IMGT/HighV-QUEST for NGS analysis of IG and TR:
statistical analysis of IMGT clonotypes (AA), novel interface and functionalities

Marianne Lèbre, Karthik Kalyan, Patrice Duroux, Véronique Giudicelli, Sofia Kossida , Marie-Paule Lefranc

IMGT®, the international ImMunoGeneTics information system®, Laboratoire d'ImmunoGénétique Moléculaire (LIGM),
Institut de Génétique Humaine (IGH), UMR 9002 CNRS-UM, Université de Montpellier (UM), Montpellier (France)

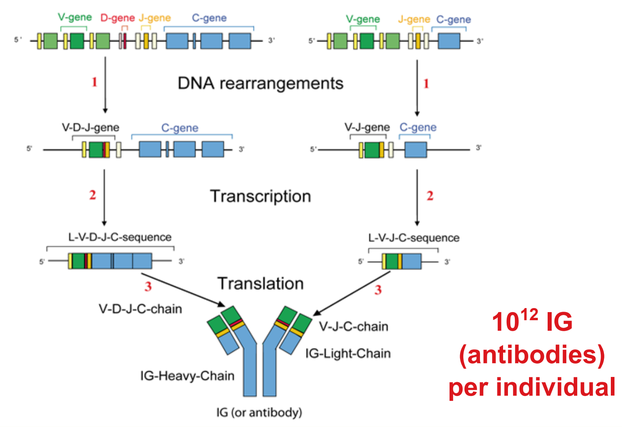
http://www.imgt.org

IMGT®, the international ImMunoGeneTics information system®, http://imgt.org/ [1], is the global reference in immunogenetics and immunoinformatics [2], founded in 1989 by Marie-Paule Lefranc at Montpellier (Université de Montpellier and CNRS). IMGT® is a high-quality integrated knowledge resource specialized in the immunoglobulins (IG) or antibodies, T cell receptors (TR), major histocompatibility (MH) of humans and other vertebrate species. IG and TR are the antigenes receptors for the adaptive immune response which characterizes the vertebrates with jaws (*Gnathostomata*) [2]. Their study in normal and pathological conditions is a challenge due to the huge diversity of the variable domain (V-DOMAIN) at the N-terminal end of each chain (10^{12} potential specificities for humans), which results from complex IG and TR synthesis. Since 2010, IMGT® has developed IMGT/HighV-QUEST [3-7], an online portal for the analysis of the IG and TR immune repertoires obtained through the next generation sequencing (NGS) tecnhologies.

[1] Lefranc M.-P et al. Nucl. Acids Res. 43:D413-422 (2015) PMID: 25378316 [2] Lefranc M.-P. Front. Immunol. 5:22 (2014) PMID: 24600447 [3] Alamyar E. et al. Abstract 60, Poster 27, JOBIM Montpellier (2010) [4] Alamyar E. et al. Immunome Res. 8:1:2 (2012) [5] Li S. et al. Nat. Commun. 4:2333 (2013) PMID: 23995877 [6] Aouinti S. et al. Front. Immunol. 7:339 (2016) PMID: 27667992 [7] Giudicelli V. et al. BMC Immunol. 18(1):35 (2017) PMID: 28651553

IG and TR synthesis

IG and TR are antigens receptors of the adaptive immune response. The synthesis of the V-DOMAIN at the N-terminal end of each IG or TR chain results from genomic DNA rearrangements of variable (V), diversity (D) and joining (J) genes and from junctional diversity.

