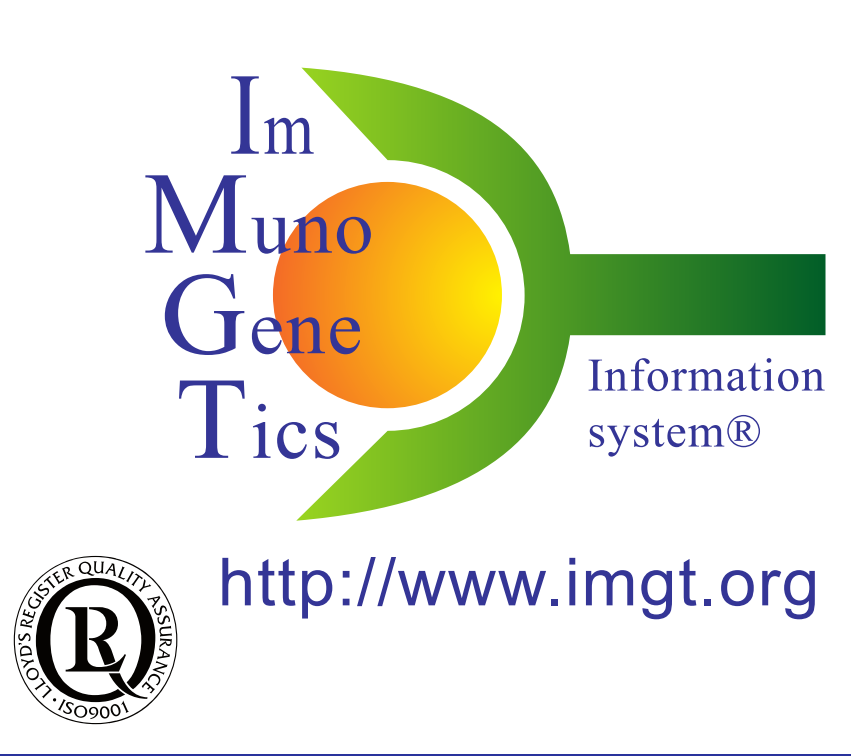


Statistical analysis of somatic mutations in immunoglobulin variable region from IMGT/HighV-QUEST output

