Strength of IMGT® standards in NGS repertoire analysis of IG and TR with IMGT/HighV-QUEST

Joumana Jabado-Michaloud, Géraldine Folch, Véronique Giudicelli,
Patrice Duroux, Eltaf Alamyar 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

ImMunoGeneTics Information system®

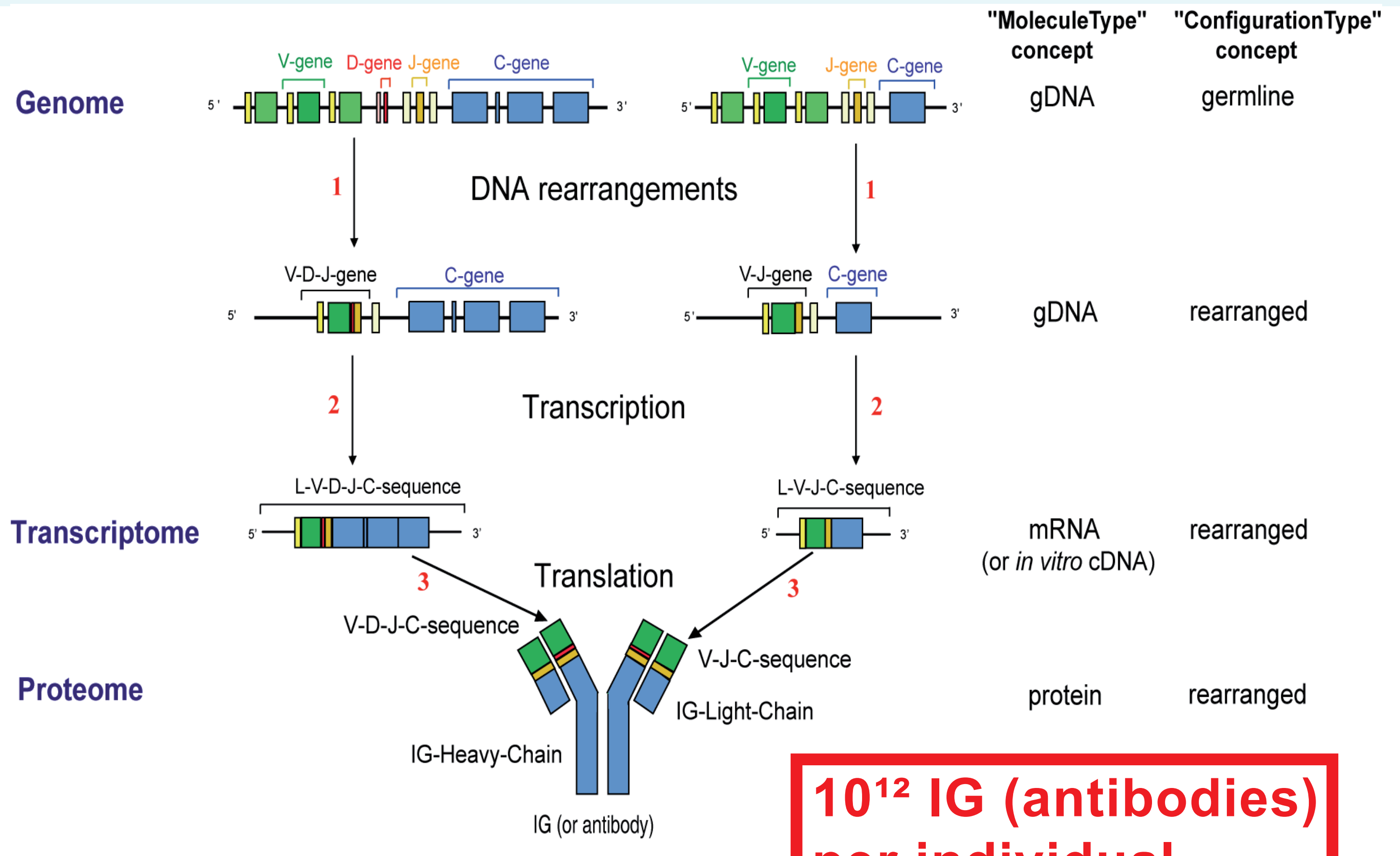
http://www.imgt.org

The analysis of expressed repertoires of antigen receptors - immunoglobulins (IG) or antibodies and T cell receptors (TR) - represents a huge challenge for the study of the adaptive immune response in normal and disease-related situations, such as viral infections. To answer that need IMGT®, the international ImMunoGeneTics information system® (http://www.imgt.org) has developed IMGT/HighV-QUEST [1,2] for the analysis of large repertoires of IG and TR sequences from NGS, which analyses up to 150,000 sequences per run, and provides statistical analysis for up to 450,000 sequences. IMGT/HighV-QUEST identifies the V, D, J genes and alleles by alignment with the germline IG and TR gene and allele sequences of the IMGT reference directory, which is constructed with data resulting from IMGT expert annotation. IMGT/HighV-QUEST integrates IMGT/JunctionAnalysis for a detailed analysis of the V-J and V-D-J junctions, and IMGT/Automat for a full V-J and V-D-J annotation. This analysis is based on IMGT-ONTOLOGY [3], the