IMGT/HighV-QUEST for NGS antibody repertoire analysis

Véronique Giudicelli, Eltaf Alamyar, Géraldine Folch, Joumana Jabado-Michaloud, Patrice Duroux, and Marie-Paule Lefranc

IMGT®, the international ImMunoGeneTics information system®,
Laboratoire d'ImmunoGénétique Moléculaire LIGM, Institut de Génétique Humaine IGH, UPR CNRS 1142,
141 rue de la Cardonille, Montpellier, 34396, France

http://www.imgt.org

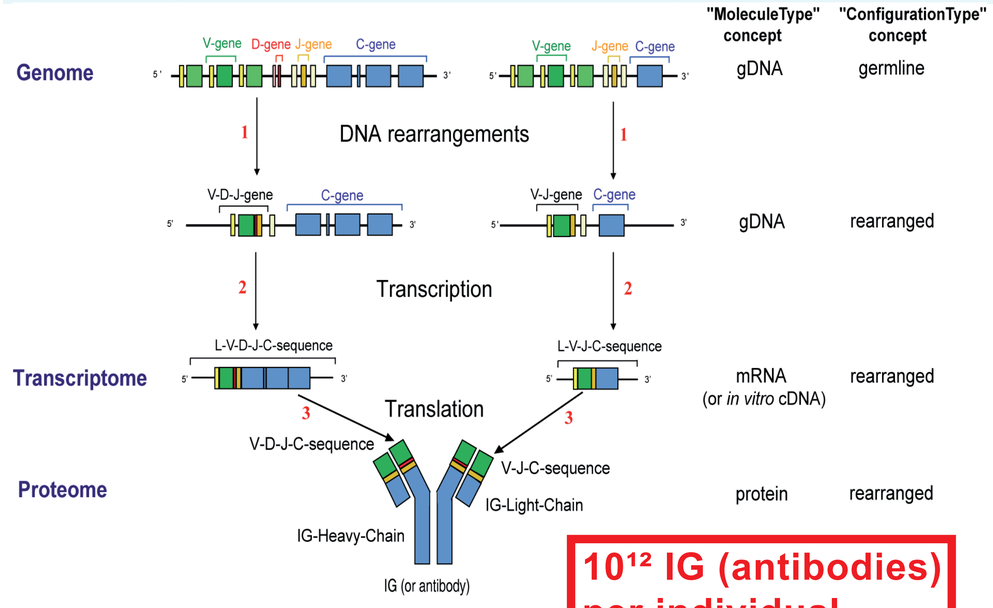
IMGT®, the international ImMunoGeneTics information system®, created in 1989 at Montpellier, France, by Marie-Paule Lefranc (CNRS and Université Montpellier 2), is at the birth of immunoinformatics [1]. IMGT® manages the immunogenetics data, and more particularly the sequences, genes and structures of immunoglobulins (IG) or antibodies and T cell receptors (TR). Standardization and data integration are obtained through the IMGT-ONTOLOGY [2] concepts of identification (IMGT standardized keywords), classification (IMGT standardized nomenclature: IMGT gene and allele names approved by HGNC and used by NCBI Gene), description (IMGT standardized labels) and numerotation (IMGT unique numbering and IMGT Colliers de Perles: widely used for antibody engineering and humanization). IMGT® comprises seven databases (including IMGT/mAb-DB), seventeen tools and more than 15,000 pages of Web resources. To answer the needs of high throughput and Next Generation Sequencing (NGS) data, the IMGT/HighV-QUEST tool [3-5] was developed which analyses up to 500,000 long 454 sequences by run. The results, based on IMGT-ONTOLOGY, include identification of the closest germline genes and alleles for genotype and haplotype analysis, and standardized characterization of the 'IMGT clonotypes (AA)' for antibody clonal diversity and expression and achieve, for the first time, a degree of resolution for NGS verifiable by the user at the sequence level. Amino acid frequency can be determined at each CDR-IMGT and FR-IMGT positions. This tool provides a paradigm for IG clonal diversity and expression repertoire analysis from NGS and high resolution results for antibody engineering and combinatorial library construction.

TAKE HOME MESSAGE: * IMGT/HighV-QUEST analyses antibody NGS data at a sequence level verifiable by the user. * Standardized characterization of IMGT clonotypes, based on the IMGT-ONTOLOGY concepts, identifies clonal diversity and expression, * Amino acid frequency can be determined at each CDR-IMGT and FR-IMGT position.

[1] Lefranc M-P, Front Immunol, 5:22, 2014. [2] Giudicelli V and Lefranc M-P, Front Genet, 3:79, 2012. [3] Alamyar E et al. Mol Biol 882:569-604, 2012. [4] Alamyar E et al. Immunome Res 8(1):26, 2012. [5] Li S et al. Nat. Commun. 4:2333 doi: 10.1038/ncomms3333 (2013).

