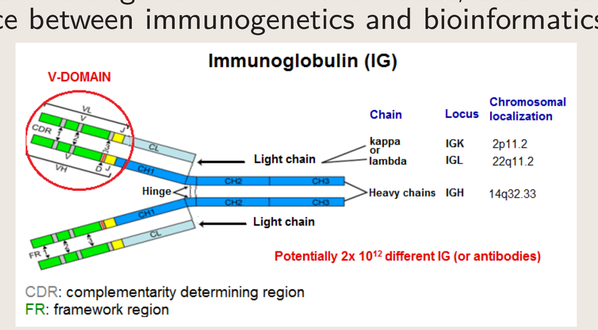
Statistical analysis of somatic mutations in immunoglobulin variable region from IMGT/HighV-QUEST output

Safa Aouinti, Véronique Giudicelli, Patrice Duroux, Sofia Kossida, Marie-Paule Lefranc

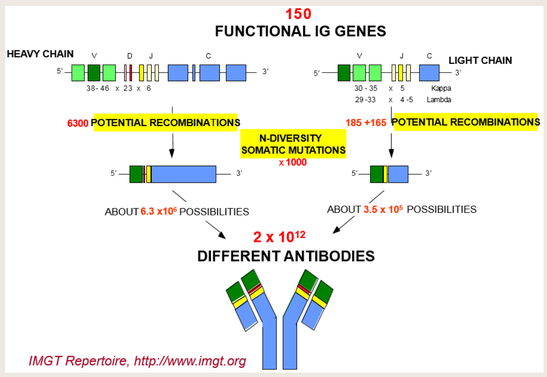
IMGT®, the international ImMunoGeneTics information system®, Laboratoire d'ImmunoGénétique Moléculaire (LIGM), Institut de Génétique Humaine (IGH), UMR 9002, CNRS, Montpellier University, Montpellier (France)

Biological context

The adaptive immune response is our ability to produce up to 2.10^{12} different immunoglobulins (IG) or antibodies and T cell receptors (TR) per individual to fight pathogens. IMGT®, the international ImMunoGeneTics information system®, was created in 1989 in Montpellier, France (CNRS and Montpellier University) to manage the huge and complex diversity of these antigen receptors [1-2], and is at the origin of immunoinformatics, a science at the interface between immunogenetics and bioinformatics [2].



1- V-Domain combinatorial diversity



2- V-Domain junctional diversity

- V-D-J and V-J rearrangements in bone marrow.

Input	V name	3'V-REGION	N1	D-REGION	N2	5'J-REGION	J name	D name
AJ399822	Homsap IGHV1-3*01	tgtgcgagaga	att	.....ctatggt...	g	...acttggctactgg	Homsap IGHJ4*02	Homsap IGHD5-18*01
AB012909	Homsap IGHV3-30*04	tgtgcgag...	g	.....acagctttctttta.	acgc	....ctatgactttctgg	Homsap IGHJ4*02	Homsap IGHD5-18*01
AB027443	Homsap IGHV3-21*01	tgtgcgagag.		.gtacgtaggga.....	t	...actggttcgacccctgg	Homsap IGHJ5*02	Homsap IGHD1-26*01
AF015123	Homsap IGHV3-23*01	tgtgcga....	tcaacacaca	...gatatagtggcta.....		.....cgacctctgg	Homsap IGHJ5*02	Homsap IGHD5-12*01
AB021521	Homsap IGHV3-23*01	tgtgcgaaaga		.....ctatgatagtagtggttat.....	cc	.....tgactactgg	Homsap IGHJ4*02	Homsap IGHD3-22*01
AB063683	Homsap IGHV1-18*01	tgtgcgagag.		...gagttgg.....	ttc	....ctttgactactgg	Homsap IGHJ4*02	Homsap IGHD5-24*01

Enzymes: exonuclease (deletion of nucleotides (nt)).
TdT (addition of nt) → N-REGION (N1, N2).