first ONTOLOGY in immunogenetics and immunoinformatics. IMGT-ONTOLOGY includes concepts of identification (IMGT standardized keywords), description (IMGT standardized labels), classification (IMGT standardized nomenclature: IMGT gene and allele names approved by HGNC and used by NCBI Gene) and numerotation (IMGT unique numbering and IMGT Colliers de Perles: widely used for antibody engineering and humanization). IMGT® standards are the basis of IMGT® biocuration. Based on them, IMGT/HighV-QUEST analyses NGS sequences of the expressed repertoires of antigen receptors with the same degree of accuracy and detailed annotation (539 columns) as IMGT/V-QUEST online. Since October 2010, more than 311 millions of sequences from 496 users (21/03/13) have been analysed.

[1] Alamyar E et al. Mol Biol 882:569-604, 2012. *[2] Alamyar E et al. Immunome Res 8(1):26, 2012.* *[3] Giudicelli V and Lefranc M-P, Front Genet, 3:79, 2012.*

Biological Context

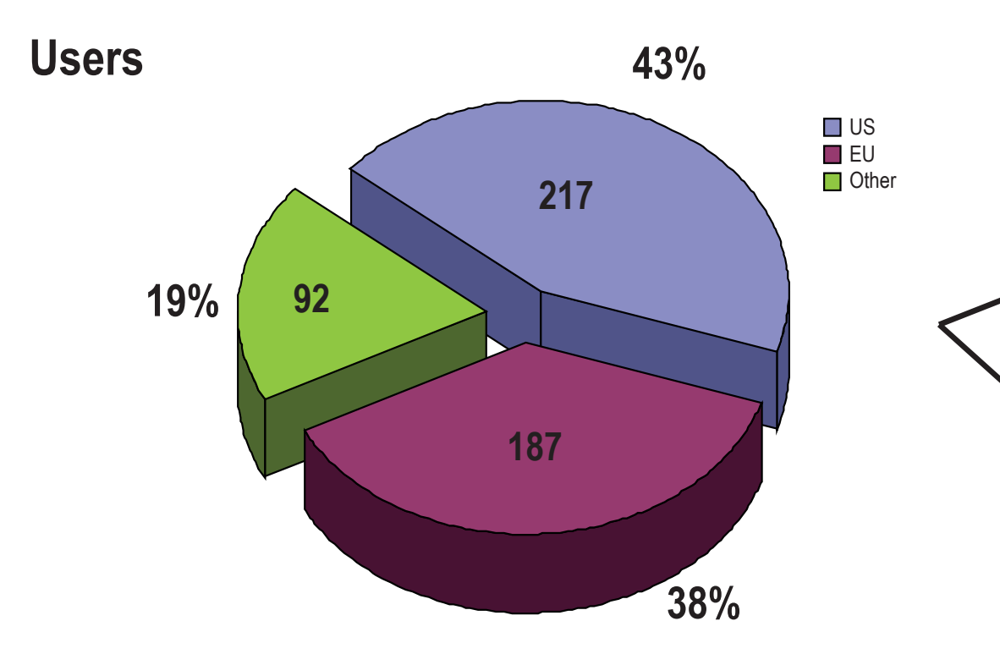
The adaptive immune response is characterized by an extreme diversity of the specific antigen receptors that comprise the immunoglobulins (IG) or antibodies and the T cell receptors (TR) (10^{12} different IG and 10^{12} different TR per individual, in humans). The complex molecular mechanisms (DNA rearrangements,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