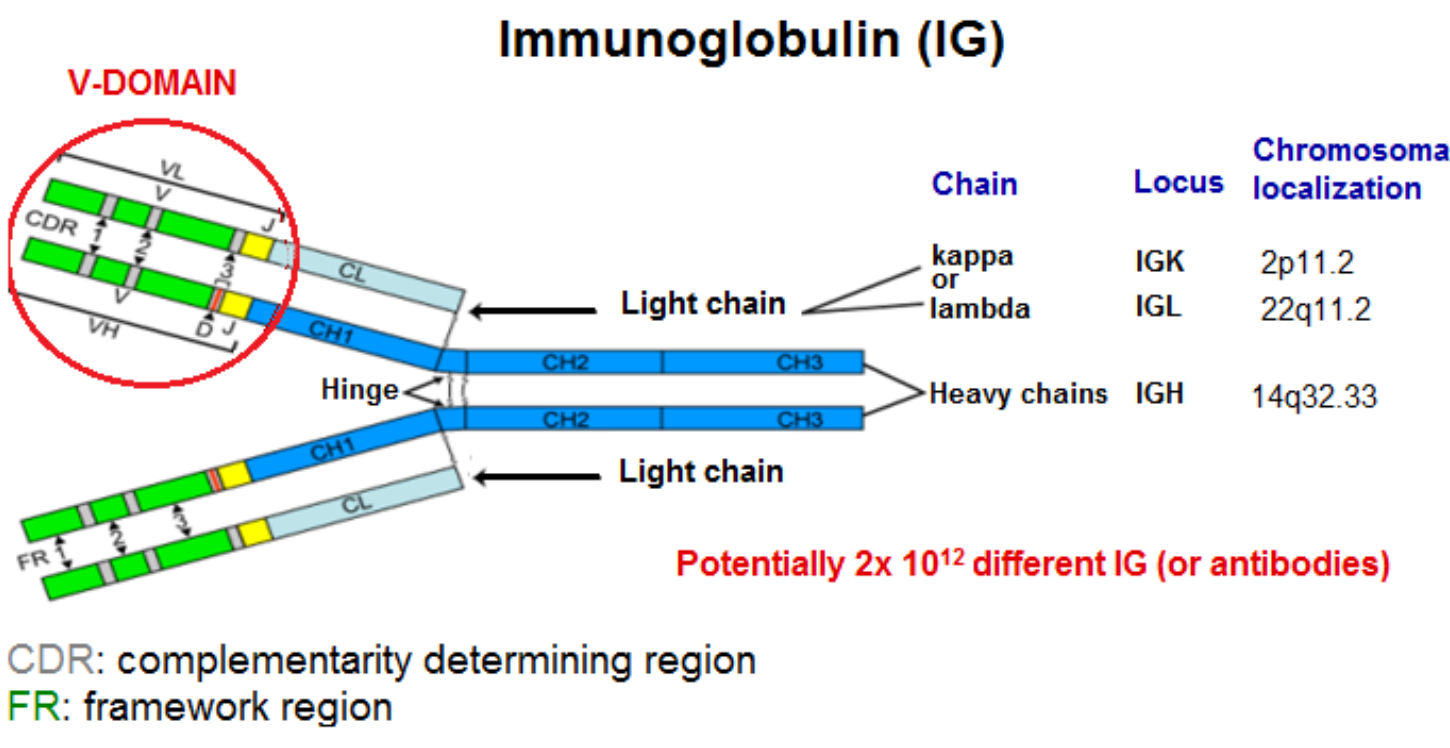
Safa Aouinti, Véronique Giudicelli, Patrice Duroux, Sofia Kossida, Marie-Paule Lefranc



