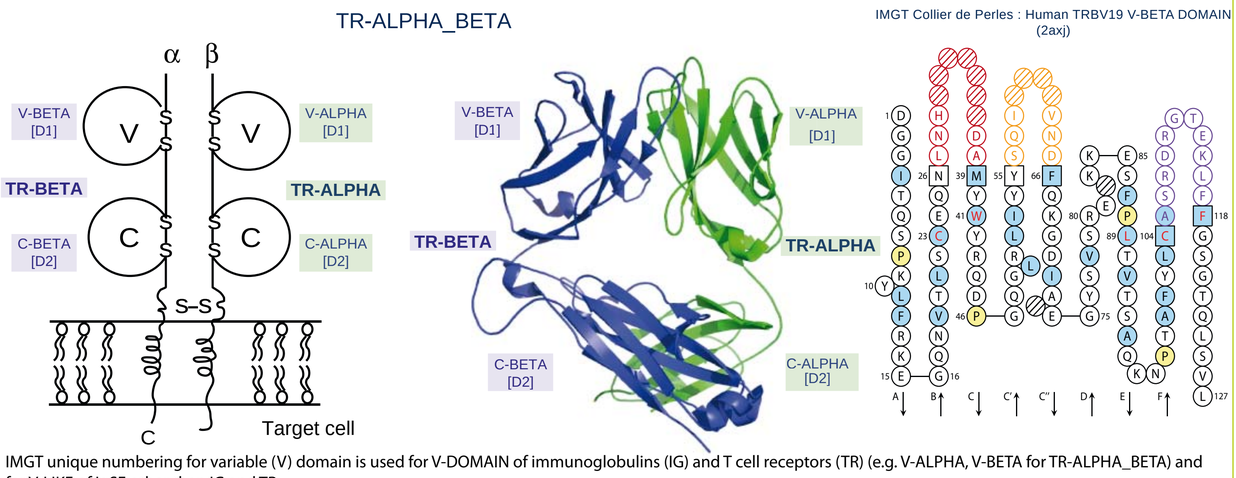
http://www.imgt.org

IMGT®, the international ImMunoGeneTics information system®, http://www.imgt.org, the global reference in immunogenetics and immunoinformatics, provides an integrated approach to biocuration of two-dimensional (2D) and three-dimensional (3D) structures in IMGT/2Dstructure-DB and IMGT/3Dstructure-DB, that allows one to annotate 2D structures even if 3D structures are not yet available. This approach is based on two major concepts of numerotation of IMGT-ONTOLOGY, generated from the NUMEROTATON axiom: the IMGT unique numbering and the IMGT Collier de Perles. These concepts have been defined for the variable (V), the constant (C) and the groove (G) domains. The V domain includes the V-DOMAIN of the immunoglobulins (IG) and T cell receptors (TR) and the V-LIKE-DOMAIN of the immunoglobulin superfamily (IgSF) proteins other than IG and TR. The C domain includes the C-DOMAIN of the IG and TR and the C-LIKE-DOMAIN of the IgSF proteins other than IG and TR. The G domain includes the G-DOMAIN of the major histocompatibility (MH) proteins and the G-LIKE-DOMAIN of the MhSF proteins other than MH. The IMGT unique numbering always assigns the same position to conserved amino acids in the domain, whatever the chain type and the species. For example, a V or a C domain is characterized by 4 (5 in a V-DOMAIN) conserved amino acids, 1st-CYS 23, CONSERVED-TRP 41, hydrophobic amino acid 89, 2nd-CYS 104, and for a V-DOMAIN, J-PHE or J-TRP 118. In addition, the IMGT Collier de Perles allows to visualize the standardized strands and loops of the V and C domains and strands and helix of the G domain and to precisely delimit the framework regions (FR-IMGT) and complementarity determining regions (CDR-IMGT) of the V-DOMAIN. Whereas IMGT/3Dstructure-DB contains data specific to 3D structures (contact analysis, peptide/MH (pMH) complexes, antigen receptor/antigen complexes (IG/Ag, TR/pMH)), it also contains data similar to IMGT/2Dstructure-DB. Both databases contain IG or antibodies, TR, MH, related proteins of the immune system (RPI). Biocuration is performed using IMGT/DomainGapAlign that provides IMGT gene and allele names (CLASSIFICATION), region and domain delimitations (DESCRIPTION), and amino acid positions according to the IMGT unique numbering (NUMEROTATION). Coupled to the IMGT/Collier-de-Perles tool, IMGT/DomainGapAlign provides an invaluable help for antibody engineering and humanization design of therapeutical antibodies based on CDR grafting. Indeed they allow to precisely define and to easily compare amino acid sequences of the FR-IMGT and CDR-IMGT, between the nonhuman (mouse, rat…) and the closest human V domains. It also facilitates the identification of potential immunogenic residues at given positions in chimeric or humanized antibodies, including those of the C domains. These therapeutic applications emphasize the importance of the IMGT® standardized approach that bridges the gap between sequences and structures whatever the species. Since 2008, amino acid sequences of monoclonal antibodies (mAb, suffix -mab), of fusion proteins for immune applications (FPIA, suffix -cept) and of composite proteins for clinical applications (CPCA) from the WHO/International Nonproprietary Name (INN) programme have been entered in IMGT®.