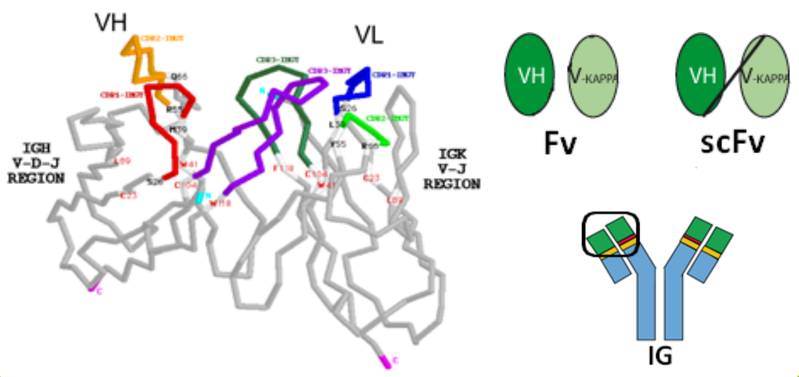


10^{12} IG (antibodies) per individual

V-DOMAIN

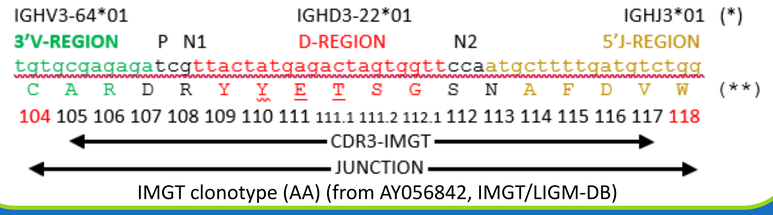
The antigen-binding site is formed by the six complementarity determining regions or CDR-IMGT of two non covalently paired V-DOMAIN (VH/VL for IG, V-ALPHA/V-BETA or V-GAMMA/V-DELTA for TR), defined as Fragment variable (Fv).

An *in vitro* engineered chain made of two V-DOMAIN connected by a linker is a single chain Fragment variable (scFv).



IMGT clonotypes

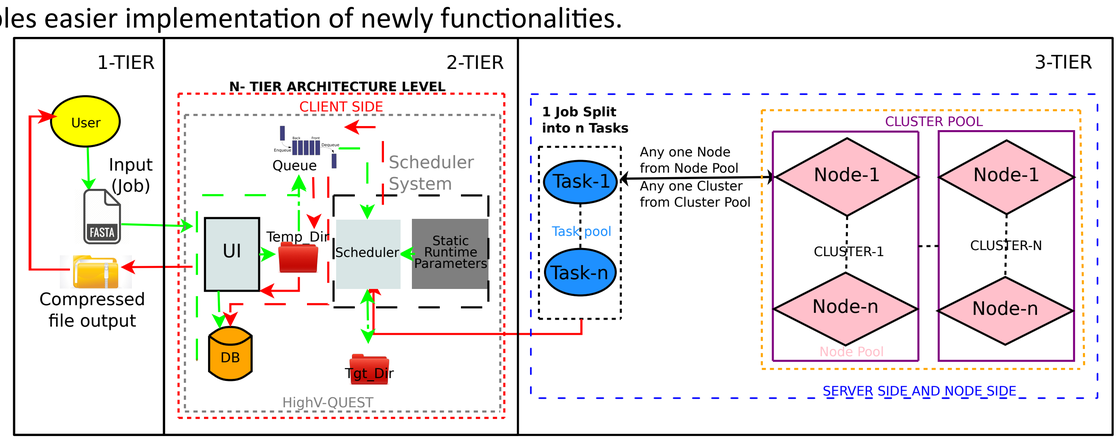
The 'IMGT clonotype' concept provides a molecular characterization of the IG and TR V-DOMAIN repertoires. IMGT clonotypes are identified by analysis of the nucleotide (nt) sequences of the V-DOMAIN (V-(D)-J-REGION), using IMGT/HighV-QUEST for NGS. An 'IMGT clonotype (AA)' is defined as a unique V-(D)-J rearrangement (with IMGT gene and allele names (*)) and a unique CDR3-IMGT amino acid (AA) sequence (**) (from positions 105 to 117, the anchors of the JUNCTION being 2nd-CYS 104 and J-PHE or J-TRP 118). For a given IMGT clonotype (AA), sequences which differ by CDR3-IMGT nt differences are defined as 'IMGT clonotypes (nt)'.



IMGT clonotype (AA) (from AY056842, IMGT/LIGM-DB)

IMGT/HighV-QUEST architecture system

IMGT/HighV-QUEST architecture has moved from a 2-tier to a 3-system: web user interface (UI), database and scheduling- system. The scheduling-sytem is now a standalone system (shell scripts and cron) which has the possibility to be integrated to an automation tool. The 3-tier architecture enables easier implementation of newly functionalities.



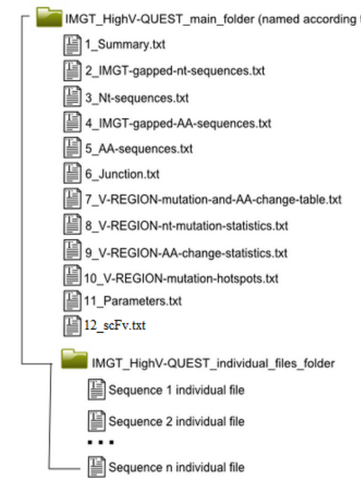
The user can follow the analysis status in the "Analysis history": **queued** (new entry created in the database when data are submitted for a job); **dispatched** (data are sent by the scheduling system for the analysis job when there is enough computational resources available); **completed** (results available and database updated by the scheduling system).