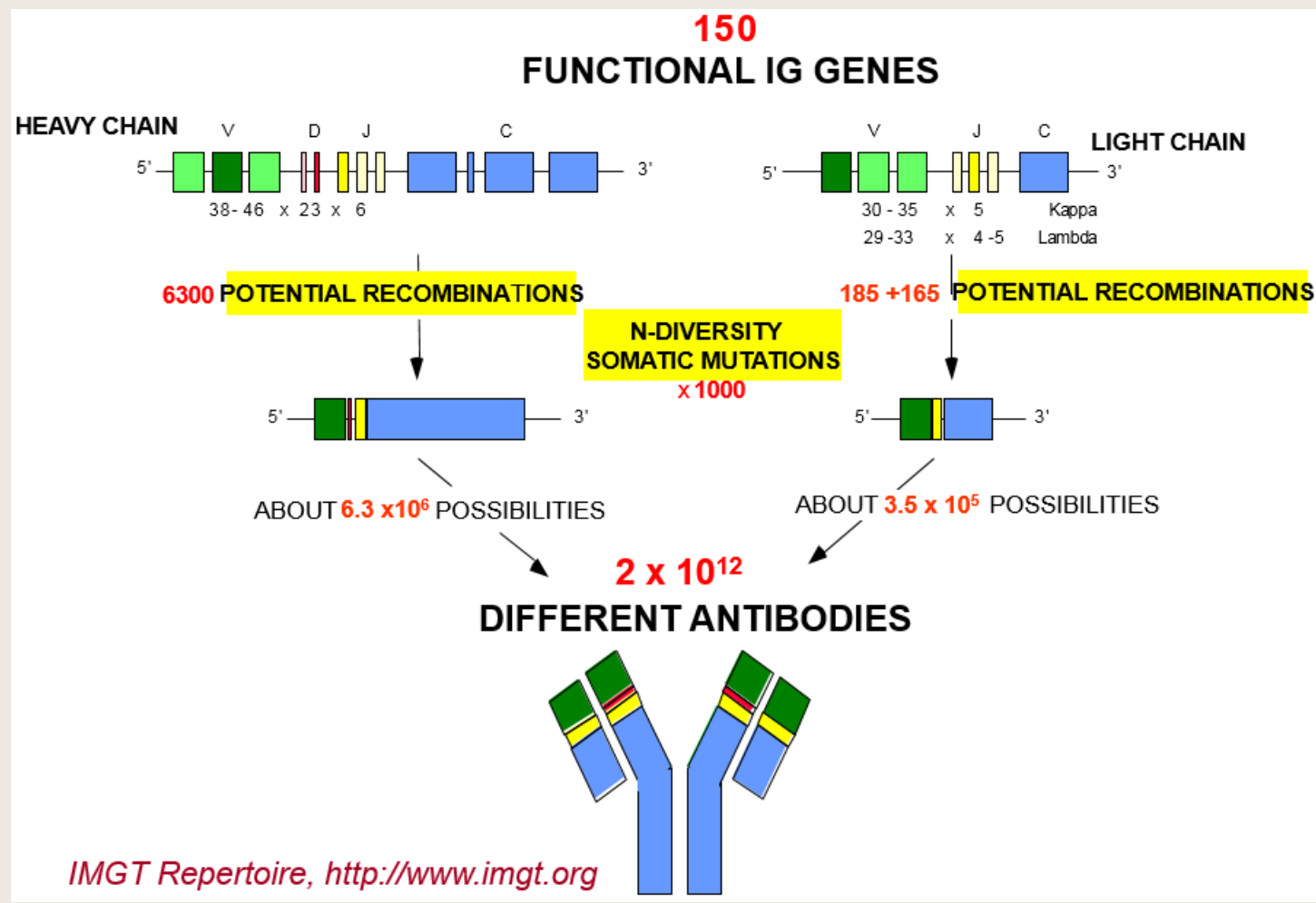
IMGT®, the international ImMunoGeneTics information system®, Laboratoire d'ImmunoGénétique Moléculaire (LIGM), Institut de Génétique Humaine (IGH), UMR 9002, CNRS, Montpellier University, Montpellier (France)

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	Hommap_IGHV1-3*01	ctctggagaga	att	ctatggtt...	g	actgggtctactgg	Hommap_IGHD1*02	Hommap_IGHD5-18*01
AB012909	Hommap_IGHV3-30*04	ctctggagaga	g	acagattggtt...	agcg	ctatggtctactgg	Hommap_IGHD1*02	Hommap_IGHD5-18*01
AB027443	Hommap_IGHV3-21*01	ctctggagaga	g	ctatggtt...	t	ctatggtctactgg	Hommap_IGHD1*02	Hommap_IGHD5-18*01
AF015123	Hommap_IGHV3-23*01	ctctggagaga	ctacacacaa	ctatggtt...	cc	ctatggtctactgg	Hommap_IGHD1*02	Hommap_IGHD5-18*01
AB021021	Hommap_IGHV3-23*01	ctctggagaga	g	ctatggtt...	cc	ctatggtctactgg	Hommap_IGHD1*02	Hommap_IGHD5-18*01
AB026983	Hommap_IGHV3-18*01	ctctggagaga	g	ctatggtt...	ctc	ctatggtctactgg	Hommap_IGHD1*02	Hommap_IGHD5-18*01

