IG and TR single chain fragment variable (scFv) sequence analysis: a new advanced functionality of IMGT/V-QUEST and IMGT/HighV-QUEST

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

Giudicelli et al. BMC Immunology (2017) 18:35
DOI 10.1186/s12865-017-0218-8

BMC Immunology

METHODOLOGY ARTICLE Open Access

IG and TR single chain fragment variable (scFv) sequence analysis: a new advanced functionality of IMGT/V-QUEST and IMGT/HighV-QUEST

Véronique Giudicelli*, Patrice Duroux*, Sofia Kossida* and Marie-Paule Lefranc*

Abstract

Background: IMGT®, the international ImMunoGeneTics information system® (http://www.imgt.org), was created in 1989 in Montpellier, France (CNRS and Montpellier University) to manage the huge and complex diversity of the antigen receptors, and is at the origin of immunoinformatics, a science at the interface between immunogenetics and bioinformatics. Immunoglobulins (IG) or antibodies and T cell receptors (TR) are managed and described in the IMGT® databases and tools at the level of receptor, chain and domain. The analysis of the IG and TR variable (V) domain rearranged nucleotide sequences is performed by IMGT/V-QUEST (online since 1997, 50 sequences per batch) and, for next generation sequencing (NGS), by IMGT/HighV-QUEST, the high throughput version of IMGT/V-QUEST (portal begun in 2010, 500,000 sequences per batch). In vitro combinatorial libraries of engineered antibody single chain Fragment variable (scFv) which mimic the in vivo natural diversity of the immune adaptive responses are extensively screened for the discovery of novel antigen binding specificities. However the analysis of NGS full length scFv (~850 bp) represents a challenge as they contain two V domains connected by a linker and there is no tool for the analysis of two V domains in a single chain.

Methods: The functionality "Analyis of single chain Fragment variable (scFv)" has been implemented in IMGT/V-QUEST and, for NGS, in IMGT/HighV-QUEST for the analysis of the two V domains of IG and TR scFv. It proceeds in five steps: search for a first closest V-REGION, full characterization of the first V-(D)-J-REGION, then search for a second V-REGION and full characterization of the second V-(D)-J-REGION, and finally linker delimitation.

Results: For each sequence or NGS read, positions of the 5'V-DOMAIN, linker and 3'V-DOMAIN in the scFv are provided in the 'V-orientated' sense. Each V-DOMAIN is fully characterized (gene identification, sequence description, junction analysis, characterization of mutations and amino changes). The functionality is generic and can analyse any IG or TR single chain nucleotide sequence containing two V domains, provided that the corresponding species IMGT reference directory is available.

Conclusion: The "Analysis of single chain Fragment variable (scFv)" implemented in IMGT/V-QUEST and, for NGS, in IMGT/HighV-QUEST provides the identification and full characterization of the two V domains of full-length scFv (~850 bp) nucleotide sequences from combinatorial libraries. The analysis can also be performed on concatenated paired chains of expressed antigen receptor IG or TR repertoires.

Keywords: IMGT, immunoglobulin, IG, T cell receptor, TR, single chain fragment variable, scFv, IMGT-ONTOLOGY, V-DOMAIN, adaptive immune repertoire

* Correspondence: Veronique.Giudicelli@igh.cnrs.fr; Patrice.Duroux@igh.cnrs.fr; Sofia.Kossida@igh.cnrs.fr; Marie-Paule.Lefranc@igh.cnrs.fr
IMGT®, the international ImMunoGeneTics information system®, Laboratoire d'ImmunoGénétique Moléculaire LIGM, Institut de Génétique Humaine IGH, UMR 9002, CNRS, Montpellier University, Montpellier, France

Background

The efficiency of the adaptive immune responses of humans and other jawed vertebrates (or *gnathostomata*) results from the remarkable immune specificity and memory, which are the properties of B and T cells owing to an extreme diversity of their antigen receptors [1]. The specific antigen receptors comprise the immunoglobulins (IG) or antibodies [2], expressed on the surface of the B cells and secreted by the plasmocytes, and the T cell receptors (TR) [3] expressed on the surface of T cells. The potential antigen receptor repertoire of each individual is estimated to comprise about 2 x 10^{12} different IG and TR specificities, and the limiting factor is the number of B and T cells that an organism is genetically programmed to produce [1].

IMGT®, the international ImMunoGeneTics information system® [4, 5], was created in 1989 by Marie-Paule Lefranc at Montpellier, France (CNRS and Montpellier University) to manage the huge and complex diversity of these antigen receptors, and is at the origin of immunoinformatics, a science at the interface between immunogenetics and bioinformatics [1]. IMGT® has developed IMGT-ONTOLOGY [6] to manage, reuse and share knowledge in immunoinformatics [1]. IMGT-ONTOLOGY comprises seven axioms which generated the concepts of identification, description, classification, numerotation, localization, orientation and obtention and the IMGT Scientific chart rules (keywords, labels, numbering): IDENTIFICATION (IMGT® standardized keywords) [7], DESCRIPTION (IMGT® standardized labels (in capital letters, no plural)) [8], CLASSIFICATION (IMGT® standardized gene and allele nomenclature) [9], NUMEROTATION (IMGT unique numbering [10–12] and its graphical 2D representation or IMGT Collier de Perles [13]) [14–16], LOCALIZATION, ORIENTATION and OBTENTION [17–19].

IMGT® is specialized in the IG or antibodies, TR, major histocompatibility (MH) of human and other jawed vertebrate species, and in the immunoglobulin superfamily (IgSF), MH superfamily (MhSF) and related proteins of the immune system (RPI) of vertebrates and invertebrates. IMGT® comprises 7 databases, seventeen online tools and more than 20,000 pages of Web resources, available at the IMGT® Home page [4, 5]. The databases provide IMGT biocurated and standardized information on genes (IMGT/GENE-DB [20], sequences (IMGT/LIGM-DB [21], IMGT/PRIMER-DB), two-dimensional (2D) and three-dimensional (3D) structures (IMGT/2Dstructure-DB and IMGT/3Dstructure-DB [22, 23]), therapeutic monoclonal antibodies, fusion proteins for immune applications (FPIA), composite proteins for clinical applications (CPCA) and related proteins of the immune system (RPI) (IMGT/mAb-DB [4]). The online tools are available for the analysis of nucleotide sequences (IMGT/V-QUEST [24–26], IMGT/JunctionAnalysis [27, 28], IMGT/Automat [29, 30]), next generation sequencing (NGS) nucleotide sequences (IMGT/HighV-QUEST [31–35]), amino acid sequences (IMGT/DomainGapAlign [36], IMGT/Collier-de-Perles [37]), genes (IMGT/GeneInfo [38], IMGT/LIGMotif [39], IMGT/GeneFrequency) and 2D and 3D structures (IMGT/StructuralQuery). The standalone tool, IMGT/StatClonotype [40, 41], allows statistical comparison of clonotype diversity and expression from IMGT/HighV-QUEST NGS results.

IG and TR are managed and described in the IMGT® databases and tools at the level of receptor, chain and domain [1]. A complete IgG1 is made of 12 domains belonging to two identical heavy (H) chains (4 domains each) and two identical light (L) chains (2 domains each) [1]. The N-terminal domain of each IG H and L chain is a variable domain (VH and VL, respectively) which results from the rearrangement at the DNA level of three genes for the VH (variable (V), diversity (D) and joining (J)) and of two genes for the VL (V and J). As a result a VH is encoded by a V-D-J-REGION whereas a VL is encoded by a V-J-REGION (Table 1) [2]. Similarly the N-terminal domain of each chain of a T cell receptor (TR) is a variable domain encoded by a V-D-J-REGION or a V-J-REGION (Table 1) [3].

The analysis of the IG and TR V domain rearranged nucleotide sequences is performed by IMGT/V-QUEST (online since 1997, 50 sequences per batch) and, for NGS, by IMGT/HighV-QUEST, the high throughput version of IMGT/V-QUEST (online since October 2010), maximum of 500,000 sequences per batch, set comparison of 1 million results). IMGT/V-QUEST and HighV-QUEST use the same algorithm and the same IMGT reference directories [4].

So far, the analysis has been performed on each V domain individually. The Sanger sequencing of single chain Fragment variable (scFv) was done on a case by case basis using IMGT/V-QUEST. Indeed scFv are single chains of approximate molecular weight of 26,000 Da, encoded by about 800–900 nucleotides with two V domains connected by a linker of about 45-60 nucleotides (Fig. 1), and the user could easily identify the linker by its sequence and length (for example (GSSS)3) and remove it or split the sequence in two parts preceding IMGT/V-QUEST analysis. This manual approach is cumbersome and not applicable to high-throughput sequencing. The NGS sequencing of scFv from combinatorial libraries has been limited up to now by the short length of reads, however with the availability of longer NGS reads (1000 bp and more) and the use of circular consensus sequencing (CCS) [42] as introduced by Pacific Biosciences, high quality sequencing of full-length scFv or of single cell

Table 1 V-DOMAIN types analyzed by IMGT/V-QUEST

Receptor type	V-DOMAIN description			Locus name	Chain type (transcript or protein)
	Structure labels (IMGT/3Dstructure-DB)		Sequence labels (IMGT/LIGM-DB)		
IG	VH		V-D-J-REGION	IGH	IG-Heavy
	VL[a]	V-KAPPA	V-J-REGION	IGK	IG-Light-Kappa
		V-LAMBDA	V-J-REGION	IGL	IG-Light-Lambda
		V-IOTA[b]	V-J-REGION	IGI	IG-Light-Iota
TR	V-ALPHA		V-J-REGION	TRA	TR-Alpha
	V-BETA		V-D-J-REGION	TRB	TR-Beta
	V-GAMMA		V-J-REGION	TRG	TR-Gamma
	V-DELTA		V-D-J-REGION	TRD	TR-Delta

[a]V-RHO of the IG-Light-Rho (frog) and V-SIGMA of the Ig-Light-Sigma (Chondrichthyes and frog) are not shown but will be analyzed once the genomic germline V and J genes are sequenced and available in the IMGT/V-QUEST reference directory (Correspondence between chain types and C genes: IG and TR (all vertebrate species) [60])
[b]V-IOTA is the VL of the IG-Light-Iota chain type (transcript or protein) of the Chondrichthyes and Actinopterygii (which include the Teleostei)

concatenated antigen receptor V-domain or chain pairs are expected.