Biological Context

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



10^{12} IG (antibodies) per individual

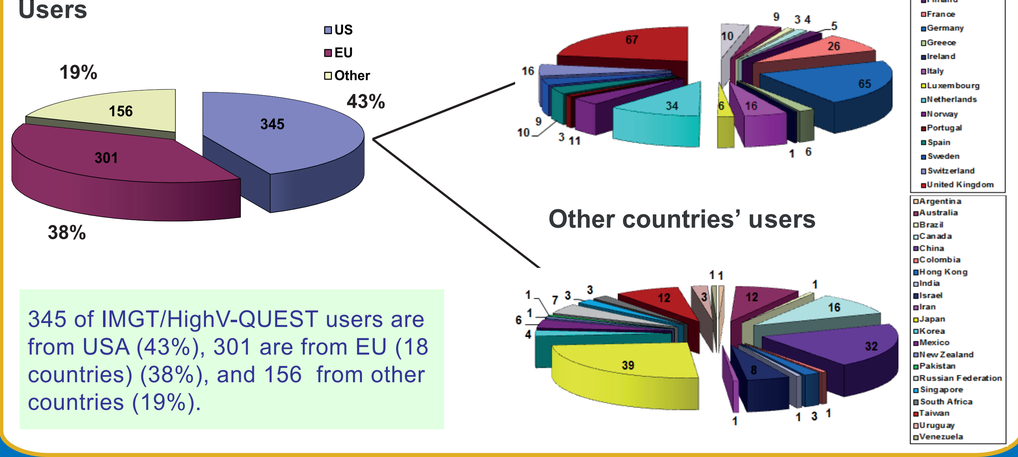
IMGT Clonotypes (AA)

In IMGT®, the clonotype designated as 'IMGT clonotype (AA)' is defined among the '1 copy' 'single allele' (for V and J) by a unique V-(D)-J rearrangement (IMGT genes and alleles determined at the nt level), conserved anchors (C104, W or F 118), and a unique CDR3-IMGT AA in frame junction [4]. Each 'IMGT clonotype (AA)' is characterized by a selected unique representative sequence.

ID		IMGT clonotype (AA) definition						IMGT clonotype (AA) representative sequence				Nb			IMGT clonotypes (nt)
#	Exp. ID	V gene and allele	D gene and allele	J gene and allele	CDR3-IMGT length (AA)	CDR3-IMGT sequence (AA)	Anchors 104,118	V %	Sequence length	Sequence ID		Total nb of '1 copy'	Total nb of 'More than 1'	Total	Sequences file ('1 copy')
Homsap IGHV1-2*02 F															
1	137-mid5	Homsap IGHV1-2*02 F	Homsap IGHD2-2*01 F	Homsap IGHJ6*03 F	22 AA	ARDLYCSSTSCYGGWYYYYMDV	C,F	95.14	425	GJNZTB402H8X9K	length=425	1	0	1	Sequences file
2	157-mid5	Homsap IGHV1-2*02 F	Homsap IGHD6-6*01 F	Homsap IGHJ6*03 F	22 AA	ARERVGRSIAARRAPDYYYMDV	C,F	97.92	426	GJNZTB402H4DK	W length=426	1	0	1	Sequences file
3	305-mid5	Homsap IGHV1-2*02 F	Homsap IGHD3-22*01 F	Homsap IGHJ4*02 F	21 AA	ARGPYHRPTYYYDSSGYYGDY	C,F	96.15	374	GJNZTB402FSHFL	length=374	1	0	1	Sequences file
4	331-mid5	Homsap IGHV1-2*02 F	Homsap IGHD3-10*02 F	Homsap IGHJ3*02 F	21 AA	ARNVGHRPGSSDAWEADAFDI	C,F	99.31	422	GJNZTB402JFJ2O	length=422	1	0	1	Sequences file
5	374-mid5	Homsap IGHV1-2*02 F	Homsap IGHD3-3*01 F	Homsap IGHJ3*02 F	21 AA	ATTHEPAITIFGVVINDAFDI	C,F	96.18	422	GJNZTB402H3QEP	length=422	1	0	1	Sequences file
6	647-mid5	Homsap IGHV1-2*02 F	Homsap IGHD6-19*01 F	Homsap IGHJ6*03 F	19 AA	AKGAIAVAGTNYYYYYMDV	C,F	97.74	330	GJNZTB402HVTXP	length=330	1	0	1	Sequences file
7	693-mid5	Homsap IGHV1-2*02 F	Homsap IGHD3-22*01 F	Homsap IGHJ5*02 F	19 AA	ARDGSTYYYDSSGYYWFDP	C,F	99.65	417	GJNZTB402FI7QZ	length=417	1	0	1	Sequences file
8	709-mid5	Homsap IGHV1-2*02 F	Homsap IGHD4-23*01 ORF	Homsap IGHJ2*01 F	19 AA	ARDMGRYGGNLRRYWYFDL	C,F	97.57	416	GJNZTB402G7EN4	length=416	1	0	1	Sequences file