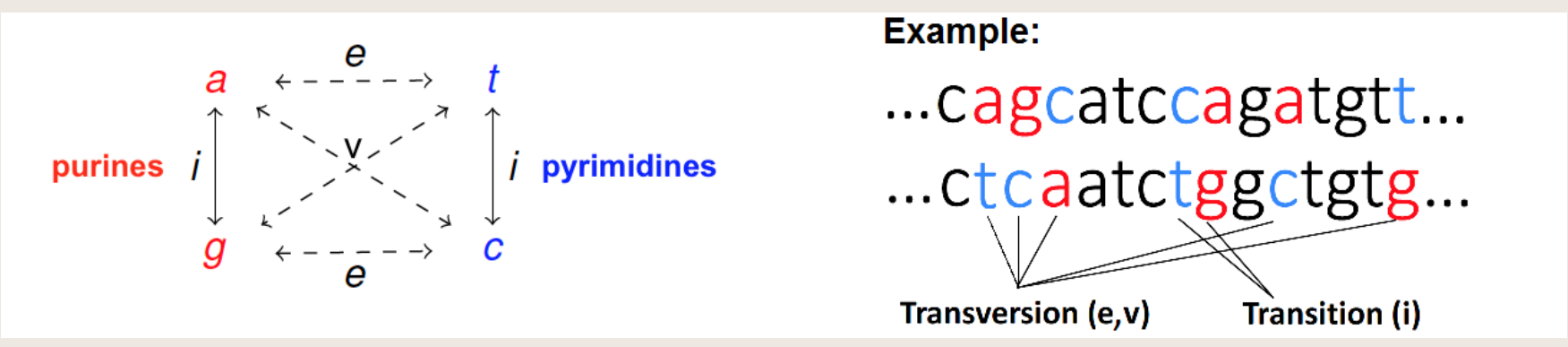
Enzymes: exonuclease (deletion of nucleotides (nt)).
TdT (addition of nt) → N-REGION (N1, N2).

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):

Codon details	Amino acid changes
1 gct gct gcc gca ggc gtt gat act cct cct	1 A A A A V G D T P S
2 gcc gct ggc gca ggc gac acc ccc tcc	2 A A A A V G E T P P
3 gca gca ggc gct gcc gca acc cca tca	3 A A A A V G E T P P
4 ggc gca ggc gct gtc ggc acc cgc tgc	4 A A A A V G E T P P
1 cgt cgt cgc cga cgc cat cct ctt tgt ggt agt	1 R R R R R P L C G C
2 cgc cgc cgt cgc cga ccc ctc tgc agc agc	2 R R R R R P L C G C
3 cga cga cgt cgc cga cca cta tga aga aga	3 R R R R R P L C G C
4 cgg cga cgt cgc cga atc atg tgg aga aga	4 R R R R R P L C G C
5 aga aga agt agc aga aca ata gga tga cga	5 R R S S S T I G A F
6 agg aga agc agt aga acg atg tgg aga aga	6 R R S S S T I G A F
1 aat aat aac aag aag agt att act gat tat cat	1 N N K K S I T D Y
2 aac aac aag aag aag agc acc acc tac cag	2 N N K K S I T D Y

IMGT 'Physicochemical' classes of the 20 common AA:

'Volume' classes	'Hydropathy' classes
Very large 189-229	Hydrophobic
Large 161-174	Neutral
Medium 138-154	Hydrophilic
Small 108-117	Aliphatic
Very small 60-90	Sulfur

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	Hommap_IGHV3-30*3*01 F	97.56	g6>c, V2; V2 gtlg 4-6>V gtlc[7>a.Q3>T(- -); Q3 cag
34	Hommap_IGHV4-61*02 F	96.56	a2>c.Q1>P(- - -); Q1 cag 1-3>P ccc[3>c.Q1>P(- - -)
64	Hommap_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 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

WELCOME!
to IMGT/StatMutation
THE INTERNATIONAL IMMUNOGENETICS INFORMATION SYSTEM

Load "1_Summary" file
1. Summary/Clonotype file
2. Nt-sequences/Clonotype file
3. Nt-sequences/Clonotype file
4. V-REGION-mutation-and-AA-change-table file
5. V-REGION-mutation-and-AA-change-table file