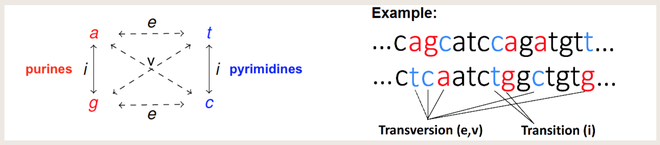
3- V-Domain mutational diversity

- Somatic mutations in B cells of lymph nodes and spleen.

Enzyme: AICDA (activation induced cytidine deaminase).

Standardized description

2 types of nucleotide mutations can be distinguished: transition (i) and transversion (e, v)

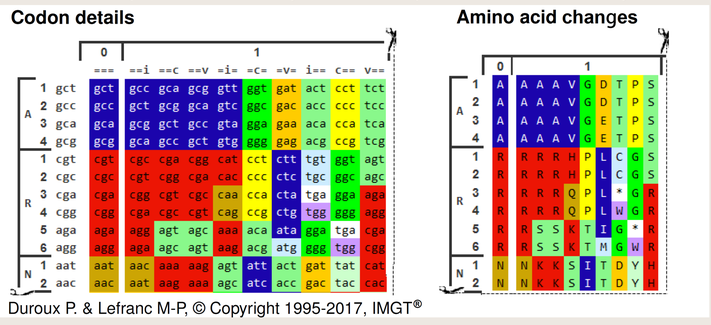


At the codon level, the consequence of the nucleotide mutation can be:

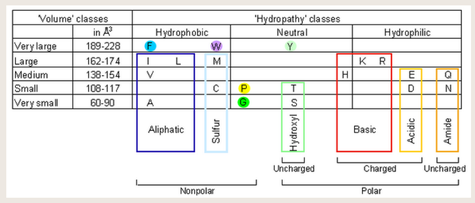
- **Silence (S):** no AA change
- **Replacement (R):** AA change

Example:
. Q H P D V .
...cag cat cca gat gtt ...
...ctc aat ctg gct gtg ...
. L N L A V .

Expected R (AA change) and S (no AA change):



IMGT 'Physicochemical' classes of the 20 common AA:



IMGT/HighV-QUEST for NGS analysis

IMGT/HighV-QUEST [3-6], the first web portal for next generation sequencing (NGS) analysis of IG and TR, provides the identification of the variable (V), diversity (D) and joining (J) genes and alleles, analysis of the V-(D)-J junction and characterization of the 'IMGT clonotype (AA)' (AA for amino acid).

Sequence nb	V-GENE and allele	V-REGION identity %	V-REGION mutations
31	Homsap IGHV3-30-3*01 F	97.56	g6>c, V2; V2 gtg 4-6>V gtc|c7>a,Q3>T(- - -); Q3 cag
34	Homsap IGHV4-61*02 F	96.56	a2>c,Q1>P(- - -); Q1 cag 1-3>P ccc|g3>c,Q1>P(- - -)
64	Homsap IGHV3-21*01 F	99.65	g1>c,E1>Q(+ + -); E1 gag 1-3>Q cag|

It describes the V-REGION mutations and determines the hotspot positions in the closest germline V gene.

Statistical procedure

Data

- 43,558 reads of plasmacytes B cells isolated from a healthy female subject
- NGS technology: Roche GLF-LX 454
- NCBI Sequence Read Archive accession code: SRR1168790

Study purpose

- Development of an algorithm for statistical analysis of somatic mutations from IMGT/HighV-QUEST output.
- Development of a user-friendly tool for users.

R and S analysis

- Calculation of observed and expected number of AA changes (R) based on germline.
- Application of binomial [8] and multinomial [9] models to test significance between observed and expected R in FR-IMGT and CDR-IMGT under the null hypothesis of no selection $\mathbf{H_0}$: the difference between the observed and expected number of R is due to chance only.

