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

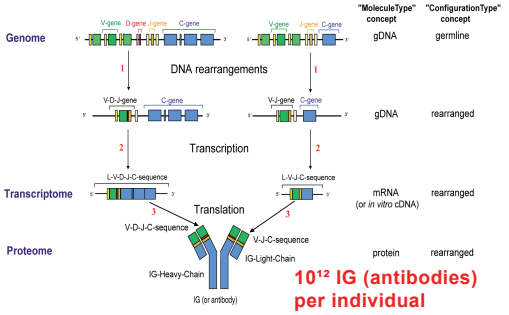
IMGT®, the international ImMunoGeneTics information system®
Université de Montpellier (UM), Laboratoire d'ImmunoGénétique Moléculaire (LIGM),
Institut de Génétique Humaine (IGH), UMR 9002 CNRS-UM, Montpellier (France)

IMGT®, the international ImMunoGeneTics information system® (http://www.imgt.org) [1], 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 [2]. Immunoglobulins (IG) or antibodies [3] and T cell receptors (TR) [4] 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) [5] and, for next generation sequencing (NGS), by IMGT/HighVQUEST, the high throughput version of IMGT/V-QUEST (portal begun in 2010, 500,000 sequences per batch) [6, 7]. The analysis of NGS scFv represents a challenge by their length (~850 bp) 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. The functionality "Analysis 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 [8]. 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. Nowadays, advances in NGS technology allow for longer reads (1000 bp and more), therefore full-length scFv. The *in vitro* combinatorial libraries which mimic the *in vivo* natural diversity of the immune adaptive responses are extensively screened for the discovery of novel antigen binding specificities and new therapeutic candidates.

[1] Lefranc M-P et al. Nucleic Acids Res 43:413-422 (2015) PMID: 25378316
[2] Lefranc M.-P. Front Immunol. 5:22 (2014) PMID: 24600447
[3] Lefranc M-P, Lefranc G. The Immunoglobulin FactsBook (2001)
[4] Lefranc M-P, Lefranc G. The T cell receptor FactsBook (2001)
[5] Brochet X. et al. Nucleic Acids Res 36:W503-8 (2008) PMID: 18503082
[6] Alamyar E. et al. Immunome Res. 8:1:2 (2012) PMID: 22647994
[7] Li S. et al. Nat. Comm. 4:2333 (2013) PMID: 23995877
[8] Giudicelli V. et al. BMC Immunol. 18:(1):35 (2017) PMID: 28651553

IG diversity

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.

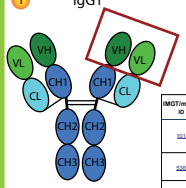


10^{12} IG (antibodies) per individual

scFv

A single chain Fragment variable (scFv) is an engineered fragment of an immunoglobulin (IG) or antibody, composed of two variable domains (V-DOMAIN) connected by a linker (2). For comparison a typical IgG1 is shown in (1). A scFv is a single chain of about 235-280 amino acids (AA) and approximate molecular weight of 26,000 Da, the two V-DOMAIN of about 110-130 AA each being connected by the linker generally of about 15-20 AA. Typically, a scFv comprises one IG heavy variable domain (VH) and one IG light variable domain (VL), in the VH-VL or VL-VH format (2). The VH-VH or VL-VL format generated in combinatorial libraries is usually not intended and artefactual.
A typical scFv with its VH and VL domains retains the full antigen-binding capability providing important applications in medicine, laboratory diagnosis and research. For example, in IMGT/mAb-DB (http://www.imgt.org/mAb-DB/), the IMGT® database for therapeutic monoclonal antibodies (mAbs), there are 14 scFv in clinical trials (phases I, II, III or M) (3).