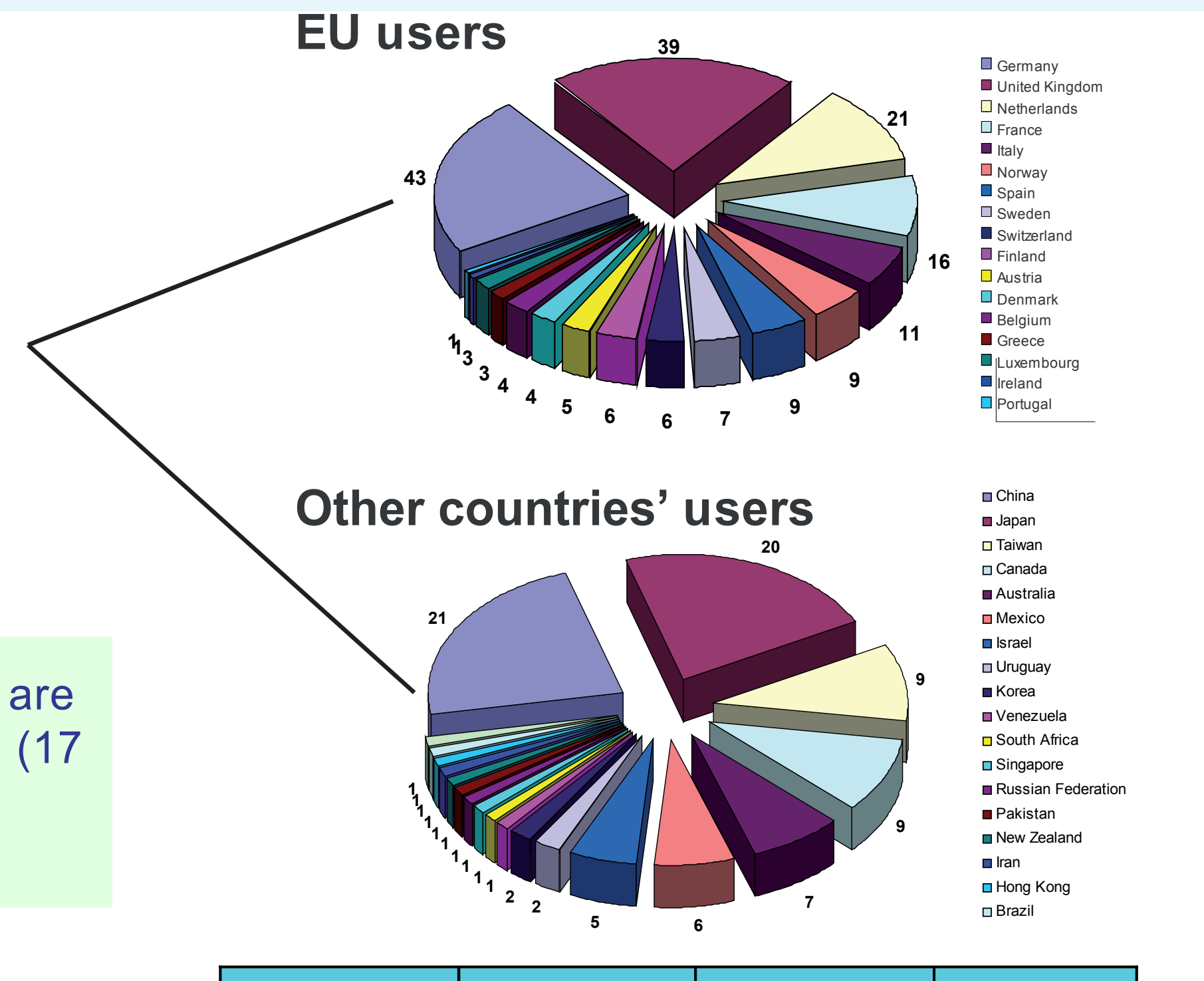


10^{12} IG (antibodies) per individual

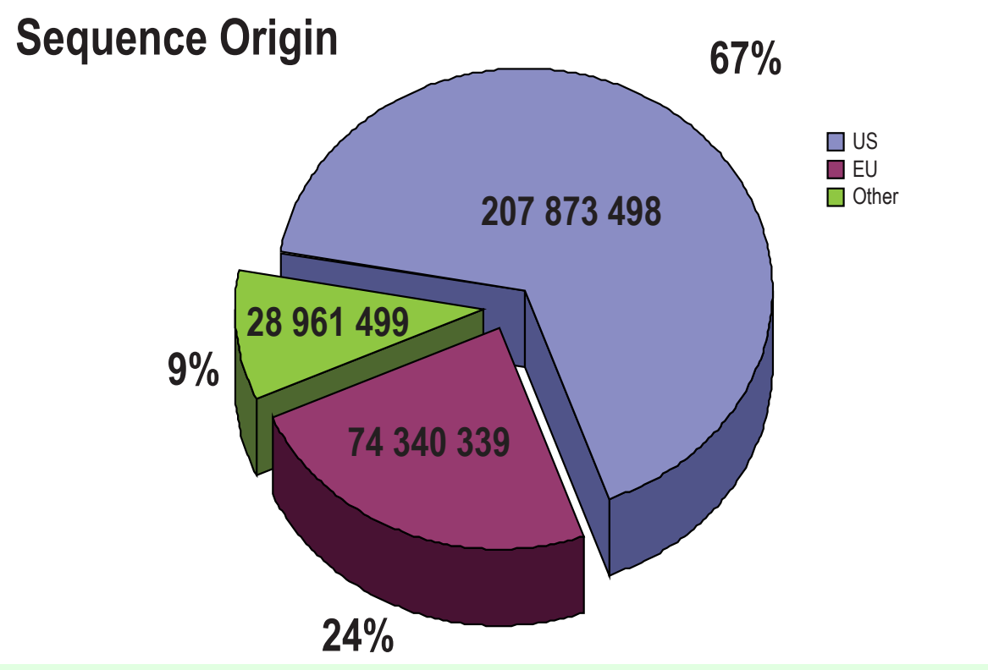
Users and Analyses

Since the availability of IMGT/HighV-QUEST in October 2010, more than 311 millions of sequences (from external users) have been submitted. They required more than 133,000 hours of computational resources. About 7.4 terabytes of results were generated.