[1] IMGT booklet (11 papers), Cold Spring Harb Protoc, 124 pages (2011) (pdf, IMGT References, http://www.imgt.org). *With generous provision from Cold Spring Harbor (CSH) Protocols.*

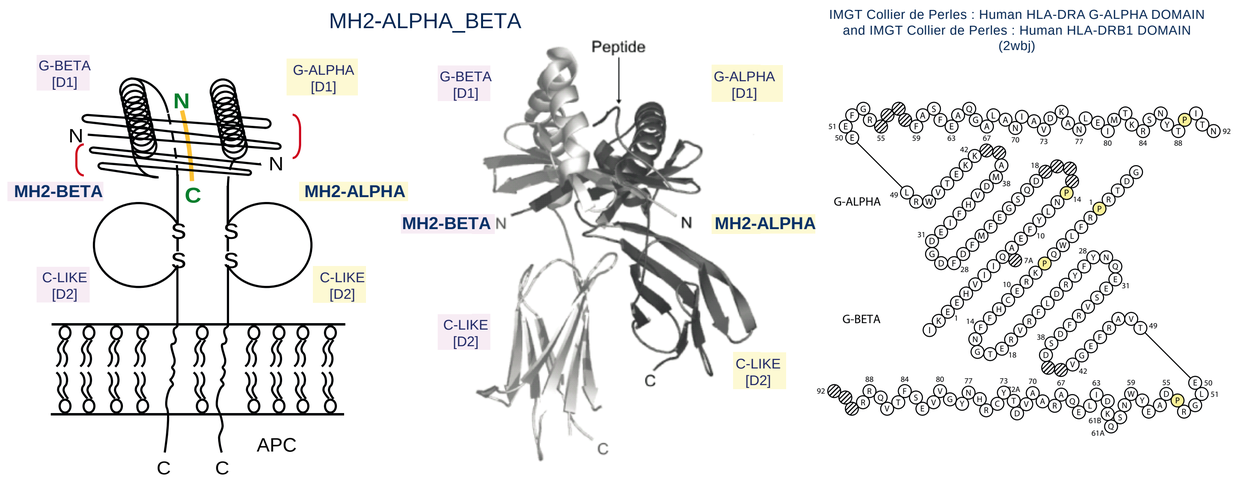
IMGT unique numbering and IMGT Colliers de Perles

V-DOMAIN and C-DOMAIN of IG and TR, V-LIKE and C-LIKE of IgSF other than IG and TR



IMGT unique numbering for variable (V) domain is used for V-DOMAIN of immunoglobulins (IG) and T cell receptors (TR) (e.g. V-ALPHA, V-BETA for TR-ALPHA_BETA) and for V-LIKE of IgSF other than IG and TR.
IMGT unique numbering for constant (C) domain is used for C-DOMAIN of immunoglobulins (IG) and T cell receptors (TR) (e.g. C-ALPHA, C-BETA for TR-ALPHA_BETA) and for C-LIKE of IgSF other than IG and TR.

G-DOMAIN of MH, G-LIKE of MhSF other than MH

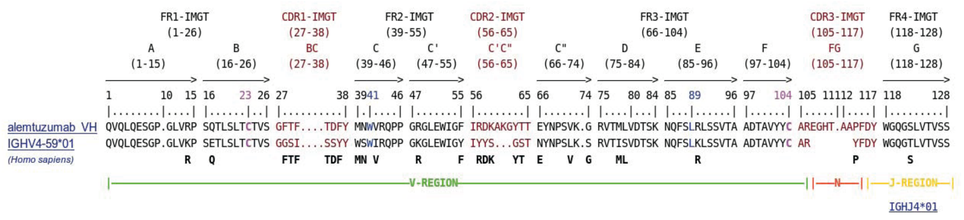


IMGT unique numbering for groove (G) domain is used for G-DOMAIN of Major histocompatibility (MH) (e.g. G-ALPHA, G-BETA for MH2-ALPHA_BETA) and for G-LIKE of MhSF other than MH. MH belongs also to IgSF by its C-LIKE.