Binomial ~ multinomial distribution model [10]:

Binomial test	Multinomial test
To calculate p-value of producing exactly the number of observed R.	To calculate p-value of producing either the observed number of R or more extreme values.
Tests both for an excess or scarcity in R for CDR-IMGT and FR-IMGT.	Specify the direction of selection when calculating the p-value: - positive selection for CDR-IMGT - negative selection for FR-IMGT

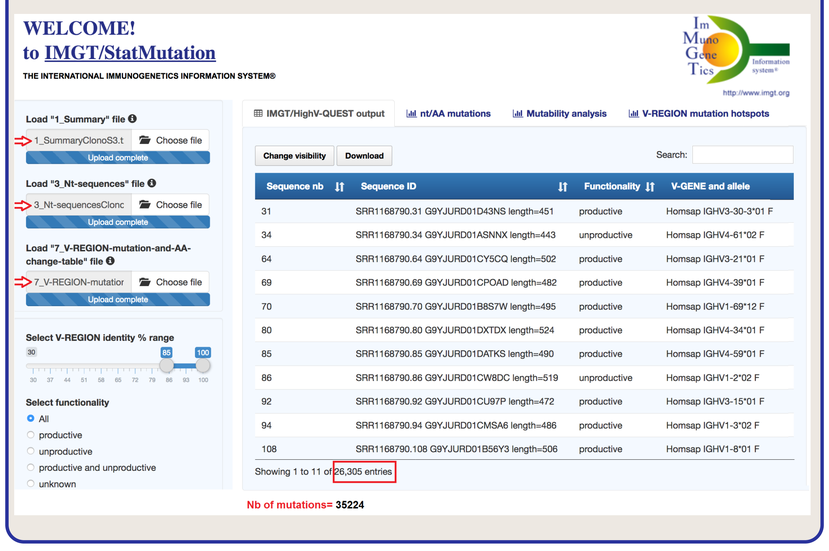
'IMGTStatMutation' R package

'IMGTStatMutation' is an R package for statistical analysis of somatic mutations in IG V-REGION from IMGT/HighV-QUEST output.
It includes a procedure to visualize mutations, calculate and detect significant differences between observed and expected number of AA changes (R) in FR-IMGT and CDR-IMGT.
'IMGTStatMutation' incorporates a userfriendly web interface, allowing use of the IMGT/StatMutation tool, in users' own browser.

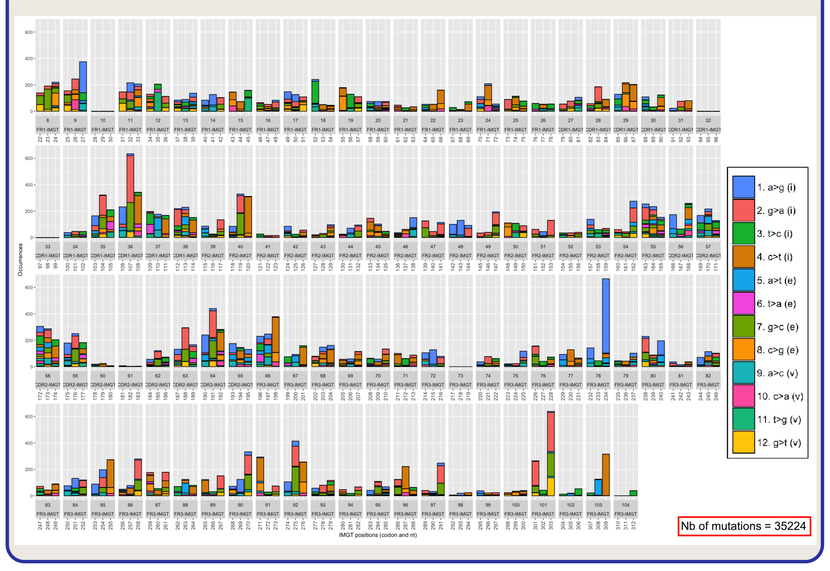
IMGT/StatMutation web tool

IMGT/StatMutation is an IMGT® [1] tool for statistical analysis of somatic mutations from IMGT/HighV-QUEST output [3-6]. IMGT/StatMutation uses a statistical procedure for detecting significant differences in observed and expected number of AA changes (R) in FR-IMGT and CDR-IMGT.
It uses the IMGT gene and allele nomenclature based on IMGT-ONTOLOGY [7] and IMGT standards in immunoinformatics [2]. IMGT/StatMutation is a user-friendly web interface in users' own browser.