In this paper, we describe a new advanced IMGT/V-QUEST functionality "Analysis of single chain Fragment variable (scFv)" for the identification and characterization of the two variable domains of scFv, generic for IG and TR, and implemented, for NGS, in IMGT/HighV-QUEST.

Methods

The algorithm proceeds in five steps (Fig. 2): search for a first closest V-REGION, full characterization of the first V-(D)-J-REGION, then search for a second V-REGION and full characterization of the second V-(D)-J-REGION, and finally linker delimitation.

Search for a first closest V-REGION

For a selected species and receptor type (IG or TR), the IMGT/V-QUEST tool first searches the submitted sequence for the closest V-REGION by comparison with the IMGT reference directory of the V groups of the selected receptor type (for the IG: IGHV, IGKV and IGLV; for the TR: TRAV, TRBV, TRGV and TRDV) [26]. The IMGT reference directories [4] are reference sequences of IG and TR IMGT genes and alleles (functional (F), open reading frames (ORF) and in-frame pseudogenes (P)), from IMGT/GENE-DB [20]. By default, the search is done on 'F + ORF + in-frame P'. The identification of the closest V-REGION determines the assignment of the genes of the V-(D)-J-REGION to a locus (IGH, IGK or IGL for IG, or TRA, TRB, TRG or TRD for TR, respectively).

The first closest V-REGION identified is the one with the highest score which would have been detected in a classical IMGT/V-QUEST analysis (i.e., without the option "Analysis of single chain Fragment variable (scFv)"). There is no search priority for a given V group or for a

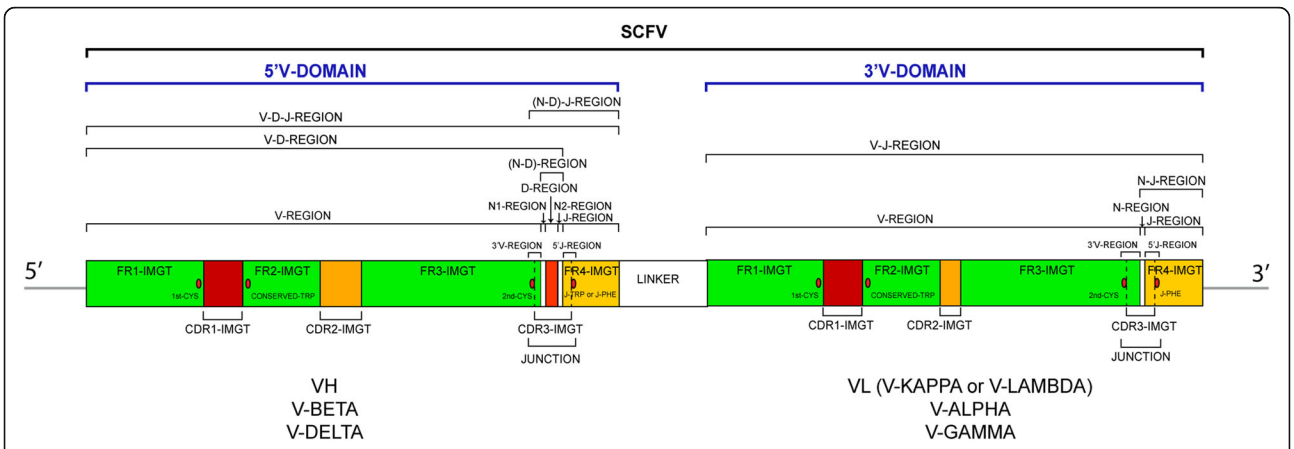

Fig. 1 Representation of a scFv with the IMGT Labels for the two V-DOMAIN. The relative position 5′ and 3′ of the two V-DOMAIN in the scFv, in this figure, is arbitrary. At least 23 and 19 labels are necessary to describe a V-D-J-REGION and a V-J-REGION, respectively in IMGT/V-QUEST [59]. The conserved 1st-CYS 23, CONSERVED-TRP 41, 2nd-CYS 104 and J-TRP or J-PHE 118 based on the IMGT unique numbering are shown in both V-DOMAIN [1, 10, 14, 16]

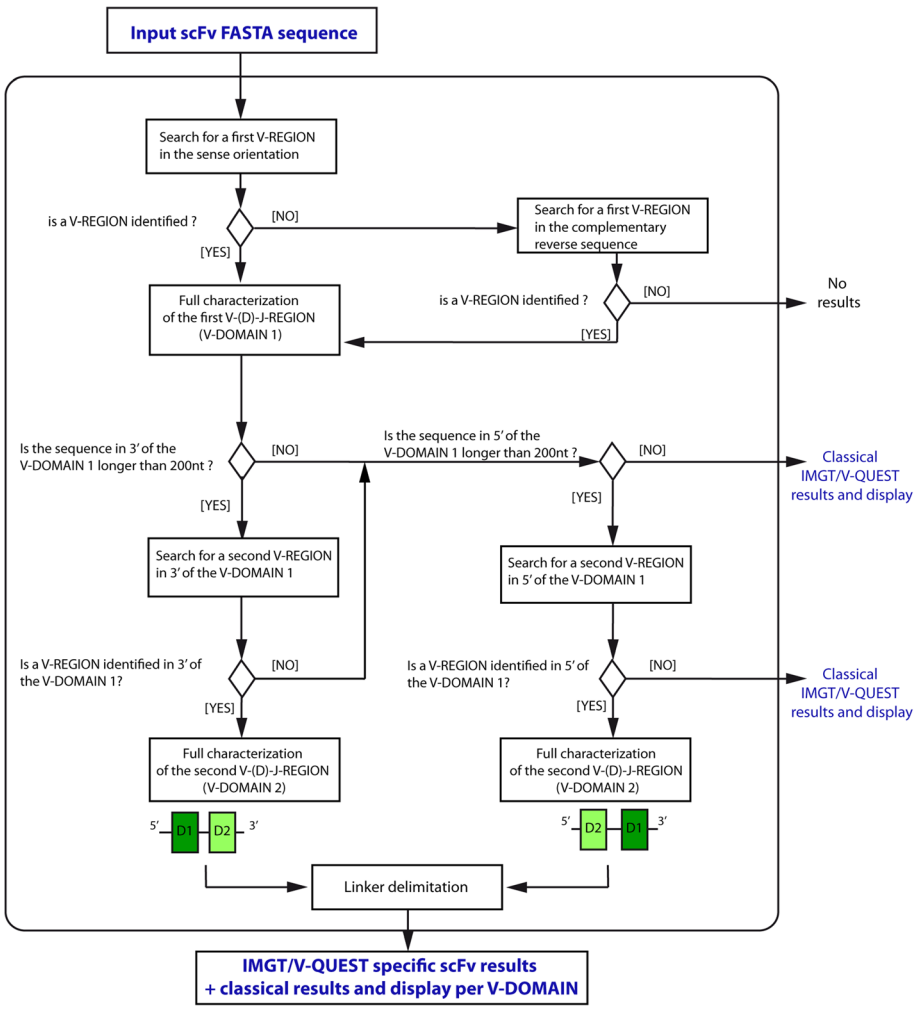

Fig. 2 Main steps of IMGT/V-QUEST algorithm for the analysis of scFv sequences. 'D1': indicates the first identified and characterized V-(D)-J-REGION (V-DOMAIN 1). 'D2': indicates the second identified and characterized V-(D)-J-REGION (V-DOMAIN 2) which can be found in 3′ or in 5′ of 'D1' in the 'V-orientated' sequence

respective order position (5′ or 3′) in the submitted sequence.

If no V-REGION is identified, IMGT/V-QUEST complementary reverses the submitted (input) sequence automatically, and the search is performed again. If a closest V-REGION is identified, this defines the complementary reverse sequence as being in the 'sense' orientation for the V-REGION (Fig. 2).

The following steps of the algorithm are performed on scFv sequences in which the V-REGION has a 'sense' orientation, and therefore are designated as 'V-orientated scFv' (either from the direct input scFv sequence or as a result of a complementary reverse step).