217 of IMGT/HighV-QUEST users are from USA (43%), 187 are from EU (17 countries) (38%), and 92 other countries (19%).



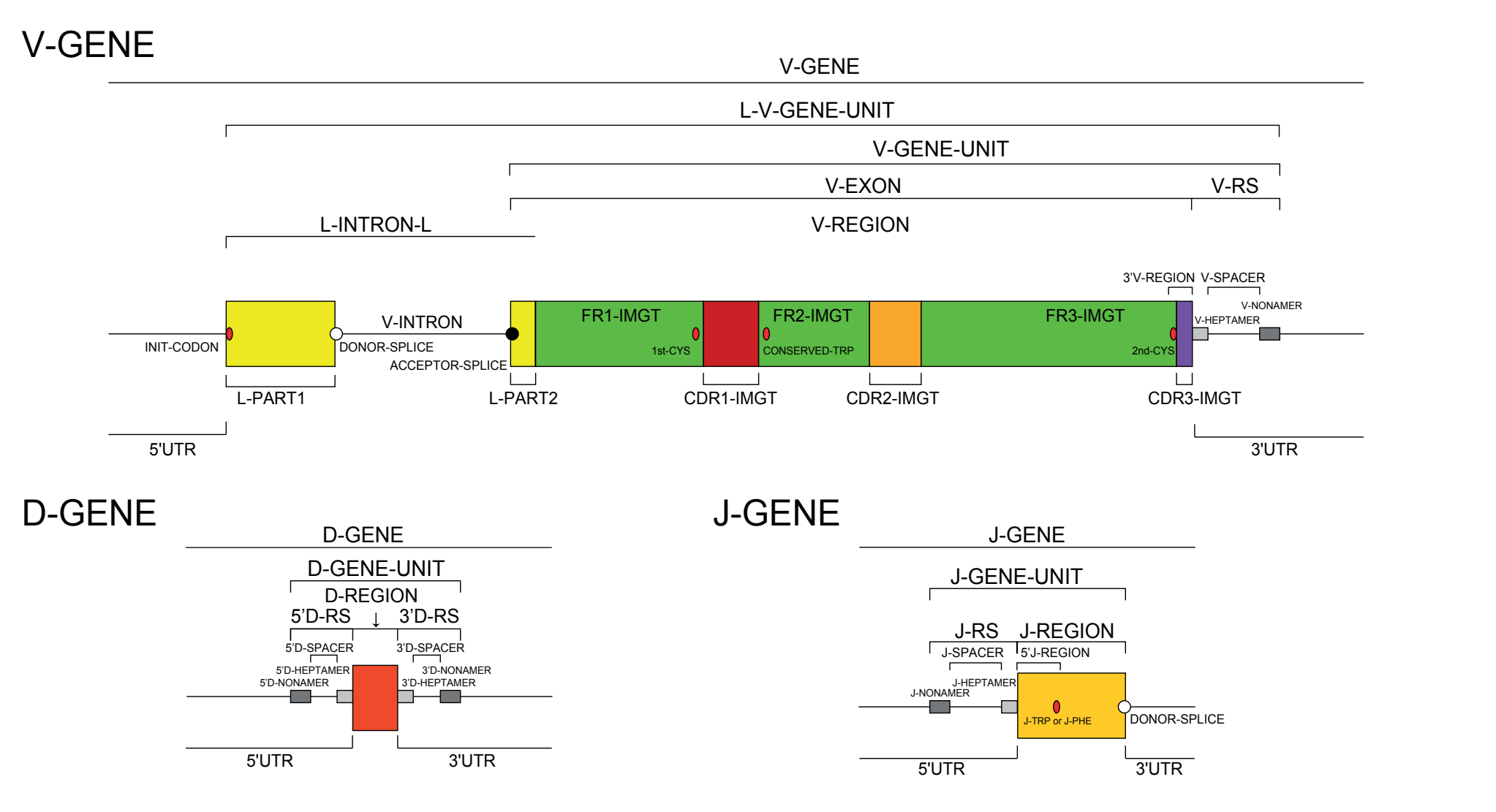
Users from USA submitted 67% of the sequences, users from EU submitted 24%, while the remaining sequences were submitted by users from other countries.