A new client UI based on modern web technologies (Bootstrap, Struts2 and Tiles3) has recently been made available. It also integrates a new e-mail update functionality.

IMGT/HighV-QUEST analysis

IMGT/HighV-QUEST analyzes up to 500,000 IG or TR rearranged sequences per run, with the same degree of resolution and high-quality results as IMGT/V-QUEST (same algorithm, same IMGT reference directories). The tool:

1) identifies, by default, the insertions/deletions (indels) and correct them
2) numbers the user sequences and introduces gaps (IMGT unique numbering)
3) identifies the V, D and J GENE and alleles
4) characterizes IG somatic hypermutations (nt and AA)
5) describes the JUNCTION (IMGT/JunctionAnalysis)
6) provides a complete annotation with IMGT labels (IMGT/Automat)
7) as an option, analyses scFv (added in 2017 [7]).



IMGT/HighV-QUEST results consist of 11 CSV files (or 12 with the scFv option) provided as an archive file. Files #1 to #10 comprise systematically sequence identification, i.e. the sequence name, the functionality, the names of the closest V-GENE and allele, and files #1 to #6 also include the D and J GENEs and alleles. The files #7 to #10 that report the analysis of mutations are used mostly for IG. Files #1 to #10 include one line per submitted sequence, and together may comprise up to 539 columns for a complete results report.

IMGT/HighV-QUEST statistical analysis: IMGT clonotypes (AA) and (nt)

The statistical analysis applies a filter on the IMGT/HighV-QUEST results: only the ones characterized by a V-GENE and allele (single or several alleles), a JUNCTION and a J-GENE and allele (single or several alleles) are filtered-in for statistical analysis. The IMGT/HighV-QUEST statistical analysis, which allows the identification and characterization of the clonotypes [5], may analyse up to one million IMGT/HighV-QUEST results.

A IMGT clonotype (AA and nt) results per locus

The statistical results are provided in 10 sections (HTML pages):
IMGT clonotypes (AA) per Nb (1) without or (2) with detailed clonotypes (nt)
IMGT clonotypes (AA) per V gene (3) without or (4) with detailed clonotypes (nt)
IMGT clonotypes (AA) per CDR3-IMGT length (AA) (5) without or (6) with detailed clonotypes (nt)
IMGT clonotypes (AA) by CRD3-IMGT sequence (AA) alphabetical order with detailed clonotypes (nt) (7)
IMGT clonotype (AA) diversity and expression histograms: per V,(D),J-GENE and per CDR3-IMGT length (8)
IMGT clonotype (AA) diversity and expression tables: per V, (D), J-GENE and per CDR3-IMGT length (9)
V gene and allele table: Rearrangements, Nb of sequences and Nb IMGT clonotypes (AA) per V-GENE and allele (10)

(1)