Full characterization of the first V-(D)-J-REGION

The full characterization of the first identified V-(D)-J-REGION ('V-DOMAIN 1' or 'D1' in Fig. 2) is performed through a set of methods described previously [24–26]. In summary, IMGT/V-QUEST

i. identifies the names of the closest germline V-GENE and allele and J-GENE and allele, with score and percent (%) of identity [24–26] by alignments with the IMGT reference directory [4].
ii. adds gaps according to the IMGT unique numbering [10] and determines the lengths of the four framework regions (FR) FR1-IMGT to FR4-IMGT, and those of the three complementarity determining regions (CDR), CDR1-IMGT to CDR3-IMGT [1].
iii. delimits the V-(D)-J-REGION, i.e., the V-DOMAIN (V-D-J-REGION for the IGH, TRB and TRD loci or V-J-REGION for the IGK, IGL, TRA and TRG loci),
iv. provides a detailed analysis of the V-(D)-J junction and the identification of the D genes and alleles for

IGH, TRB and TRD performed by the integrated IMGT/JunctionAnalysis tool [27, 28],
v. provides an extensive analysis of the nucleotide (nt) mutations and amino acid (AA) changes, resulting for the IG from somatic hypermutations, by comparison with the closest V-REGION,
vi. localizes the mutation hotspots in the closest germline V gene and allele,
vii. and finally, annotates the V-(D)-J-REGION identified with IMGT labels using IMGT/Automat [29, 30].

Links to the IMGT/Collier-de-Perles tool graphical representation [34] are only incorporated for IMGT/V-QUEST results online.

Search for a second V-REGION and full characterization of the second V-(D)-J-REGION

Following the complete characterization of a first V-(D)-J-REGION ('V-DOMAIN 1' or 'D1' in Fig. 2), a second V-REGION is searched by comparison with the V groups of the same receptor type (IG or TR) and species as previously selected, from the IMGT reference directory. The search is performed on a 'V-orientated' sequence, which is either the input sequence if the 'D1' has a 'sense' orientation or the complementary reverse sequence if the 'D1' has an "antisense" orientation.

The search is performed first between the 3′end of 'D1' and the 3′end of the V-orientated sequence, provided that this part has at least a length of 200 nt (Fig. 2). If a V-REGION is identified, the full characterization of the second V-(D)-J-REGION ('V-DOMAIN 2′ or 'D2' in Fig. 2) is performed by IMGT/V-QUEST similarly to that of the first V-(D)-J-REGION, as described above [24–30] (Fig. 2).

If IMGT/V-QUEST does not find a second V-REGION in 3′ of the V-orientated sequence (either sequence shorter than 200 nt or absence of results), a similar search is performed between the 5′ end of the V-orientated sequence and the 5′ end of 'D1', provided that this part of the sequence has at least a length of 200 nt. If a V-REGION is identified, the full characterization of the second V-(D)-J-REGION ('V-DOMAIN 2′ or 'D2' in Fig. 2) is performed by IMGT/V-QUEST similarly to that of the first V-(D)-J-REGION, as described above [24–30] (Fig. 2).

Linker delimitation

When two V-(D)-J-REGION ('V-DOMAIN 1' and 'V-DOMAIN 2') are characterized, the sequence between them is delimited and defined as 'linker' (Fig. 2). The linker length and positions in the sequence are delimited by the 3′ end of 'V-DOMAIN 1′ and the 5′ end of 'V-DOMAIN 2'. There is no further characterization of the linker sequence.

Results

IMGT/V-QUEST user submission for "Analysis of single chain Fragment variable (scFv)"

The IMGT/V-QUEST novel functionality for "Analysis of single chain Fragment variable (scFv)" is freely available online for academics (since May 10, 2016). The scFv sequences are submitted in FASTA format (up to 50 sequences, or a maximum of 10 sequences with the option 'Search for insertions and deletions').

The "Analysis of single chain Fragment variable (scFv)" is selected as an option in "Advanced functionalities" at the bottom of the IMGT/V-QUEST Search page (Fig. 3).

The user can choose one of the three displays: A. Detailed view. B. Synthesis view. C. Excel file. Selecting the "Analysis of single chain Fragment variable (scFv)" adds automatically the file '12' in the display C. Excel file.

The results are displayed even if the V-(D)-J-REGION (V-DOMAIN) are partial and/or not fully characterized, the only requirement being that at least the V-REGION has been identified, which is the condition for IMGT/V-QUEST to give results.

IMGT/V-QUEST Detailed view results for scFv

The top of the page IMGT/V-QUEST Detailed view results for scFv (Fig. 4a) recalls the IMGT/V-QUEST program version, IMGT/V-QUEST reference directory release, and then the selected parameters: Species, Receptor type or locus (IG or TR), IMGT reference directory set (e.g., F + ORF + in-frame P), and options 'Search for insertions and deletions' (yes or no) and, for the current purpose, 'Analysis of scFv' (yes).

The Detailed results show the Number of analysed sequences (here, 3) and the Number of analysed V-DOMAIN (here, 6). The table "Identified scFv" comprises one line per identified scFv in the submitted sequence set. It includes the sequence ID (the one from the flat file header), the 5′V-DOMAIN ID, positions and length, the linker positions and length, the 3′V-DOMAIN ID, positions and length (Fig. 4a).

The 5′V-DOMAIN ID and 3′V-DOMAIN ID consists of the sequence ID, followed by an underscore and a capital letter for the locus as identified by IMGT/V-QUEST (H, K, L for IGH, IGK and IGL, and A, B, D and G for TRA, TRB, TRD and TRG, respectively), and preceded by a number which indicates the V domain analysis order in the submitted set. Clicking on the 5′V-DOMAIN ID or 3′V-DOMAIN ID link leads to the corresponding classical detailed view (Fig. 4b).

The complete V-orientated sequence of the scFv is shown in the result of each domain, with the corresponding analysed V domain being highlighted in green.

If the option 'Search for insertions and deletions' has been selected, insertions detected are reported in capital letters in the sequence of the corresponding V domain(s).

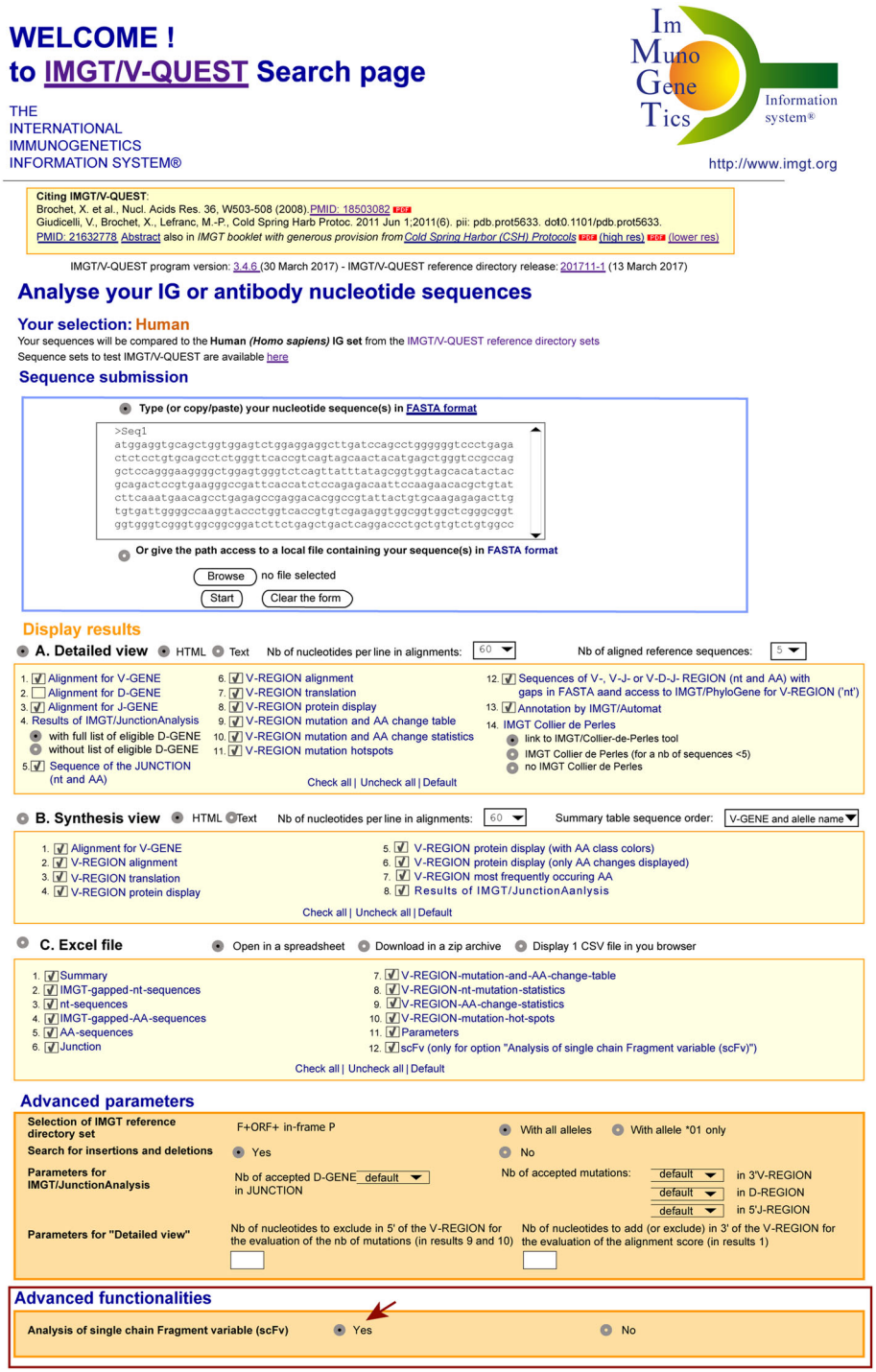

Fig. 3 IMGT/V-QUEST Search page. The analysis of scFv is an option of IMGT/V-QUEST available in the section Advanced functionalities at the bottom of the IMGT/V-QUEST Search page. This option is not selected by default

The IMGT/V-QUEST results per domain are given after filling the deletions and removing the insertions [25].