Users and Analyses

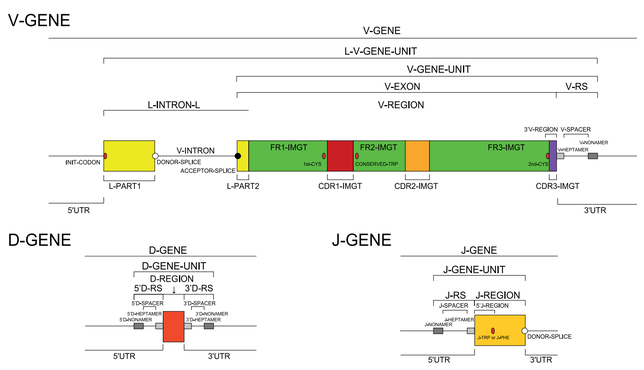
Since the availability of IMGT/HighV-QUEST in October 2010, more than 2.2 billions of sequences (from external users) have been submitted. They required more than 1.1 million hours of computational resources. About 52 terabytes of results were generated.



345 of IMGT/HighV-QUEST users are from USA (43%), 301 are from EU (18 countries) (38%), and 156 from other countries (19%).

IMGT-ONTOLOGY concepts

Prototypes of IG and TR V, D, J genes



Prototypes are graphical representation based on the concepts of description

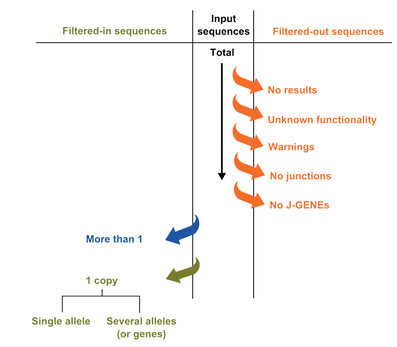
IMGT-ONTOLOGY Concepts

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

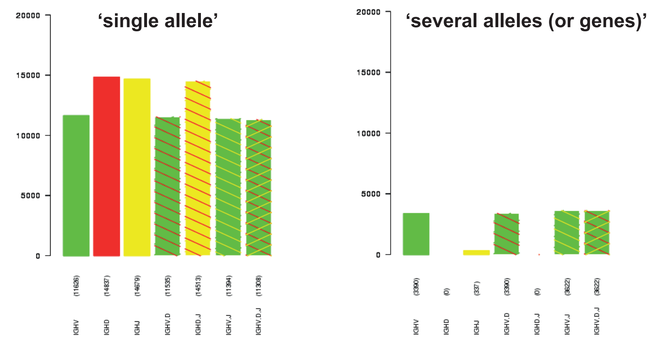
CLASSIFICATION
The concepts of classification allowed to classify and name the human IG and TR genes and alleles which were approved by HGNC and endorsed by WHO-IUIS. They provide the frame for the standardized IG and TR nomenclature of jawed vertebrates.

NUMEROTATION
The concepts of numerotation comprise the 'IMGT unique numbering' and 'IMGT Collier de Perles'.

IMGT/HighV-QUEST based on IMGT® standard

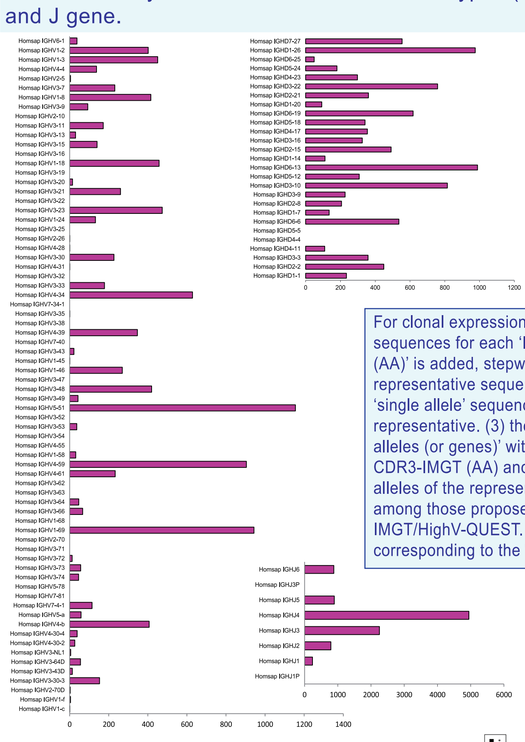


The filtered-in sequences comprise the '1 copy' and the 'More than 1'.
The '1 copy' includes the 'single allele' and 'several alleles (or genes)'. A high proportion of 'single allele' is a good indicator of sequence length and quality.