Statistics in 2013 show an increasing number of IMGT/HighV-QUEST users and a growing analysis demand compared with 2012 (150% increase in the number of submitted sequences and 61 new users were registered and activated during 2013 first quarter)

Country	Nb of submitted sequences	Computational resources (hours)	Size of generated files (Gbytes
United States	207 873 498,00	89 498,27	4 956,11
Germany	25 767 134,00	11 093,83	614,34
France	6 227 824,00	2 681,34	148,48
Canada	9 947 639,00	4 282,88	237,17
United Kingdom	17 868 752,00	7 693,25	426,03
Belgium	4 184 507,00	1 801,61	99,77
Ireland	2 367 729,00	1 019,41	56,45
Portugal	1 650 000,00	710,39	39,34
Mexico	2 069 967,00	891,21	49,35
Israel	1 626 694,00	700,36	38,78
China	6 460 347,00	2 781,45	154,03
Austria	1 638 553,00	705,47	39,07
Netherlands	4 439 821,00	1 911,53	105,85
Switzerland	2 802 384,00	1 206,54	66,81
Luxembourg	334 673,00	144,09	7,98
Japan	2 964 961,00	1 276,54	70,69
Spain	736 582,00	317,13	17,56
Hong Kong	1 344 548,00	578,88	32,06
Italy	200 115,00	86,16	4,77
Norway	5 661 440,00	2 437,49	134,98
Greece	46 240,00	19,91	1,10
Australia	181 439,00	78,12	4,33
Taiwan	4 118 234,00	1 773,07	98,19
Sweden	1 552,00	0,67	0,04
Argentina	9,00	0,00	0,00
Korea	30,00	0,01	0,00
Denmark	409 616,00	176,36	9,77
Venezuela	47 556,00	20,47	1,13
Finland	3 417,00	1,47	0,08
Russian Federation	200 000,00	86,11	4,77
Uruguay	75,00	0,03	0,00
Total	**311 175 336,00**	**133 974,05**	**7 419,02**

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 a standardized nomenclature for any vertebrate species.

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

IMGT/HighV-QUEST based on IMGT® standard

1. Selection of results for statistical analysis

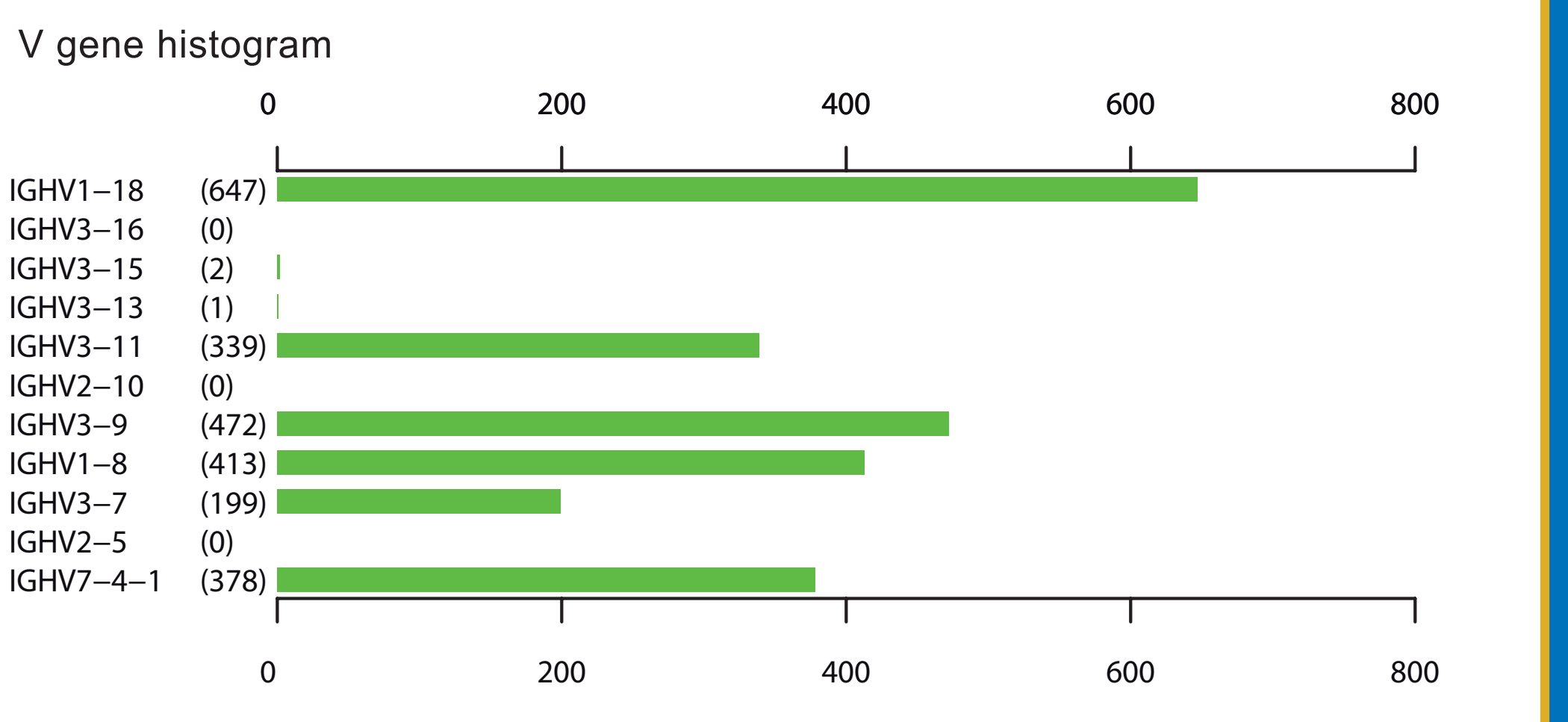
Statistical analyses are performed on results selected as '1 copy' (redundancies are recorded but not processed), and with quality criteria (identification of a single gene/allele, known functionality, absence of IMGT/V-QUEST warnings regarding the CDR1-IMGT and CDR2-IMGT lengths and the percentage of identity).