The Result summary table of each V domain (Fig. 4b) is followed by the 14 classical displays of Detailed view results (not shown) [32].

IMGT/V-QUEST Synthesis view results for scFv

The top of the page IMGT/V-QUEST Synthesis view results for scFv (Fig. 5) recalls, as for the Detailed view results above, the IMGT/V-QUEST program version, IMGT/V-QUEST reference directory release, and then

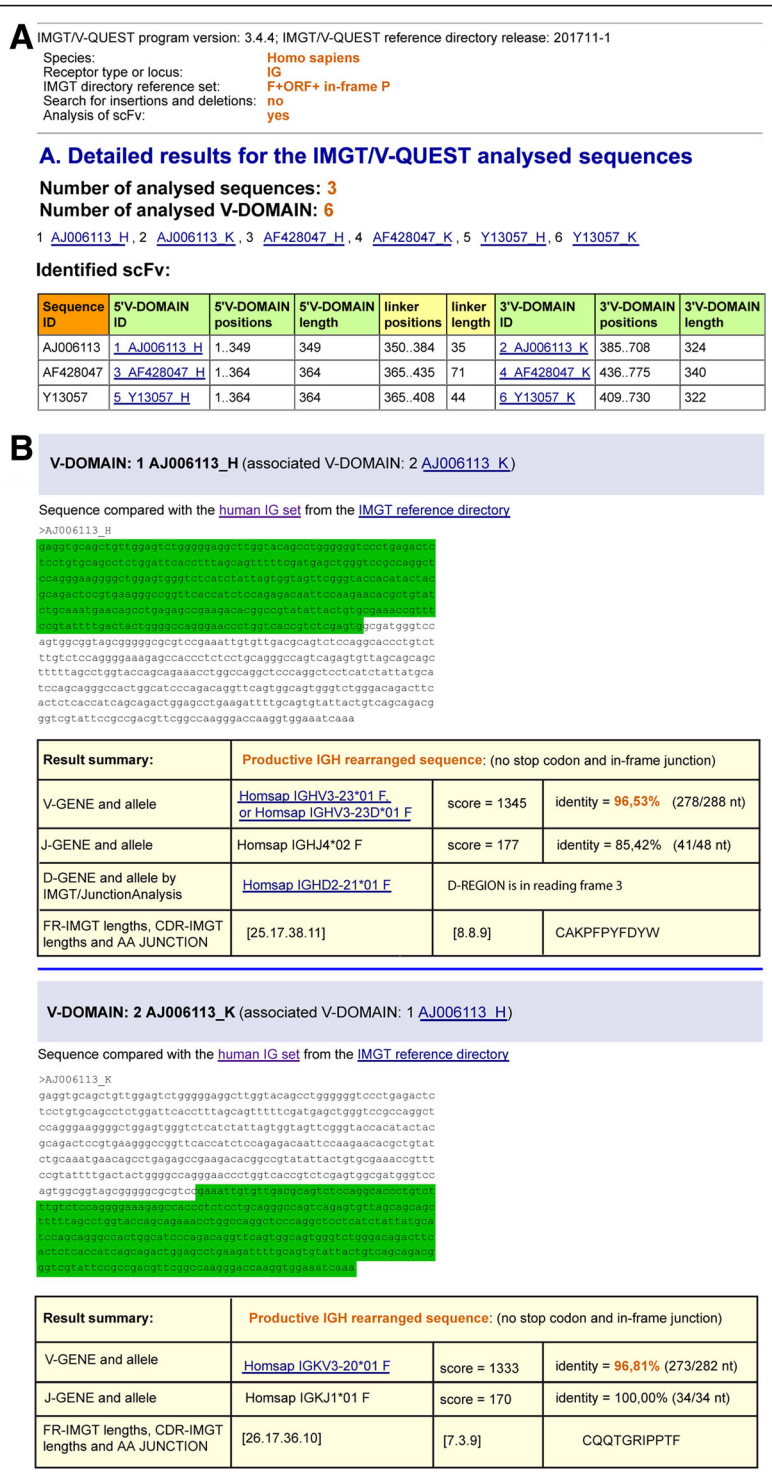

Fig. 4 IMGT/V-QUEST Detailed view results for IG scFv sequences. **a** The top of the Detailed view results recalls the parameters for the analysis and provides the number of analyzed sequences (here, 3) and the number of analyzed V-DOMAIN (here, 6). The "Identified scFv" table indicates, for each identified scFv in the submitted sequence set, the positions and length of the 5′V-DOMAIN, linker and 3′V-DOMAIN in the 'V-orientated' scFv. Clicking on the 5′V-DOMAIN ID or 3′V-DOMAIN ID leads to the corresponding detailed analysis. **b** Sequence and Result summary for the two V-(D)-J-REGION (V-DOMAIN) of a scFv are shown. The part of the scFv FASTA sequence colored in green corresponds to the analyzed V-DOMAIN. AJ006113, AF428047, Y13057 are accession numbers in the IMGT/LIGM-DB database [21]. Other detailed results for each V-DOMAIN comprise 14 displays (not shown) as listed in Detailed view

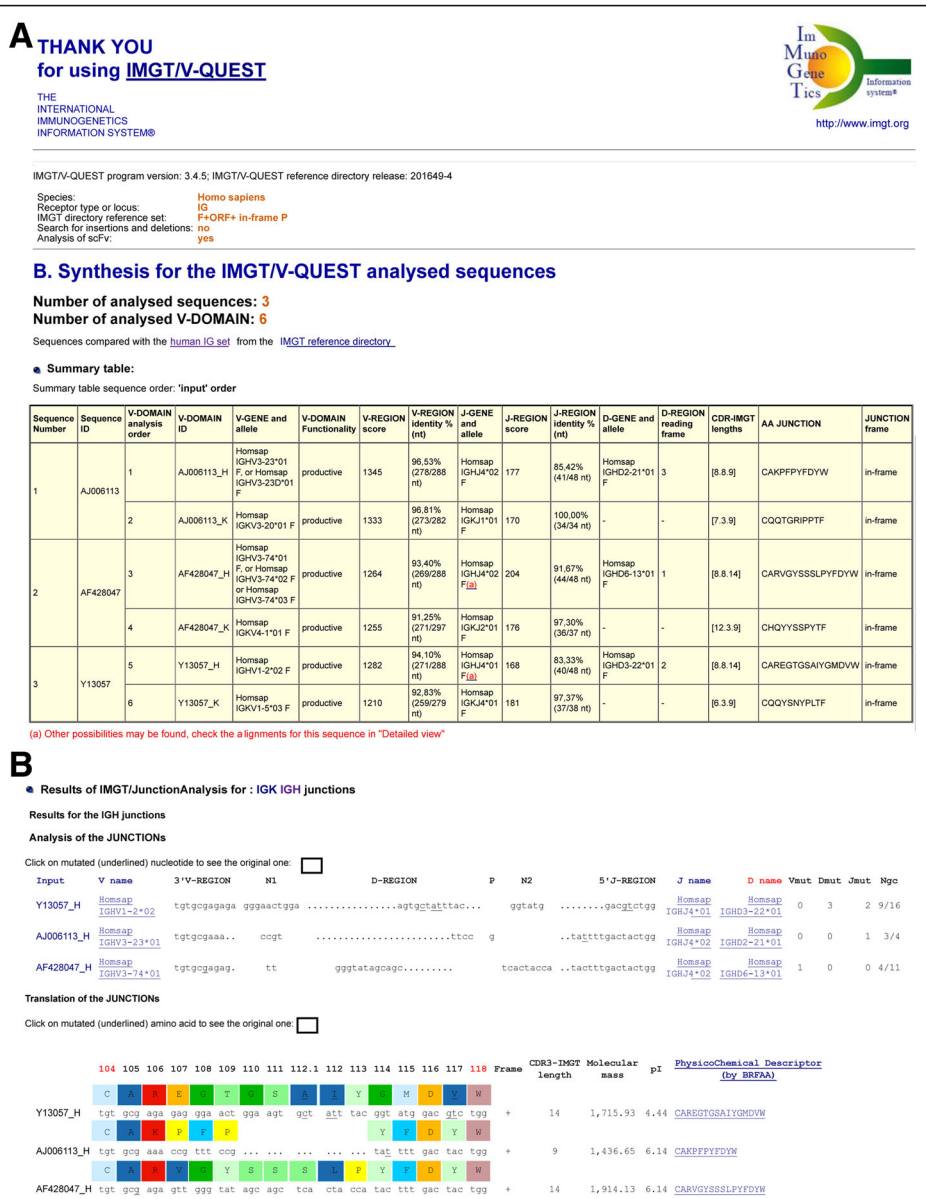

Fig. 5 IMGT/V-QUEST Synthesis view results for IG scFv sequences. **a** Three scFv were analyzed with the option "Analysis of single chain Fragment variable (scFv)". The top of the Synthesis view recalls the parameters for the analysis and provides the number of analyzed sequences (here, 3) and the number of analyzed V-DOMAIN (here, 6). The Summary table includes one line per V-DOMAIN and 2 lines per identified scFv in the submitted sequence set, identified by their order number in the set and their sequence ID. The order of V-DOMAIN analysis is indicated on the left of the V-DOMAIN ID. AJ006113, AF428047, Y13057 are accession numbers in the IMGT/LIGM-DB database [21]. Other results for each V-DOMAIN comprise 8 displays (not shown) as listed in Synthesis view. **b** Results of IMGT/JunctionAnalysis for the VH domain of the 3 scFv

the selected parameters: Species, Receptor type or locus (IG or TR), IMGT reference directory set (e.g., F + ORF + in-frame P, and options 'Search for insertions and deletions' (yes or no) and, for the current purpose, 'Analysis of scFv' (yes).