IMGT/StatMutation web interface

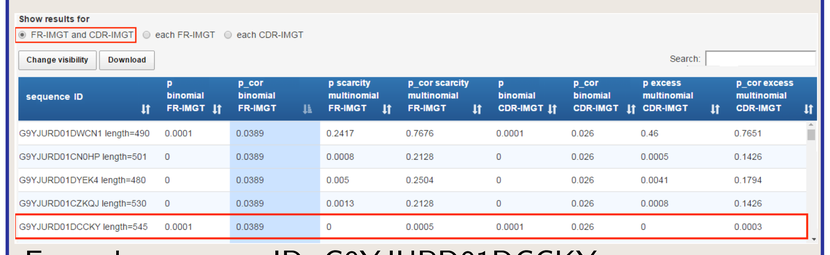


Nucleotide mutation types



Replacement (R) and Silence (S) occurrences

R and S analysis results



Example sequence ID: G9YJURD01DCCKY

Results for CDR-IMGT

N	R observed CDR-IMGT	R expected CDR-IMGT	p binomial CDR-IMGT	p_cor binomial CDR-IMGT	p excess multinomial CDR-IMGT	p_cor excess multinomial CDR-IMGT
21	3	2.9457	0.2417	0.7676	0.46	0.7651
14	7	1.8212	0.0008	0.2128	0.0005	0.1426
24	8	2.9861	0.005	0.2504	0.0041	0.1794
14	7	1.9616	0.0013	0.2128	0.0008	0.1426
18	12	2.2025	0	0.0005	0	0.0003

Results for FR-IMGT

N	R observed FR-IMGT	R expected FR-IMGT	p binomial FR-IMGT	p_cor binomial FR-IMGT	p scarcity multinomial FR-IMGT	p_cor scarcity multinomial FR-IMGT
21	4	12.9039	0.0001	0.0389	0.0001	0.026
14	1	8.6472	0	0.0389	0	0.026
24	5	14.9509	0	0.0389	0	0.026
14	1	8.5151	0	0.0389	0	0.026
18	3	11.1894	0.0001	0.0389	0.0001	0.026

Conclusion

IMGT/StatMutation answers the need for somatic mutations statistical analysis from IMGT/HighV-QUEST output of high throughput IG repertoire. It provides a standardized study of mutations and AA changes which are of prime importance for the specificity and affinity of antibodies during protective (vaccination, cancers and infections) or pathogenic (autoimmunity and lymphoproliferative disorders) immune responses.

References: [1] Lefranc M-P et al., Nucleic Acids Res., 43:413-422, 2015. [2] Lefranc M-P, Front. Immunol., 5:22, 2014. [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, 2013. [6] Giudicelli V et al., AutoImmun Infec. Dis., 1(1), 2015. [7] Giudicelli V and Lefranc M-P, Front. Genet., 3:79, 2012. [8] Chang B and Casali P, Immunol. Today, 15(8):367-373, 1994. [9] Lossos I.S et al., J., Immunol., 165(9):5122-6, 2000. [10] Hershberg U et al., Int. Immunol., 20(5):683-69, 2008.

Acknowledgments: this work was granted access to the HPC resources of HPC@LR and of CINES and TGCC-CEA under the allocation 036029-(2010-2017) made by GENCI.

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 (Veroniqe.Giudicelli@igh.cnrs.fr)
Computer manager: Patrice Duroux (Patrice.Duroux@igh.cnrs.fr)