#	ID Exp. ID	Nb Total Nb of 1 copy	Nb Total Nb of More than 1	Nb Total	IMGT clonotype (AA) definition: V gene and allele	D gene and allele	J gene and allele	CDR3-IMGT length (AA)	CDR3-IMGT sequence (AA)	Anchors 104,118	IMGT clonotype (AA) representative sequence: V %	Sequence length	Functionality	Sequence ID	IMGT clonotypes (nt) Sequences file (1 copy)
1	6874-S5	851	14	865	Homsap IGHV4-39*01 F	Homsap IGHD2-15*01 F, or Homsap IGHD2-2*01 F	Homsap IGHJ3*02 F	12 AA	ARLVASAPGLFF	C,W	95.17	496	productive	SRR1168788.23654 G9YJURD01EUXEQ length=496 NA	Sequences file
2	1501-S5	132	0	132	Homsap IGHV4-34*01 F	Homsap IGHD2-15*01 F	Homsap IGHJ4*02 F	20 AA	ARGRYCSGGSCRMTQSYFDY	C,W	100	519	productive	SRR1168788.64718 G9YJURD01B5B0 length=519 NA	Sequences file
3	6146-S5	124	7	131	Homsap IGHV1-18*01 F, or Homsap IGHV1-18*04 F	Homsap IGHD2-8*01 F, or Homsap IGHD3-10*01 F	Homsap IGHJ4*02 F	13 AA	ARDIMASTRGLFY	C,W	89.58	468	productive	SRR1168788.35516 G9YJURD01EUECZ length=468 NA	Sequences file
4	7616-S5	105	1	106	Homsap IGHV4-59*01 F	Homsap IGHD5-24*01 F, or Homsap IGHD5-12*01 F	Homsap IGHJ3*02 F	10 AA	ARVGSNAFDI	C,W	100	489	productive	SRR1168788.46912 G9YJURD01BZPNA length=489 NA	Sequences file
5	6881-S5	102	2	104	Homsap IGHV4-39*01 F	Homsap IGHD2-15*01 F	Homsap IGHJ4*02 F	12 AA	ARLVASAPGXFF	C,W	94.85	501	productive	SRR1168788.16705 G9YJURD01EWEJ3 length=501 NA	Sequences file
6	3668-S5	94	0	94	Homsap IGHV1-8*01 F	Homsap IGHD3-3*01 F	Homsap IGHJ3*02 F	16 AA	AKGGLWSGSRDTFNM	C,W	88.89	501	productive	SRR1168788.13130 G9YJURD01BGFNZ length=501 NA	Sequences file
7	7372-S5	92	0	92	Homsap IGHV4-31*03 F	Homsap IGHD2-2*01 F	Homsap IGHJ4*02 F	11 AA	ARVVPAHNDY	C,W	100	494	productive	SRR1168788.1255 G9YJURD01CLTF7 length=494 NA	Sequences file

(4)

4089	48	1	gcgagacatcccgggglancagtggctgglgtgatgctttlgatatcg	0	Homsap IGHV4-39*01 F	Homsap IGHD6-19*01 F	Homsap IGHJ3*01 F, or Homsap IGHJ3*02 F	C,W	98.97	299	74	34	499	1	0	1					
4090	4090-S5	16 AA	ARHPPLGWRGVGDFQH	Homsap IGHV4-59*08 F	Homsap IGHD2-21*01 F	Homsap IGHJ1*01 F	C,W	87.68	502	productive	SRR1168788.53896 G9YJURD01C7TBO length=502 NA			101	Sequences file						
4090	48	1	gcgagacatccgccctlgggglggagaggcgttggagacttccagcac	0	Homsap IGHV4-59*08 F	Homsap IGHD2-21*01 F	Homsap IGHJ1*01 F	C,W	87.68	287	71.15	50	502	1	0	1					
4091	4091-S5	16 AA	ARHQRSGGNDFSPFDY	Homsap IGHV5-51*01 F	Homsap IGHD5-12*01 F	Homsap IGHJ4*02 F	C,W	87.55	518	productive	SRR1168788.49302 G9YJURD01DK9DE length=518 NA			202	Sequences file						
4091	48	2	gcgagacatcagcgatctggtggcaacgatttctctcccttgacttc	1	Homsap IGHV5-51*01 F	Homsap IGHD5-12*01 F	Homsap IGHJ4*02 F, or Homsap IGHJ4*03 F	C,W	86.11	296	81.25	44	512	1	0	1					
			gcgagacatcagcgatctggtggcaacgatttctctcctttgacttc	0	Homsap IGHV5-51*01 F	Homsap IGHD5-12*01 F	Homsap IGHJ4*02 F, or Homsap IGHJ4*03 F	C,W	87.5	296	81.25	43	518	1	0	1					
4092	4092-S5	16 AA	ARHRPYSSGWYVYFDY	Homsap IGHV4-59*08 F	Homsap IGHD6-19*01 F	Homsap IGHJ4*02 F	C,W	88.95	503	productive	SRR1168788.62333 G9YJURD01D5CB1 length=503 NA			160	16	Sequences file					
4092	48	1	gcgagacataggccgtatagcagtggctggtacgtctacttlgactac	0	Homsap IGHV4-59*08 F	Homsap IGHD6-19*01 F	Homsap IGHJ4*02 F, or Homsap IGHJ4*03 F	C,W	98.62	292	96.53	47	498	15	0	15					
					Homsap IGHV4-59*08 F	Homsap IGHD6-19*01 F	Homsap IGHJ4*02 F, or Homsap IGHJ4*03 F	C,W	100	293	70.83	47	507	1	0	1					

B IMGT clonotype (AA) results comparison per locus

(5)