Each identified V-(D)-J-REGION (V-DOMAIN) appears individually on a different line of the Summary table. Pairs of V-DOMAIN belonging to the same scFv, are identified by having the same sequence ID (Fig. 5). For each identified V-(D)-J-REGION (V-DOMAIN), the classical results are displayed (V-GENE and allele, V-DOMAIN Functionality, V-REGION score, V-REGION identity % (nt), J-GENE and allele, J-REGION score, J-REGION identity % (nt), D-GENE and allele, D-REGION reading frame, CDR-IMGT lengths, AA JUNCTION, JUNCTION frame). Below the Summary table, Results of IMGT/JunctionAnalysis (comparison of JUNCTION of V-DOMAIN belonging to the same locus, e.g., IGH in Fig. 5) and Alignment with the closest alleles (comparison of V-DOMAIN expressing the same V

gene and allele) are provided for an analysis between V-DOMAIN of different scFv.

IMGT/V-QUEST Excel file or IMGT/HighV-QUEST CSV files results for scFv

Classically, the IMGT/V-QUEST Excel file or, for NGS, the IMGT/HighV-QUEST comma separated values (CSV) files results include eleven Excel spreadsheets or CSV files, respectively [35]. Typically, the first ten Excel spreadsheets and CSV files include one line per identified and analyzed V-(D)-J-REGION (V-DOMAIN) (for scFv, there are therefore two lines corresponding to the two V domains for each scFv). The file 11_Parameters indicates the number of submitted scFv sequences and the number of analyzed V-DOMAIN.

An additional file "12_scFv" (Fig. 6) is specific to the scFv analysis and is automatically included in the results if the option "Analysis of single chain Fragment variable (scFv)" has been selected. The "12_scFv" file includes a single line per submitted sequence identified as an scFv (i.e., with two V-DOMAIN, or at least two V-REGION, identified in the same sequence). Each line crosses two sets of 19 columns, prefixed by "1_" and by "2_", respectively, which correspond to the results of the two V domains of the scFv, with between them, two columns for the 'linker positions' and the 'linker length' in the V-orientated scFv sequence. It should be noted that the assignment "1_" and "2_" in this file is arbitrary (it is independent on the V domain analysis order number and on the relative positions of the V domains in the V-orientated scFv). In order to facilitate data extraction and reuse, VH, V-BETA and V-DELTA are in the "1_" column set for scFv which contain paired VH-VL (V-KAPPA or V-LAMBDA), V-ALPHA-V-BETA or V-GAMMA-V-DELTA (Table 2).

The IG scFv are usually made of a VH and of VL. The order of the V-DOMAIN in an scFv can be either VH-linker-VL or VL-linker-VH. Both V domains of an scFv are transcribed in a single chain and are necessarily in the same orientation of transcription. Different associations of V domains (the two domains being in the same orientation) are possible for the IG or for the TR. Analysis of scFv V-DOMAIN by IMGT/V-QUEST is done against a single species. For the TR, other V domain associations per permutation between any two V domains (not shown) are also analyzable.

1	2	3	4	5	6	7	8	9
1_V-DOMAIN analysis order	1_V-DOMAIN ID	1_V-DOMAIN positions	1_V-DOMAIN length	1_V-DOMAIN Functionality	1_V-GENE and allele	1_V-REGION score	1_V-REGION identity %	1_V-REGION identity nt
1	AJ634527_H	82..430	349	productive	Homsap IGHV3-23*01 F, or Homsap IGHV3-23D*01 F	1327	95.83	276/288 nt
3	BD190528_H	1..370	370	productive	Homsap IGHV1-69*01 F, or Homsap IGHV1-69D*01 F	1426	99.65	287/288 nt
5	CQ812868_H	1..373	373	productive	Homsap IGHV3-23*01 F, or Homsap IGHV3-23D*01 F	1426	99.65	287/288 nt
7	CS537354_H	434..782	349	productive	Homsap IGHV5-51*01 F	1390	98.26	283/288 nt
9	CS610265_H	1..349	349	productive	Homsap IGHV3-23*01 F, or Homsap IGHV3-23D*01 F	1435	100.00	288/288 nt

10	11	12	13	14	15	16	17	18	19
1_J-GENE and allele	1_J-REGION score	1_J-REGION identity %	1_J-REGION identity nt	1_D-GENE and allele	1_D-REGION reading frame	1_CDR_lengths	1_AA JUNCTION	1_JUNCTION frame	1_comments
Homsap IGHJ4*02 F	159	81.25	39/48 nt	Homsap IGHD3-22*01 F	2	8.8.9	CAKYDSSFDYW	in-frame	
Homsap IGHJ3*02 F	196	88.00	44/50 nt	Homsap IGHD3-3*01 F	3	8.8.16	CARSRRITIFGGGAFDIW	in-frame	
Homsap IGHJ4*02 F	141	77.08	37/48 nt	Homsap IGHD2-15*01 F	2	8.8.17	CERMGPCTGGSCYDTLGNW	in-frame	
Homsap IGHJ4*02 F	195	89.58	43/48 nt	Homsap IGHD7-27*01 F	3	8.8.9	CTRGDRGVDYW	in-frame	
Homsap IGHJ4*02 F	159	81.25	39/48 nt	Homsap IGHD3-16*02 F	2	8.8.9	CAKSYGAFDYW	in-frame	

20	21	22	23	24	25	26	27	28
linker positions	linker length	2_V-DOMAIN analysis order	2_V-DOMAIN ID	2_V-DOMAIN positions	2_V-DOMAIN length	2_V-DOMAIN Functionality	2_V-GENE and allele	2_V-REGION score
431..477	47	2	AJ634527_K	478..799	322	productive	Homsap IGKV1-39*01 F, or Homsap IGKV1D-39*01 F	1300
371..414	44	4	BD190528_L	415..745	331	productive	Homsap IGLV2-14*01 F	1354
374..420	47	6	CQ812868_L	421..757	337	productive	Homsap IGLV1-40*01 F	1273
390..433	44	8	CS537354_K	68..389	322	productive	Homsap IGKV1D-16*01 F	1372
350..396	47	10	CS610265_K	397..718	322	productive	Homsap IGKV1-39*01 F, or Homsap IGKV1D-39*01 F	1390

29	30	31	32	33	34	35	36	37	38	39	40
2_V-REGION identity %	2_V-REGION identity nt	2_J-GENE and allele	2_J-REGION score	2_J-REGION identity %	2_J-REGION identity nt	2_D-GENE and allele	2_D-REGION reading frame	2_CDR_lengths	2_AA JUNCTION	2_JUNCTION frame	2_comments
96.42	269/279 nt	Homsap IGKJ1*01 F	158	91.89	34/37 nt			6.3.9	CQQTADAPNTF	in-frame	
96.53	278/288 nt	Homsap IGLJ3*02 F	162	94.44	34/36 nt			9.3.10	CSSYTTRSTRVF	in-frame	
93.40	269/288 nt	Homsap IGLJ3*02 F	145	86.84	33/38 nt			9.3.12	CQSYDSSLSGSKVF	in-frame	
99.28	277/279 nt	Homsap IGKJ1*01 F	171	97.22	35/36 nt			6.3.9	CQQYNSYPRTF	in-frame	
100.00	279/279 nt	Homsap IGKJ1*01 F	158	91.89	34/37 nt			6.3.9	CQQSYSTPNTF	in-frame	

Fig. 6 IMGT/V-QUEST 12_scFv spreadsheet of the Excel file. Only available for the option "Analysis of single chain Fragment variable (scFv)"). It includes one line per identified scFv in the submitted sequence set with the characterization of the two V-DOMAIN by 19 fields each prefixed by "1" or "2". In order to facilitate data extraction, VH (for IG), V-BETA or V-DELTA (for TR) domains are always displayed from column 1 to 19 (prefixed by "1") and VL (for IG), V-ALPHA or V-GAMMA (for TR) are always displayed from column 22 to 40 (prefixed by "2") (with exception for scFv composed of 2 VH or of 2 VL domains). Columns 20 and 21 correspond to linker positions and linker length, respectively. Yellow columns have been specifically created for scFv analysis. 1_V-DOMAIN analysis order and 2_V-DOMAIN analysis order columns indicate the V-DOMAIN analysis order which is also reported in the 10 classical spreadsheets of Excel file in IMGT/V-QUEST and in the CSV files in IMGT/HighV-QUEST