Select V-REGION identity % range
100%
99%
98%
97%
96%
95%
94%
93%
92%
91%
90%
89%
88%
87%
86%
85%
84%
83%
82%
81%
80%
79%
78%
77%
76%
75%
74%
73%
72%
71%
70%
69%
68%
67%
66%
65%
64%
63%
62%
61%
60%
59%
58%
57%
56%
55%
54%
53%
52%
51%
50%
49%
48%
47%
46%
45%
44%
43%
42%
41%
40%
39%
38%
37%
36%
35%
34%
33%
32%
31%
30%
29%
28%
27%
26%
25%
24%
23%
22%
21%
20%
19%
18%
17%
16%
15%
14%
13%
12%
11%
10%
9%
8%
7%
6%
5%
4%
3%
2%
1%
0%

Select functionality
All
productive
unproductive
productive and unproductive
unknown

IMGT/HighV-QUEST output
nt/AA mutations
Mutability analysis
V-REGION mutation hotspots

Change visibility
Download

Sequence nb
31
34
64
69
70
80
85
86
92
94
108

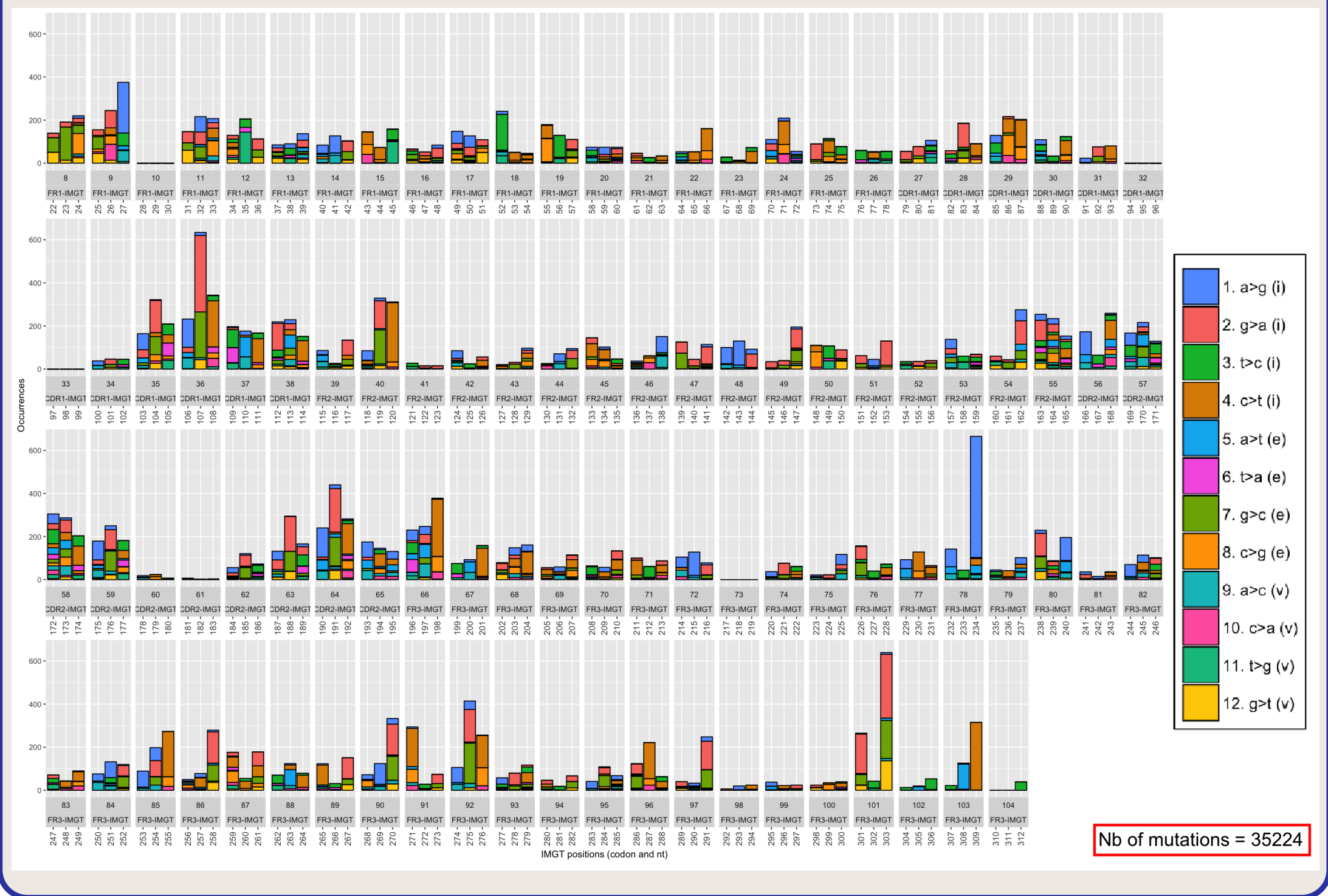
Sequence ID
SRR1168790.31.GBYJURD01D43N1 length=451
SRR1168790.34.GBYJURD01D43N1 length=443
SRR1168790.64.GBYJURD01CYSQ length=502
SRR1168790.69.GBYJURD01CPOAD length=482
SRR1168790.70.GBYJURD01BBS7W length=495
SRR1168790.80.GBYJURD01DXTDK length=524
SRR1168790.85.GBYJURD01D43N1 length=490
SRR1168790.86.GBYJURD01CWBDC length=519
SRR1168790.92.GBYJURD01C9U7P length=472
SRR1168790.94.GBYJURD01CMSA6 length=488
SRR1168790.108.GBYJURD01B56Y3 length=508