ARGGKVMITPVYWHFDL											
34	7090-S2	17 AA	ARGGKVMITPVYWHFDL	Homsap IGHV1-2*02 F	Homsap IGHD3-16*01 F	Homsap IGHJ2*01 F	C104,W118	92.01	505	productive	SRR1168750 G9YJURD01 length=505
	10676-S3	17 AA	ARGGKVMITPVYWHFDL	Homsap IGHV1-2*02 F	Homsap IGHD3-16*01 F	Homsap IGHJ2*01 F	C104,W118	93.06	496	productive	SRR1168750 G9YJURD01 length=496
ARGHCSGGSCFSGTFDY											
35	7112-S2	17 AA	ARGHCSGGSCFSGTFDY	Homsap IGHV4-31*03 F	Homsap IGHD2-15*01 F	Homsap IGHJ4*02 F	C104,W118	95.88	513	productive	SRR1168750 G9YJURD01 length=513
	10729-S3	17 AA	ARGHCSGGSCFSGTFDY	Homsap IGHV4-31*03 F	Homsap IGHD2-15*01 F	Homsap IGHJ4*02 F	C104,W118	95.88	509	productive	SRR1168750 G9YJURD01 length=509
	10730-S3	17 AA	ARGHCSGGSCFSGTFDY	Homsap IGHV4-31*02 F	Homsap IGHD2-15*01 F	Homsap IGHJ4*02 F	C104,W118	95.88	510	productive	SRR1168750 G9YJURD01 length=510
ARHVRTRDWNDMYYFDY											
36	7283-S2	17 AA	ARHVRTRDWNDMYYFDY	Homsap IGHV4-38*01 F	Homsap IGHD1-1*01 F	Homsap IGHJ4*02 F	C104,W118	95.88	501	productive	SRR1168750 G9YJURD01 length=501
	11097-S3	17 AA	ARHVRTRDWNDMYYFDY	Homsap IGHV4-39*01 F	Homsap IGHD1-1*01 F	Homsap IGHJ4*02 F	C104,W118	95.88	510	productive	SRR1168750 G9YJURD01 length=510
ARVSPNNWNDDGYYMDV											
37	7533-S2	17 AA	ARVSPNNWNDDGYYMDV	Homsap IGHV4-61*02 F	Homsap IGHD1-1*01 F	Homsap IGHJ6*03 F	C104,W118	95.53	506	productive	SRR1168750 G9YJURD01 length=506
	11556-S3	17 AA	ARVSPNNWNDDGYYMDV	Homsap IGHV4-61*02 F	Homsap IGHD1-1*01 F	Homsap IGHJ6*03 F	C104,W118	93.1	498	productive	SRR1168750 G9YJURD01 length=498

The IMGT clonotypes (AA) comparison per locus has several displays. The 'Full results' display provides the same results as 'IMGT clonotypes (AA) per CDR3-IMGT length (AA)' (5) but sorted here by IMGT clonotypes (AA) present in a single batch and IMGT clonotypes (AA) common to 2 (or more) batches (lightpink and light yellow lines).

C IMGT/StatClonotype

IMGT/StatClonotype [6] is a tool, downloadable on the IMGT® site, which allows evaluating and exploring, between sets, the significance of pairwise comparison of IMGT clonotype (AA) diversity and expression per V, D and J gene. The IMGT/HighV-QUEST statistical ouput contains, in 'data' directory, txz file(s) designated as stats_xxx, where 'xxx' is the batch name and the locus type. At least two 'stats_xxx' files are needed to launch a comparative analysis in IMGT/StatClonotype. Integrated in the R package "IMGTStatClonotype", the tool offers a graphical interface to visualize pair wise comparison, per IMGT genes and alleles, of the IMGT clonotype (AA) diversity or expression of any IG or TR immunoprofiles of any species.

IMGT® founder and executive director emeritus: Marie-Paule Lefranc (Marie-Paule.Lefranc@igh.cnrs.fr)
IMGT® director: Sofia Kossida (Sofia.Kossida@igh.cnrs.fr)
Bioinformatics manager: Véronique Giudicelli (Veronique.Giudicelli@igh.cnrs.fr)
Computer manager: Patrice Duroux (Patrice.Duroux@igh.cnrs.fr)

This work is granted access to HPC@LR and to the HPC ressources of the CINES and to the TGCC of the CEA under the allocation [036029] (2010-2018) made by GENCY.