Table 2 Associations of scFv V-DOMAIN analyzed by IMGT/V-QUEST

IG scFv		TR scFv	
5'V-DOMAIN	3'V-DOMAIN	5'V-DOMAIN	3'V-DOMAIN
VH	V-KAPPA	V-ALPHA	V-BETA
VH	V-LAMBDA	V-BETA	V-ALPHA
VH	VH	V-GAMMA	V-DELTA
V-KAPPA	VH	V-DELTA	V-GAMMA
V-KAPPA	V-LAMBDA		
V-KAPPA	V-KAPPA		
V-LAMBDA	VH		
V-LAMBDA	V-KAPPA		
V-LAMBDA	V-LAMBDA		

Each set of 19 columns for each V domain comprises the following fields:

1. V-DOMAIN analysis order: order in the submitted scFv sequences set (as in the ten first spreadsheets or files),
2. V-DOMAIN ID: V-DOMAIN identifier (SequenceName_LocusLetter),
3. V-DOMAIN positions: begin and end position of the identified V-DOMAIN in the V-orientated scFv sequence,
4. V-DOMAIN length: as determined by the begin and end positions,
5. V-DOMAIN Functionality: productive or unproductive,
6. V-GENE and allele: IMGT gene and allele name of the closest germline V-REGION,
7. V-REGION score: alignment score with the closest germline V-REGION,
8. V-REGION identity %: identity percentage with the closest germline V-REGION,
9. V-REGION identity nt: number of identical nt with the closest germline V-REGION,
10. J-GENE and allele: IMGT gene and allele name of the closest germline J-REGION,
11. J-REGION score: alignment score with the closest germline J-REGION,
12. J-REGION identity %: identity percentage with the closest germline J-REGION,
13. J-REGION identity nt: number of identical nt with the closest V germline J-REGION,
14. D-GENE and allele: IMGT gene and allele name of the closest germline D-REGION (as identified by IMGT/JunctionAnalysis),
15. D-REGION reading frame: reading frame 1, 2 or 3 (as identified by IMGT/JunctionAnalysis),
16. CDR_lengths: length of the 3 CDR-IMGT,
17. AA JUNCTION: amino acid sequence of the junction,
18. JUNCTION frame: frame of the junction (in-frame or out-of-frame),
19. Comments: to highlight the particularities of the V-DOMAIN, if any.

It should be noted that sequences not identified as scFv (i.e., for which only a single (or no) V-DOMAIN or V-REGION is identified) are not integrated in the "12_scFv" spreadsheet or file, so this spreadsheet or file may be empty if none of the submitted sequences are identified as scFv.

As the online version of IMGT/V-QUEST can analyze 50 sequences per run, the results for scFv analysis may potentially include the analysis of up to 100 V-DOMAIN. With the option "Search for insertions and deletions", the number of submitted sequences is restricted to 10, and the results for scFv may include the analysis of up to 20 V-DOMAIN.

In the IMGT/HighV-QUEST, the option "Search for insertions and deletions" is selected by default and the analysis includes all the identified V-DOMAIN. The new advanced functionality "Analysis of single chain Fragment variable (scFv)" provides the identification and characterization of, theoretically, up to one million domains for 500,000 submitted scFv sequences. This functionality has introduced, for the first time, the possibility of analysing simultaneously the two V domains of large scFv data sets from combinatorial libraries.

Discussion

In antibodies and T cell receptors, the antigen binding sites comprise two V-DOMAIN which are paired at the N-terminal end of the heavy and light chains for the IG and of the alpha and beta (or gamma and delta) chains for the TR [1–3]. The pairing of the two V-DOMAIN is reproduced in scFv in which the two V-DOMAIN are connected by a peptide linker. These engineered monovalent molecules were first expressed in *Escherichia coli* [43, 44] and then at the surface of filamentous phages. This methodology combined with the polymerase chain reaction (PCR) amplification of variable domains was the starting point of the construction of scFv phage combinatorial libraries [45–47], by-passing hybridoma technology and animal antibody humanization. The scFv can be expressed in various systems (bacteria, phages, yeast, plant, mammalian cells), leading to the generation of many different scFv combinatorial libraries and to the development of various technologies (such as phage or ribosome display) as an efficient tool for the screening, selection and enrichment of antibodies with a given specificity. The selection from scFv combinatorial libraries is

widely used for the discovery of novel antibody specificities for diagnostic and therapy [48–51].

Next generation sequencing (NGS) has recently emerged as a new method for the high-throughput characterization of IG and TR immune repertoires both in vivo and in vitro. Currently available NGS platforms allow the simultaneous sequencing of millions of reads. However, two challenges remain for the NGS sequencing of scFv: first, the scFv length is > 800 bp, which is too long for most NGS platforms; and second, there is no tool for the analysis of two V domains in a single chain. Up to now, NGS methods have only provided reads encompassing one V domain (400 bp), therefore losing a critical piece of information found in scFv sequences, that of the association of two specific V domains (VH and VL for the IG) by the peptide linker. Although a few approaches have been proposed, retrieving information regarding V domain association has still not been solved [52–54].

As reliable data depend on high-quality and long enough sequences to contain the full-length scFv, the new functionality "Analysis of single chain Fragment variable (scFv)" was implemented for providing the identification and full characterization of the two V domains in scFv sequences or NGS reads fulfilling these criteria.

Conclusions

The functionality "Analysis of single chain Fragment variable (scFv)" provides the identification and full characterization of the two V domains of full-length scFv in IMGT/V-QUEST online or, for NGS, in IMGT/HighV-QUEST. The functionality was used to analyse >450,000 reads of about 1000 bp, obtained from a combinatorial library, generated with the Pacific Biosciences (PacBio) RS II platform using single molecule, real-time (SMRT) circular consensus sequencing (CCS). The two V domains were identified and characterized in all reads of high-quality and sufficient length. The "Analysis of single chain Fragment variable (scFv)" will facilitate and improve the description of the scFv content of combinatorial libraries, a key information in therapeutic antibody discovery, selection and development.

The need for the analysis of sequences containing two V domains from expressed repertoires is also rapidly rising. NGS single-cell sequencing of paired chains have been obtained by a technology comprising flow focusing and encapsulation of single cells in emulsions containing magnetic beads for mRNA capture, reverse-transcription of mRNA transcripts, physical linkage of the partners by overlap extension PCR, and NGS sequencing [55]. Other developments of paired IG and TR sequences include paired recovery of transcripts and concatenation per single cell [56], single cell paired sequencing [57], capture strategies [58]. IMGT/V-QUEST and IMGT/HighV-QUEST perform classically on sequences of paired chains identified by bar-coding of single cells, each chain having a single V-DOMAN. In contrast, if the sequences of the paired chains are physically linked, the functionality "Analysis of single chain Fragment variable (scFv)" should be selected in order to identify and describe the two V-DOMAIN. Indeed, this functionality for scFv sequence analysis is generic for IG and TR and can be used without modification for libraries of single B or T cell concatenated paired expressed chains, and will facilitate the identification of novel paratopes in infections, cancers, autoimmune diseases or neurodegenerative diseases.

Abbreviations

2D: Two-dimensional; 3D: Three-dimensional; AA: Amino acid; bp: Base pair; C: Constant; CCS: Circular consensus sequencing; CDR: Complementarity determining region; CPCA: Composite protein for clinical applications; CSV: Comma separated values; D: Diversity; Da: Dalton; F: Functional; FPIA: Fusion protein for immune applications; FR: Framework region; H: Heavy; ID: Identifier; IG: Immunoglobulin; IgSF: Immunoglobulin superfamily; J: Joining; L: Light; MH: Major histocompatibility; MhSF: Major histocompatibility superfamily; NGS: Next generation sequencing; nt: Nucleotide; ORF: Open reading frame; P: Pseudogene; RPI: Related protein of the immune system; scFv: Single chain Fragment variable; SMRT: Single molecule, real-time; TR: T cell receptor; V: Variable; VH: Variable heavy; VL: Variable light

Acknowledgements

We thank Gisèle Clofent-Sanchez and Audrey Hemadou for scFv sequences samples and their implication in the development of the project. We are grateful to Gérard Lefranc, for helpful comments and to the IMGT team members for their constant motivation. We thank Géraldine Folch, Joumana Jabado-Michaloud, Safa Aouinti, Mélissa Cambon, Imène Chently, Karthik Kalyan, Anjana Kushwaha, Arthur Lavoie, Claudio Lorenzi, Perrine Pégorier, Laurène Picandet, Saida Hadi-Saljoqi, Mélanie Arrivet, Pascal Bento and Marine Peralta. IMGT® is Academic Institutional Member of the International Medical Informatics Association (IMIA) and of the Global Alliance for the Genomics and Health (GA4GH).

Funding

IMGT® is currently supported by the Centre National de la Recherche Scientifique (CNRS); the Ministère de l'Enseignement Supérieur et de la Recherche (MESR); the Montpellier University, France; the Agence Nationale de la Recherche (ANR) Labex MabImprove [ANR-10-LABX-5301]; BioCampus Montpellier; Région Languedoc-Roussillon (Grand Plateau Technique pour la Recherche (GPTR)). This work was granted access to the HPC@LR and to the High Performance Computing (HPC) resources of the Centre Informatique National de l'Enseignement Supérieur (CINES) and to Très Grand Centre de Calcul (TGCC) of the Commissariat à l'Energie Atomique et aux Energies Alternatives (CEA) under the allocation [036029] (2010-2017) made by GENCI (Grand Equipement National de Calcul Intensif). Funding for open access charge: IMGT (Montpellier University and CNRS). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Availability of data and materials

The functionality is available as an option at http://www.imgt.org/IMGT_vquest/vquest in Advanced Functionalities. Data for testing the functionality are available at: http://www.imgt.org/IMGT_vquest/share/textes/testsets.html#set3.