2. Tables and histograms for each gene (V, D and J)

For each gene, number of sequences, average sequence length, average V-, D-, J-REGION length, and number of sequences with an identity percentage of 100% by comparison with the germline, are provided.

V gene and allele table

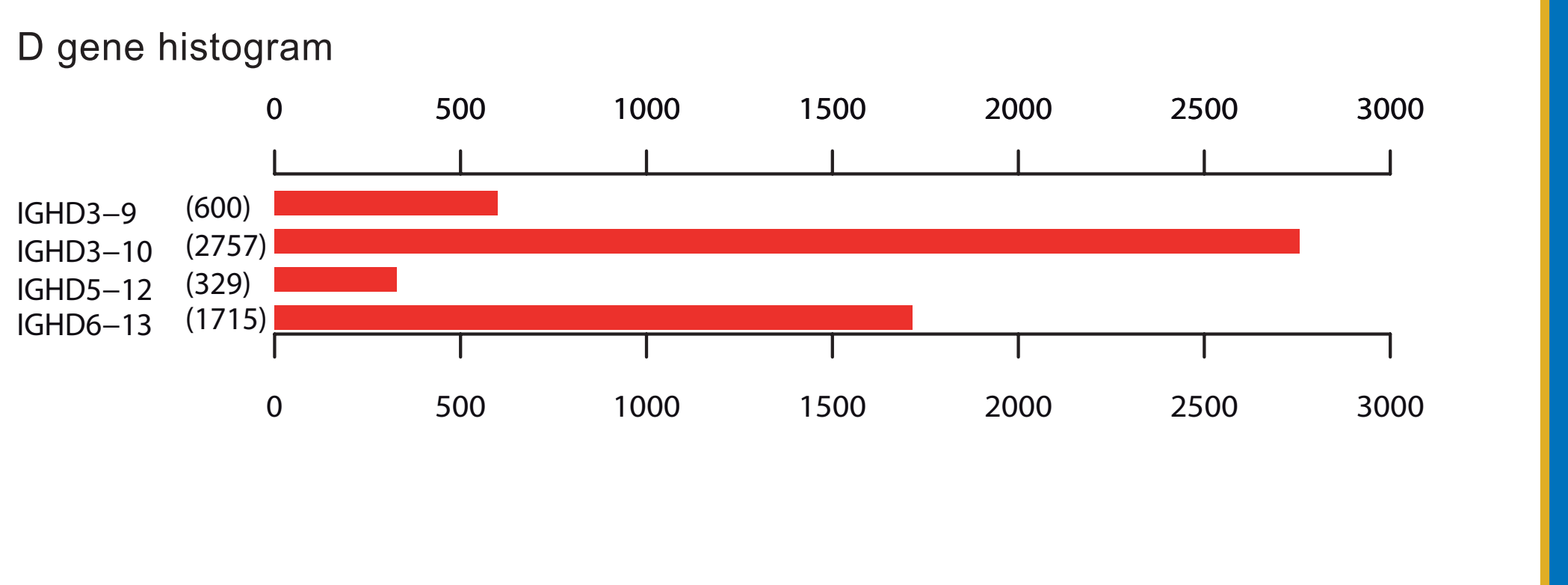
#	IMGT gene and allele	Total	Average sequence length	Average V-REGION length	id=100% nb (%)
1	IGHV1-18	647	243	166	455 (70.32%)
	IGHV1-18*01	647	243	166	455 (70.32%)
9	IGHV3-11	339	242	166	253 (74.63%)
	IGHV3-11*01	339	242	166	253 (74.63%)
10	IGHV3-13	1	223	158	1 (100.0%)
	IGHV3-13*01	1	223	158	1 (100.0%)
11	IGHV3-15	2	266	173	1 (50.0%)
	IGHV3-15*04	1	283	173	0 (0.0%)
	IGHV3-15*07	1	248	173	1 (100.0%)