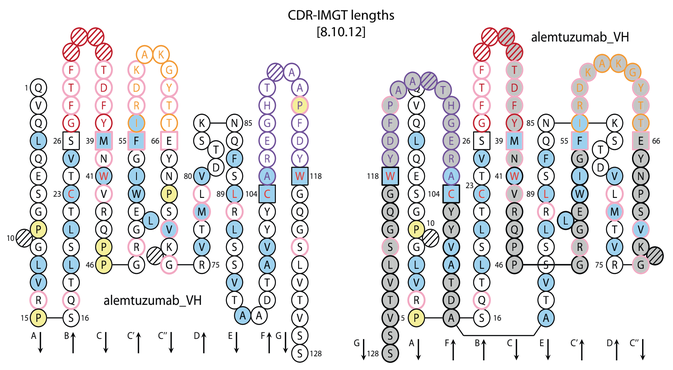
IMGT/2Dstructure-DB and IMGT/DomainGapAlign

IMGT/DomainGapAlign identifies the closest germline V-REGION (for 'V'), C-DOMAIN (for 'C') or G-DOMAIN (for 'G'), creates gaps according to the IMGT unique numbering and highlights differences with the closest reference(s). For an antibody V domain sequence, the tool identifies the closest germline V-REGION and J-REGION, and provides a delimitation of the strands, framework regions (FR-IMGT) and CDR-IMGT. The gene and allele name of the closest sequence(s) from the IMGT domain directory is provided with a percentage of identity and a Smith-Waterman score. Regions and domains are highlighted using the IMGT color menu and IMGT Colliers de Perles are generated from the gapped sequences provided by the tool.

V Gene and allele	Species	Domain	Smith-Waterman	Identity percentage	Overlap
IGHV4-59*01	*Homo sapiens*	1	494	73.0	100
J Gene and allele	**Species**	**Domain**	**Smith-Waterman**	**Identity percentage**	**Overlap**
IGHJ4*01	*Homo sapiens*	1	94	92.9	14



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_VH	73.0%	[8,10,12]	11	[25.17.38.11] = 91 AA	14	25



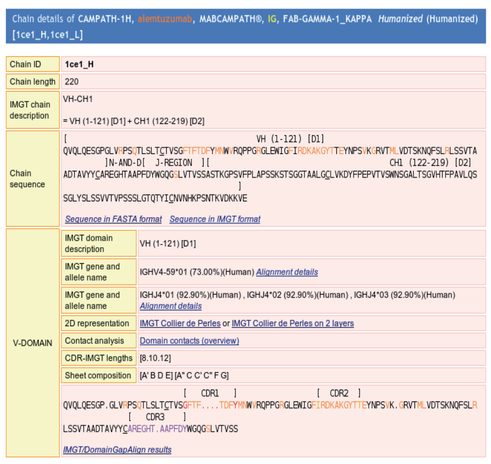
The circles bordered in pink indicate amino acid differences with the closest domain.

A standardized comparison of V domain sequences for antibody humanization includes the delimitation of the FR-IMGT and CDR-IMGT, the determination of the CDR-IMGT lengths (e.g. [8.8.13]), the percentage of identity between FR-IMGT [calculated on 91 amino acids for VH (FR1-, FR2-, FR3-, FR4-IMGT: 25, 17, 38, 11) and 89 for V-KAPPA (FR1-, FR2-, FR3-, FR4-IMGT: 26, 17, 36, 10)] and the number of IMGT amino acid physicochemical class changes.

	Humanized antibody	CDR-IMGT antibody	Closest *Homo sapiens* gene and allele	FR-IMGT identity	AA with IMGT class change in FR-IMGT
VH	alemtuzumab	[8.10.12]	IGHV4-59*01	84.61% (77/91)	14
	bevacizumab	[8.8.16]	IGHV7-4-1*02	74.72% (68/91)	24
	trastuzumab	[8.8.13]	IGHV3-66*01	90.10% (82/91)	9
VL	alemtuzumab	[6.3.9]	IGKV1-33*01	97.75% (87/89)	2
	bevacizumab	[6.3.9]	IGKV1-33*01	92.13% (82/89)	7
	trastuzumab	[6.3.9]	IGKV1-39*01	93.25% (83/89)	5