Authors' contributions

VG and MPL conceived and designed the experiments. VG designed the algorithm and implemented the tool. PD implemented the functionality for scFv in IMGT/HighV-QUEST. VG and MPL wrote the paper. VG, PD, SK and MPL supervised the project. All the authors have read and approved the final manuscript.

Competing interests
The authors declare that they have no competing interests.

Consent for publication
Not applicable.

Ethics approval and consent to participate
Not applicable.

Publisher's Note

Received: 3 April 2017 Accepted: 16 June 2017
Published online: 26 June 2017

References

1. Lefranc M-P. Immunoglobulin and T cell receptor genes: IMGT® and the birth and rise of immunoinformatics. Front Immunol. 2014;5:22. doi:10.3389/fimmu.2014.00022.
2. Lefranc M-P, Lefranc G. The immunoglobulin FactsBook. London: Academic; 2001. p. 1–458.
3. Lefranc M-P, Lefranc G. The T cell receptor FactsBook. London: Academic; 2001. p. 1–398.
4. Lefranc M-P, Giudicelli V, Duroux P, Jabado-Michaloud J, Folch G, Aouinti S, Carillon E, Duvergey H, Houles A, Paysan-Lafosse T, Hadi-Saljoqi S, Sasorith S, Lefranc G, Kossida S. IMGT®, the international ImMunoGeneTics information system® 25 years on. Nucleic Acids Res. 2015;43:D413–22. doi:10.1093/nar/gku1056.
5. IMGT®, the international ImMunoGeneTics information system®. http://www.imgt.org. Accessed 27 Mar 2017.
6. Giudicelli V, Lefranc M-P. IMGT-ONTOLOGY 2012. Front Genet. 2012;3:79. doi:10.3389/fgene.2012.00079.
7. Lefranc M-P. From IMGT-ONTOLOGY IDENTIFICATION axiom to IMGT standardized keywords: for immunoglobulins (IG), T cell receptors (TR), and conventional genes. Cold Spring Harb Protoc. 2011;6:604–13. doi:10.1101/pdb.ip82.
8. Lefranc M-P. From IMGT-ONTOLOGY DESCRIPTION axiom to IMGT standardized labels: for immunoglobulin (IG) and T cell receptor (TR) sequences and structures. Cold Spring Harb Protoc. 2011;6:614–26. doi:10.1101/pdb.ip83.
9. Lefranc M-P. From IMGT-ONTOLOGY CLASSIFICATION axiom to IMGT standardized gene and allele nomenclature: for immunoglobulins (IG) and T cell receptors (TR). Cold Spring Harb Protoc. 2011;6:627–32. doi:10.1101/pdb.ip84.
10. Lefranc M-P, Pommié C, Ruiz M, Giudicelli V, Foulquier E, Truong L, Thouvenin-Contet V, Lefranc G. IMGT unique numbering for immunoglobulin and T cell receptor variable domains and Ig superfamily V-like domains. Dev Comp Immunol. 2003;27:55–77. doi:10.1016/S0145-305X(02)00039-3.
11. Lefranc M-P, Pommié C, Kaas Q, Duprat E, Bosc N, Guiraudou D, Jean C, Ruiz M, Da Piedade I, Rouard M, Foulquier E, Thouvenin V, Lefranc G. IMGT unique numbering for immunoglobulin and T cell receptor constant domains and Ig superfamily C-like domains. Dev Comp Immunol. 2005;29:185–203. doi:10.1016/j.dci.2004.07.003.
12. Lefranc M-P, Duprat E, Kaas Q, Tranne M, Thiriot A, Lefranc G. IMGT unique numbering for MHC groove G-DOMAIN and MHC superfamily (MhcSF) G-LIKE-DOMAIN. Dev Comp Immunol. 2005;29:917–38. doi:10.1016/j.dci.2005.03.003.
13. Ruiz M, Lefranc M-P. IMGT gene identification and colliers de perles of human immunoglobulin with known 3D structures. Immunogenetics. 2002;53:857–83. doi:10.1007/s00251-001-0408-6.
14. Lefranc M-P. IMGT unique numbering for the variable (V), constant (C), and groove (G) domains of IG, TR, MH, IgSF, and MhSF. Cold Spring Harb Protoc. 2011;6:633–42. doi:10.1101/pdb.ip85.
15. Lefranc M-P. IMGT collier de perles for the variable (V), constant (C), and groove (G) domains of IG, TR, MH, IgSF, and MhSF. Cold Spring Harb Protoc. 2011;6:643–51. doi:10.1101/pdb.ip86.
16. Lefranc M-P. Immunoinformatics of the V, C, and G domains: IMGT® definitive system for IG, TR and IgSF, MH, and MhSF. Methods Mol Biol. 2014;1184:59–107. doi:10.1007/978-1-4939-1115-8_4.
17. Lefranc M-P, Clément O, Kaas Q, Duprat E, Chastellan P, Coelho I, Combres K, Ginestoux C, Giudicelli V, Chaume D, Lefranc G. IMGT-choreography for immunogenetics and immunoinformatics. In Silico Biology. 2005;5:45–60.
18. Duroux P, Kaas Q, Brochet X, Lane J, Ginestoux C, Lefranc M-P, Giudicelli V. IMGT-kaleidoscope, the formal IMGT-ONTOLOGY paradigm. Biochimie. 2008;90:570–83. doi:10.1016/j.biochi.2007.09.003.
19. Lefranc M-P, Giudicelli V, Regnier L, Duroux P. IMGT®, a system and an ontology that bridge biological and computational spheres in bioinformatics. Brief Bioinform. 2008;9(4):263–75. doi:10.1093/bib/bbn014.
20. Giudicelli V, Chaume D, Lefranc M-P. IMGT/GENE-DB: a comprehensive database for human and mouse immunoglobulin and T cell receptor genes. Nucleic Acids Res. 2005;33:D256–61. doi:10.1093/nar/gkh412.
21. Giudicelli V, Duroux P, Ginestoux C, Folch G, Jabado-Michaloud J, Chaume D, Lefranc M-P. IMGT/LIGM-DB, the IMGT comprehensive database of immunoglobulin and T cell receptor nucleotide sequences. Nucleic Acids Res. 2006;34:D781–4. doi:10.1093/nar/gkj088.
22. Ehrenmann F, Kaas Q, Lefranc M-P. IMGT/3Dstructure-DB and IMGT/DomainGapAlign: a database and a tool for immunoglobulins or antibodies, T cell receptors, MHC. IgSF and MhcSF Nucleic Acids Res. 2010;38:D301–7. doi:10.1093/nar/gkp946.
23. Ehrenmann F, Lefranc M-P. IMGT/3Dstructure-DB: querying the IMGT database for 3D structures in immunology and immunoinformatics (IG or antibodies, TR, MH, RPI, and FPIA). Cold Spring Harb Protoc. 2011;6:750–61. doi:10.1101/pdb.prot5637.
24. Giudicelli V, Chaume D, Lefranc M-P. IMGT/V-QUEST, an integrated software program for immunoglobulin and T cell receptor V-J and V-D-J rearrangement analysis. Nucleic Acids Res. 2004;32:W435–40. doi:10.1093/nar/gkh412.
25. Brochet X, Lefranc M-P, Giudicelli V. IMGT/V-QUEST: the highly customized and integrated system for IG and TR standardized V-J and V-D-J sequence analysis. Nucleic Acids Res. 2008;36:W503–8. doi:10.1093/nar/gkn316.
26. Giudicelli V, Brochet X, Lefranc M-P. IMGT/V-QUEST: IMGT standardized analysis of the immunoglobulin (IG) and T cell receptor (TR) nucleotide sequences. Cold Spring Harb Protoc. 2011;6:695–715. doi:10.1101/pdb.prot5633.
27. Yousfi Monod M, Giudicelli V, Chaume D, Lefranc M-P. IMGT/JunctionAnalysis: the first tool for the analysis of the immunoglobulin and T cell receptor complex V-J and V-D-J JUNCTIONs. Bioinformatics. 2004;20:i379–85. doi:10.1093/bioinformatics/bth945.
28. Giudicelli V, Lefranc M-P. IMGT/JunctionAnalysis: IMGT standardized analysis of the V-J and V-D-J junctions of the rearranged immunoglobulins (IG) and T cell receptors (TR). Cold Spring Harb Protoc. 2011;6:716–25. doi:10.1101/pdb.prot5634.
29. Giudicelli V, Protat C, Lefranc M-P. The IMGT strategy for the automatic annotation of IG and TR cDNA sequences: IMGT/automat. In: Proceedings of the European conference on computational biology (ECCB 2003)," data and knowledge bases, poster DKB_31, ECCB. Paris: Institut National de Recherche en Informatique et en Automatique; 2003. p. 103–4.
30. Giudicelli V, Chaume D, Jabado-Michaloud J, Lefranc M-P. Immunogenetics sequence annotation: the strategy of IMGT based on IMGT-ONTOLOGY. Stud Health Technol Inform. 2005;116:3–8.
31. Alamyar E, Giudicelli V, Li S, Duroux P, Lefranc M-P, 1. IMGT/HighV-QUEST: the IMGT® web portal for immunoglobulin (IG) or antibody and T cell receptor (TR) analysis from NGS high throughput and deep sequencing. Immunome Res. 2012;8:2.
32. Alamyar E, Duroux P, Lefranc M-P, Giudicelli V. IMGT(®) tools for the nucleotide analysis of immunoglobulin (IG) and T cell receptor (TR) V-(D)-J repertoires, polymorphisms, and IG mutations: IMGT/V-QUEST and IMGT/HighV-QUEST for NGS. Methods Mol Biol. 2012;882:569–604. doi:10.1007/978-1-61779-842-9_32.
33. Li S, Lefranc M-P, Miles JJ, Alamyar E, Giudicelli V, Duroux P, Freeman JD, Corbin VDA, Scheerlinck J-P, Frohman MA, Cameron PU, Plebanski M, Loveland B, Burrows SR, Papenfuss AT, Gowans EJ. IMGT/HighV QUEST paradigm for T cell receptor IMGT clonotype diversity and next generation repertoire immunoprofiling. Nat Commun. 2013;4:2333. doi:10.1038/ncomms3333.
34. Giudicelli V, Duroux P, Lavoie A, Aouinti S, Lefranc M-P, Kossida S. From IMGT-ONTOLOGY to IMGT/HighVQUEST for NGS immunoglobulin (IG) and T cell receptor (TR) repertoires in autoimmune and infectious diseases. Autoimmun Infec Dis. 2015;1:1.
35. Alamyar E, Giudicelli V, Duroux P, Lefranc M-P. Antibody V and C domain sequence, structure and interaction analysis with special reference to IMGT®. In: Ossipow V, Fischer N, editors. Monoclonal antibodies: Methods and