D gene and allele table

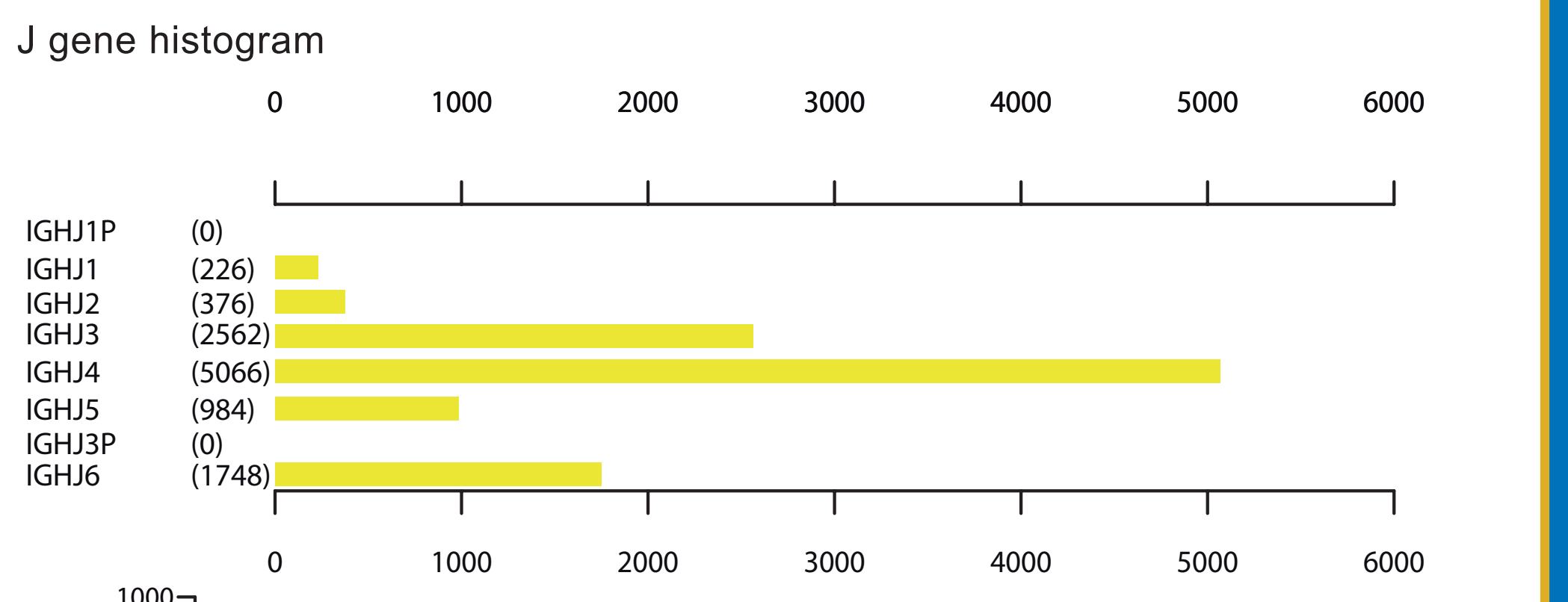
#	IMGT gene and allele	Total	Average sequence length	Average D-REGION length
10	IGHD3-10	2757	243	17
	IGHD3-10*01	2693	244	15
	IGHD3-10*02	64	242	19
14	IGHD3-9	600	246	19
	IGHD3-9*01	600	246	19
18	IGHD5-12	329	238	14
	IGHD5-12*01	329	238	14
21	IGHD6-13	1715	239	15
	IGHD6-13*01	1715	239	15

Colored lines illustrate results per gene and white lines under each gene illustrate the results per allele, individually. In the histograms, genes are ordered according to their positions from 5' to 3' in the locus.