IMGT/3Dstructure-DB

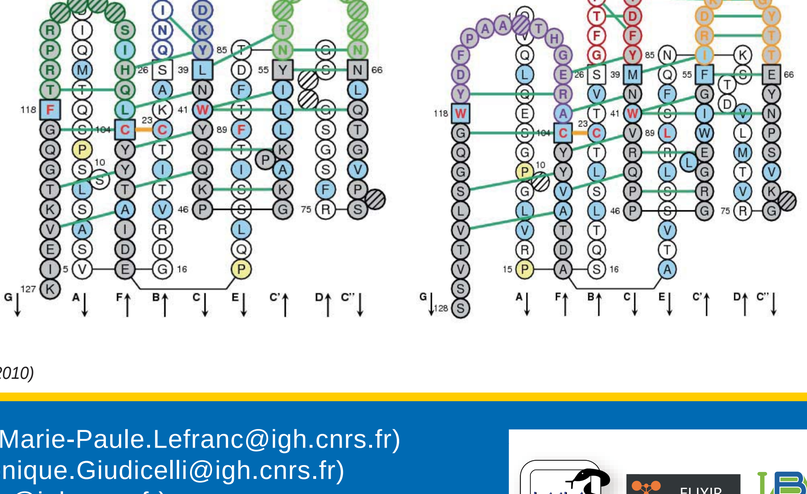
IMGT/3Dstructure-DB card gives access to chain details with individual domain description. For each domain there is a link to IMGT Collier de Perles (e.g. on one layer and two layers for V-DOMAIN) and Domain contacts (overview).

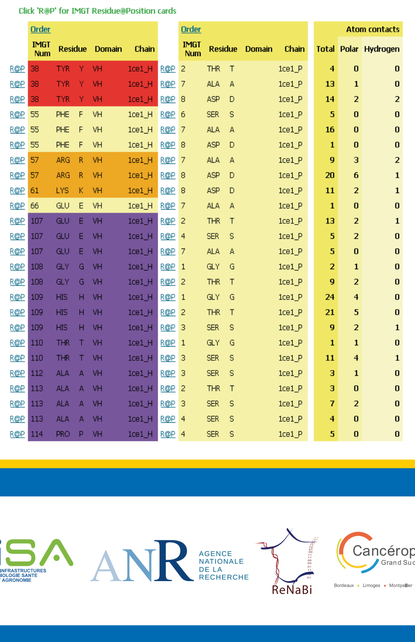


Ehrenmann F., Kaas Q. and Lefranc M.-P. Nucleic Acids Res. 38: D301-307 (2010)

If one 3D structure is available (for example 1ce1 for alemtuzumab), it is possible to query IMGT/3Dstructure-DB:

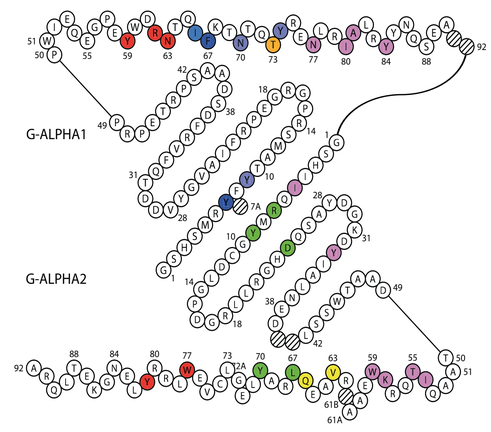
- to visualize **hydrogen bounds** in the IMGT Collier de Perles on two layers
- to check the **contact analysis** table. This table provides contacts between structural units (domains or ligand).





Order	IMGT Num	Residue	Domain	Chain	Order	IMGT Num	Residue	Domain	Chain	Atom contacts 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

IMGT pMH contact sites graphically represent, in IMGT Colliers de Perles, the MH amino acid positions that contact the peptide side chains in pMH complexes, and thus allow comparison of pMH interactions, between 3D structures of the same MH class or even between MH1 and MH2 classes.



IMGT® founder and director: Marie-Paule Lefranc (Marie-Paule.Lefranc@igh.cnrs.fr)
Bioinformatics manager: Véronique Giudicelli (Veronique.Giudicelli@igh.cnrs.fr)
Computer manager: Patrice Duroux (Patrice.Duroux@igh.cnrs.fr)
Webmaster: Chantal Ginestoux (Chantal.Ginestoux@igh.cnrs.fr)