Functionality
productive
unproductive
productive and unproductive
unknown

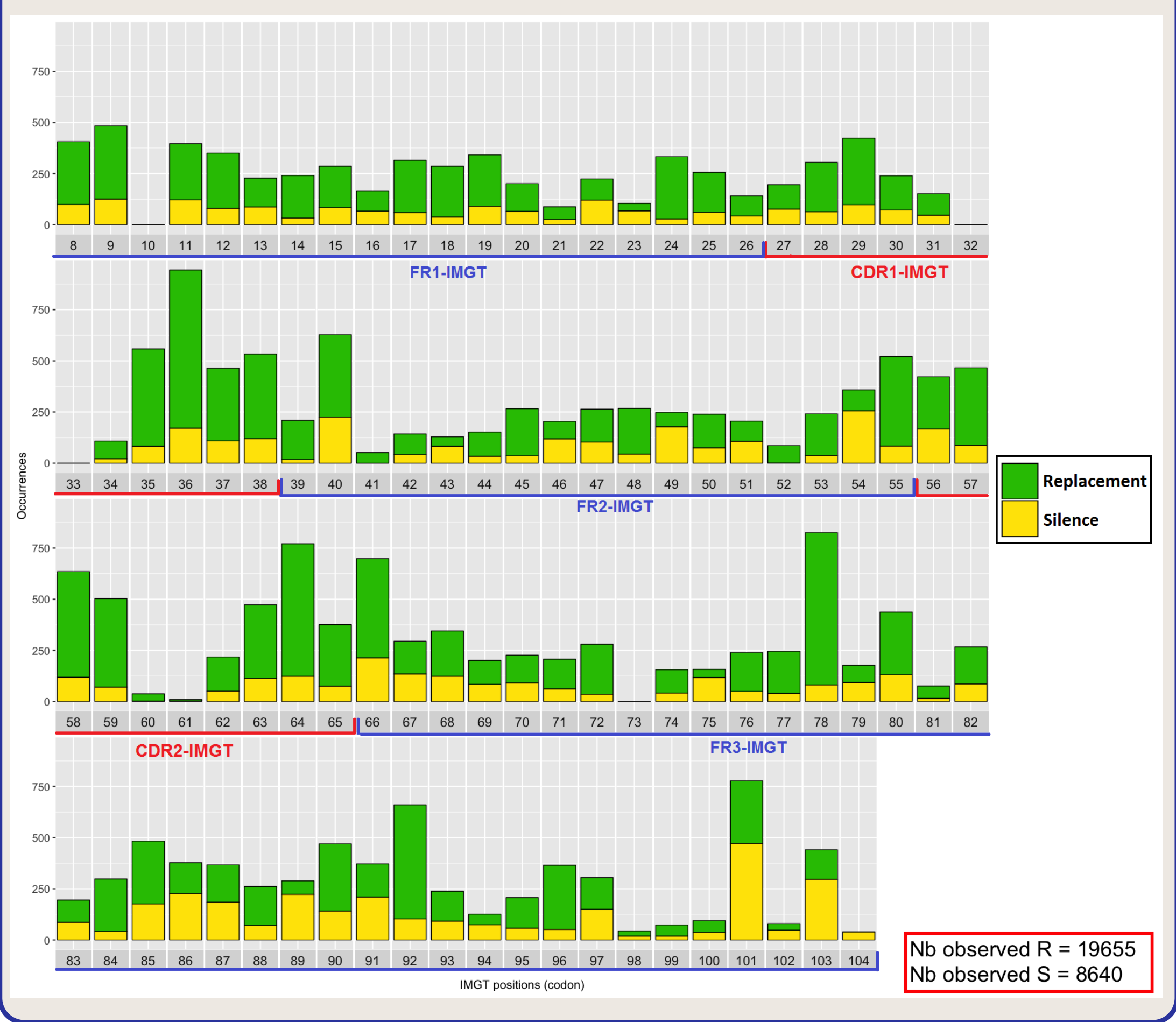
V-GENE and allele
Hommap_IGHV3-30*3*01 F
Hommap_IGHV4-61*02 F
Hommap_IGHV3-21*01 F
Hommap_IGHV4-39*01 F
Hommap_IGHV1-69*12 F
Hommap_IGHV4-34*01 F
Hommap_IGHV4-59*01 F
Hommap_IGHV1-2*02 F
Hommap_IGHV3-15*01 F
Hommap_IGHV1-3*02 F
Hommap_IGHV1-8*01 F

Nb of mutations= 35224

Nucleotide mutation types



Replacement (R) and Silence (S) occurrences



R and S analysis results

sequence ID	p binomial FR-IMGT	p binomial CDR-IMGT	p scarcity FR-IMGT	p scarcity CDR-IMGT	p binomial FR-IMGT	p binomial CDR-IMGT	p excess FR-IMGT	p excess CDR-IMGT
GBYJURD01D43N1 length=450	0.0001	0.0389	0.2417	0.7676	0.0001	0.026	0.46	0.7651
GBYJURD01D43N1 length=443	0	0.0389	0.0008	0.2128	0	0.026	0.0005	0.1426
GBYJURD01D43N1 length=443	0	0.0389	0.0008	0.2128	0	0.026	0.0041	0.1794
GBYJURD01D43N1 length=482	0	0.0389	0.0013	0.2128	0	0.026	0.0008	0.1426
GBYJURD01D43N1 length=445	0.0001	0.0389	0	0.0005	0.0001	0.026	0	0.0003

N	R observed CDR-IMGT	R expected CDR-IMGT	p binomial FR-IMGT	p binomial CDR-IMGT	p excess FR-IMGT	p excess 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

N	R observed FR-IMGT	R expected FR-IMGT	p binomial FR-IMGT	p binomial CDR-IMGT	p scarcity FR-IMGT	p scarcity CDR-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 Infect. 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., 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 (Veronique.Giudicelli@igh.cnrs.fr)

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



© Copyright 1989-2017 IMGT®, the international ImMunoGeneTics information system®