J gene and allele table

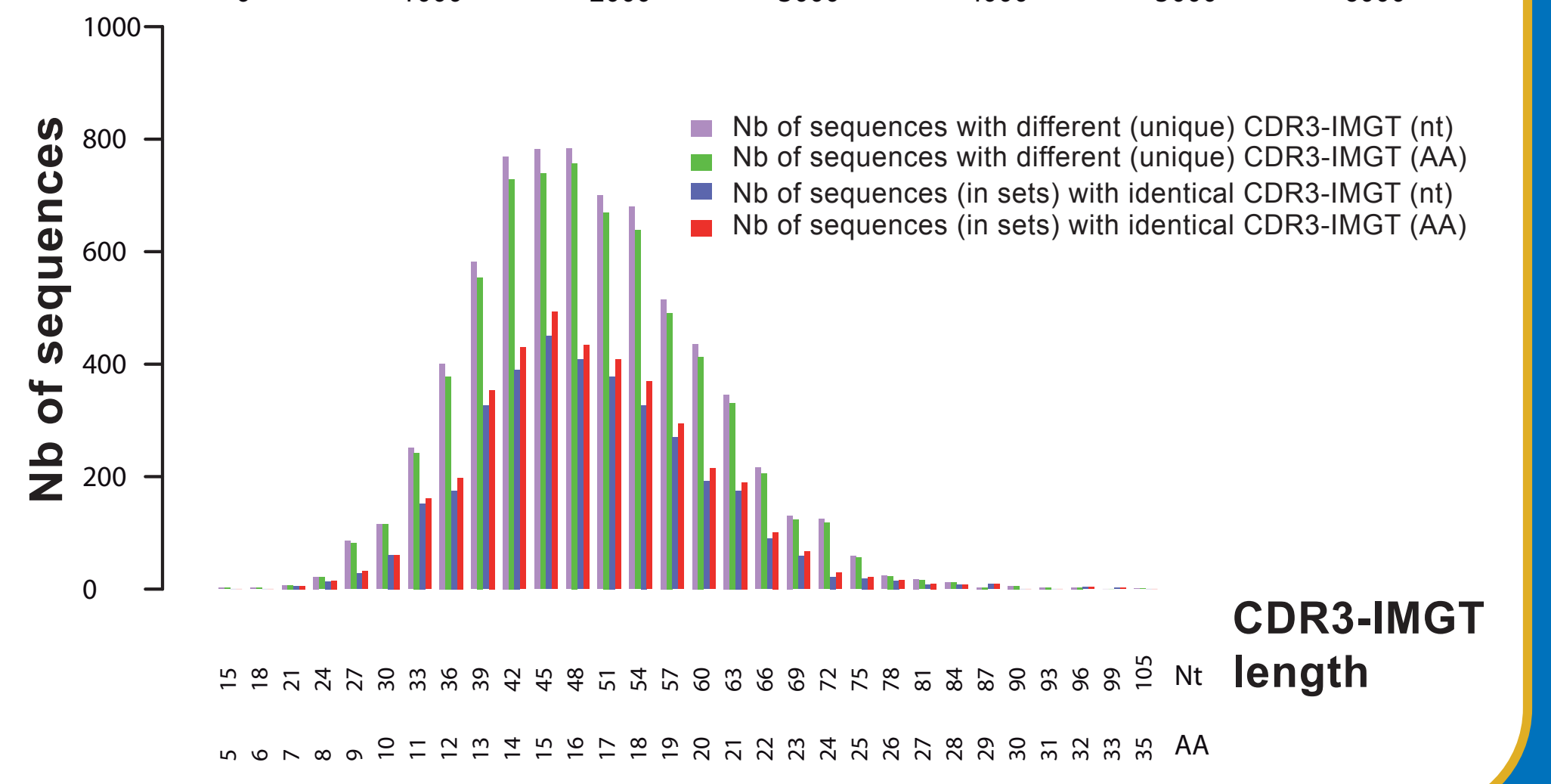
#	IMGT gene and allele	Total	Average sequence length	Average J-REGION length	id=100% nb (%)
2	IGHJ2	414	243	50	0 (0.0%)
	IGHJ2*01	414	243	50	0 (0.0%)
3	IGHJ3	2685	244	44	0 (0.0%)
	IGHJ3*01	36	245	41	0 (0.0%)
	IGHJ3*02	2649	243	48	0 (0.0%)
4	IGHJ4	5795	240	41	754 (13.01%)
	IGHJ4*01	5	239	46	3 (60.0%)
	IGHJ4*02	5708	238	33	751 (13.16%)
	IGHJ4*03	82	242	43	0 (0.0%)



3. CDR3-IMGT length analysis

Statistics provide the histogram of different and identical CDR3-IMGT sequences for each CDR3-IMGT length in nucleotides (nt) and amino acids (AA).
Results are shown as:

- Nb of sequences with different (unique) CDR3-IMGT (nt)
- Nb of sequences with different (unique) CDR3-IMGT (AA)
- Nb of sequences (in sets) with identical CDR3-IMGT (nt)
- Nb of sequences (in sets) with identical CDR3-IMGT (AA)



Acknowledgments: this work was granted access to the HPC resources of CINES under the allocation 036029-(2010-2013) 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)
Webmaster: Chantal Ginestoux (Chantal.Ginestoux@igh.cnrs.fr)