Protocols, Second edition. New York: Humana Press, Springer Science +Business Media, LLC; 2014. Methods Mol Biol. 1131:337-81. doi:10.1007/978-1-62703-992-5_21.
36. Ehrenmann F, Lefranc M-P. IMGT/DomainGapAlign: IMGT standardized analysis of amino acid sequences of variable, constant, and groove domains (IG, TR, MH, IgSF, MhSF). Cold Spring Harb Protoc. 2011;6:737–49. doi:10.1101/pdb.prot5636.
37. Ehrenmann F, Giudicelli V, Duroux P, Lefranc M-P. IMGT/collier-de-perles: IMGT standardized representation of domains (IG, TR, and IgSF variable and constant domains, MH and MhSF groove domains). Cold Spring Harb Protoc. 2011;6:726–36. doi:10.1101/pdb.prot5635.
38. Baum TP, Hierle V, Pascal N, Bellahcene F, Chaume D, Lefranc M-P, Jouvin-Marche E, Marche PN, Demongeot J. IMGT/GeneInfo: T cell receptor gamma TRG and delta TRD genes in database give access to all TR potential V(D)J recombinations. BMC Bioinformatics. 2006;7:224. doi:10.1186/1471-2105-7-224.
39. Lane J, Duroux P, Lefranc M-P. From IMGT-ONTOLOGY to IMGT/LIGMotif: the IMGT® standardized approach for immunoglobulin and T cell receptor gene identification and description in large genomic sequences. BMC Bioinformatics. 2010;11:223. doi:10.1186/1471-2105-11-223.
40. Aouinti S, Malouche D, Giudicelli V, Kossida S, Lefranc M-P. IMGT/HighV-QUEST statistical significance of IMGT clonotype (AA) diversity per gene for standardized comparisons of next generation sequencing immunoprofiles of immunoglobulins and T cell receptors. PLoS One. 2015;10(11):e0142353. doi:10.1371/journal.pone.0146702.
41. Aouinti S, Giudicelli V, Duroux P, Malouche D, Kossida S, Lefranc M-P. IMGT/StatClonotype for pairwise evaluation and visualization of NGS IG and TR IMGT clonotype (AA) diversity or expression from IMGT/HighV-QUEST. Front Immunol. 2016;7:339. doi:10.3389/fimmu.2016.00339.
42. Rhoads A, Au KF. PacBio sequencing and its applications. Genomics Proteomics Bioinformatics. 2015;13:278–89. doi:10.1016/j.gpb.2015.08.002.
43. Huston JS, Levinson D, Mudgett-Hunter M, Tai MS, Novotný J, Margolies MN, Ridge RJ, Bruccoleri RE, Haber E, Crea R. Protein engineering of antibody binding sites: recovery of specific activity in an anti-digoxin single-chain Fv analogue produced in *Escherichia coli*. Proc Natl Acad Sci U S A. 1988;85(16):5879–83.
44. Bird RE, Hardman KD, Jacobson JW, Johnson S, Kaufman BM, Lee SM, Pope SH, Riordan GS, Whitlow M. Single-chain antigen-binding proteins. Science. 1988;242(4877):423–6.
45. McCafferty J, Griffiths AD, Winter G, Chiswell DJ. Phage antibodies: filamentous phage displaying antibody variable domains. Nature. 1990; 348(6301):552–4.
46. Marks JD, Hoogenboom HR, Bonnert TP, McCafferty J, Griffiths AD, Winter G. By-passing immunization. Human antibodies from V-gene libraries displayed on phage. J Mol Biol. 1991;222(3):581–97.
47. Griffiths AD, Malmqvist M, Marks JD, Bye JM, Embleton MJ, McCafferty J, Baier M, Holliger KP, Gorick BD, Hughes-Jones NC, et al. Human anti-self antibodies with high specificity from phage display libraries. EMBO J. 1993;12(2):725–34.
48. Ahmad ZA, Yeap SK, Ali AM, Ho WY, Alitheen NBM, Hamid M. ScFv antibody: principles and clinical application. Clin Dev Immunol. 2012;2012:980250. doi:10.1155/2012/980250.
49. Deramchia K, Jacobin-Valat M-J, Laroche-Traineau J, Bonetto S, Sanchez S, Dos Santos P, Massot P, Franconi JM, Martineau P, Clofent-Sanchez G. By-passing large screening experiments using sequencing as a tool to identify scFv fragments targeting atherosclerotic lesions in a novel in vivo phage display selection. Int J Mol Sci. 2012;13(6):6902–23. doi:10.3390/ijms13066902.
50. Weber M, Bujak E, Putelli A, Villa A, Matasci M, Gualandi L, Hemmerle T, Wulhfard S, Neri D. A highly functional synthetic phage display library containing over 40 billion human antibody clones. PLoS One. 2014;9(6): e100000. doi:10.1371/journal.pone.0100000.
51. Kügler J, Wilke S, Meier D, Tomszak F, Frenzel A, Schirrmann T, Dübel S, Garritsen H, Hock B, Toleikis L, Schütte M, Hust M. Generation and analysis of the improved human HAL9/10 antibody phage display libraries. BMC Biotechnol. 2015;15:10. doi:10.1186/s12896-015-0125-0.
52. Larman HB, Xu GJ, Pavlova NN, Elledge SJ. Construction of a rationally designed antibody platform for sequencing-assisted selection. Proc Natl Acad Sci U S A. 2012;109:18523–8. doi:10.1073/pnas.1215549109.
53. Ravn U, Gueneau F, Baerlocher L, Osteras M, Desmurs M, Malinge P, Magistrelli G, Farinelli L, Kosco-Vilbois MH, Fischer N. By-passing in vitro screening–next generation sequencing technologies applied to antibody display and in silico candidate selection. Nucleic Acids Res. 2010;38:e193. doi:10.1093/nar/gkq789.
54. Ravn U, Didelot G, Venet S, Ng K, Gueneau F, Rousseau F, Calloud S, Kosco-Vilbois M, Fischer N. Deep sequencing of phage display libraries to support antibody discovery. Methods. 2013;60(1):99–110. doi:10.1016/j.ymeth.2013.03.001.
55. McDaniel JR, DeKosky BJ, Tanno H, Ellington AD, Georgiou G. Ultra-high-throughput sequencing of the immune receptor repertoire from millions of lymphocytes. Nat Protoc. 2016;11(3):429–42. doi:10.1038/nprot.2016.024. Epub 2016 Feb 4.
56. Redmond D, Poran A, Elemento O. Single-cell TCRseq: paired recovery of entire T-cell alpha and beta chain transcripts in T-cell receptors from single-cell RNAseq. Genome Med. 2016;8(1):80.
57. Knies D, Klobuch S, Xue SA, Birtel M, Echchannaoui H, Yildiz O, Omokoko T, Guillaume P, Romero P, Stauss H, Sahin U, Herr W, Theobald M, Thomas S, Voss RH. An optimized single chain TCR scaffold relying on the assembly with the native CD3-complex prevents residual mispairing with endogenous TCRs in human T-cells. Oncotarget. 2016;7(16):21199–221.
58. Hanson WM, Chen Z, Jackson LK, Attaf M, Sewell AK, Heemstra JM, Phillips JD. Reversible oligonucleotide chain blocking enables bead capture and amplification of T-cell receptor α and β chain mRNAs. J Am Chem Soc. 2016;138(35):11073–6.
59. IMGT/LIGM-DB labels. http://www.imgt.org/ligmdb/label#. Accessed 27 Mar 2017.
60. Correspondence between chain types and C genes: IG and TR (all vertebrate species). http://www.imgt.org/IMGTrepertoire/LocusGenes/correspondencedesign/corresdesign.html. Accessed 27 Mar 2017.