VH clonal diversity

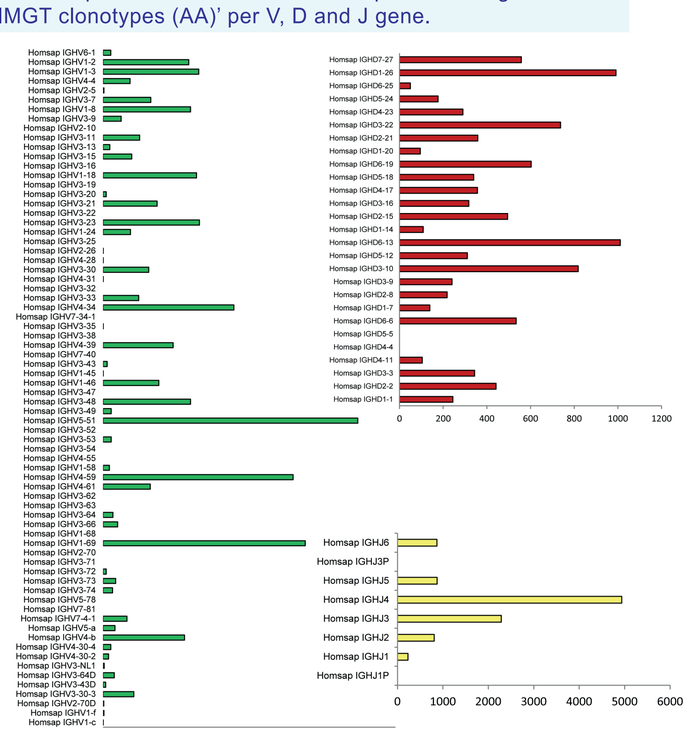
Clonal diversity is the number of 'IMGT clonotypes (AA)' per V, D and J gene.

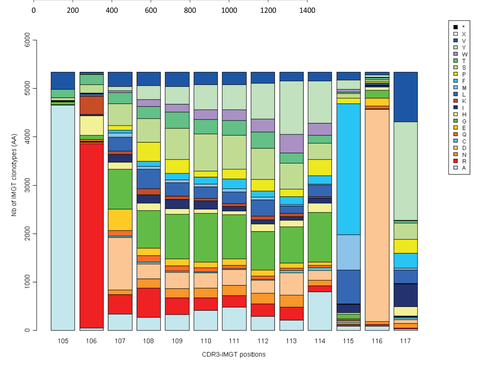


For clonal expression,the number of sequences for each 'IMGT clonotypes (AA)' is added, stepwise: (1) the '1 copy' representative sequence. (2) the '1 copy' 'single allele' sequences not selected as representative. (3) the '1 copy' 'several alleles (or genes)' with the same CDR3-IMGT (AA) and the same V and J alleles of the representative sequence among those proposed by IMGT/HighV-QUEST. (4) the 'More than 1' corresponding to the '1 copy' of steps 1-3.

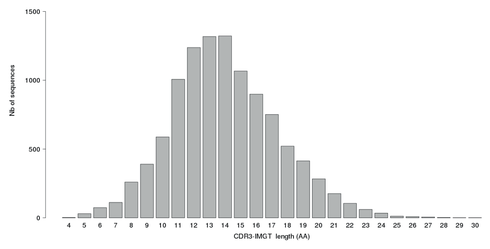
VH clonal expression

Clonal expression is the number of sequences assigned to 'IMGT clonotypes (AA)' per V, D and J gene.





Analysis of the CDR3-IMGT can be performed for clonal diversity and clonal expression. Amino acid frequency can be determined, with their IMGT AA physicochemical classes (Pommié C et al. 2004), at each CDR-IMGT and FR-IMGT position.



Acknowledgments: this work was granted access to the HPC resources of CINES under the allocation 036029-(2010-2014) made by GENCI (Grand Equipement National de Calcul Intensif).

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)

© Copyright 1995-2014 IMGT®, the international ImMunoGeneTics information system®

©2014 Véronique Giudicelli, Eltaf Alamyar, Géraldine Folch, Joumana Jabado-Michaloud and Marie-Paule Lefranc