1 IgG1



2 scFv formats

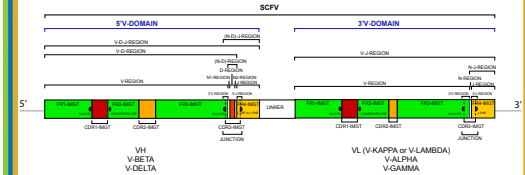
3 Examples of scFv with clinical applications

IMGT/mAb-DB ID	INN (International Nonproprietary Name)	INN Num.	Format	Clinical indication	Development status	Clinical domain
301	blinatumomab	9029		Lymphoma, B cell	Phase II	Oncology
				B cell acute lymphoblastic leukemia (B cell ALL)	Phase M	Hematology Oncology
336	brolucizumab	10053		Age-related macular degeneration (AMD), neovascular (wet)	Phase II	Ophthalmology
237	oportuzumab monatox	9045		Cancers, head and neck	Phase II/III	Oncology
				Cancers, bladder	Phase II	Oncology

scFv analysis challenges

This is a three fold issue. Firstly, experimentally, the traditional approach of Sanger sequencing of scFv is limited to a couple of dozens sequences. Nowadays, this has been overcome by the next generation sequencing (NGS) approach, whereas a much greater number of sequences can be evaluated. Secondly, owing to technological limitation the size of the NGS reads was not sufficient to cover the full length of the scFv. Recently, this has also been overcome by the introduction of PacBio SMRT sequencing, for instance. Thirdly, the computational challenge is that up until now, there was no functionality available to the scientific community to analyse the scFv. IMGT®, with the current work, introduced this novel functionality in IMGT/V-QUEST and, for NGS, in IMGT/HighV-QUEST.

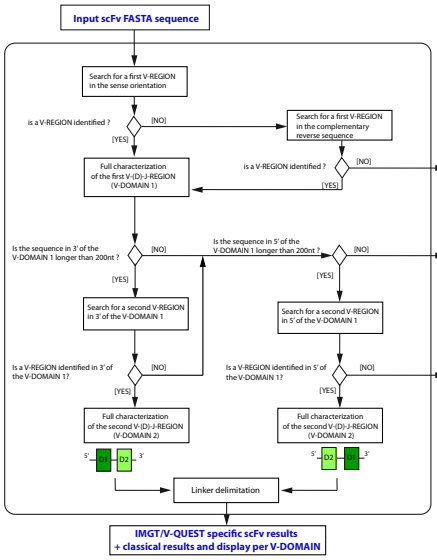
scFv prototype and IMGT labels for a complete description



Algorithm and online display of results

The "Analysis of single chain Fragment variable (scFv)" is an option of IMGT/V-QUEST (http://www.imgt.org/IMGT_vquest/vquest) and IMGT/HighV-QUEST (http://www.imgt.org/HighV-QUEST/login.action). Once selected, up to 50 scFv FASTA sequences for IMGT/V-QUEST and 500,000 for IMGT/HighV-QUEST can be analysed per run.

1 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.



2 IMGT/V-QUEST Detailed view results for scFv. (A) 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.

A

IMGT/V-QUEST program version: 3.4.4; IMGT/V-QUEST reference directory release: 201711-1

Species: Homo sapiens
Receptor type or locus: IG
IMGT directory reference set: F+ORF+ in-frame P
Search for insertions and deletions: no
Analysis of scFv: yes

A. Detailed results for the IMGT/V-QUEST analysed sequences

Number of analysed sequences: 3
Number of analysed V-DOMAIN: 6

1 AJ006113_H , 2 AJ006113_K , 3 AF428047_H , 4 AF428047_K , 5 Y13057_H , 6 Y13057_K

Identified scFv:

Sequence ID	5'V-DOMAIN ID	5'V-DOMAIN positions	5'V-DOMAIN length	linker positions	linker length	3'V-DOMAIN ID	3'V-DOMAIN positions	3'V-DOMAIN length
AJ006113	1_AJ006113_H	1..349	349	350..384	35	2_AJ006113_K	385..708	324
AF428047	3_AF428047_H	1..364	364	365..435	71	4_AF428047_K	436..775	340
Y13057	5_Y13057_H	1..364	364	365..408	44	6_Y13057_K	409..730	322

B

V-DOMAIN: 1 AJ006113_H (associated V-DOMAIN: 2 AJ006113_K)

Sequence compared with the human IG set from the IMGT reference directory

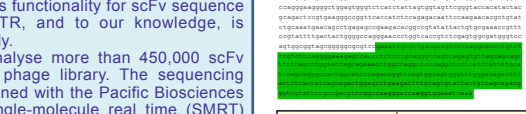
>AJ006113_H

Result summary:	Productive IGH rearranged sequence: (no stop codon and in-frame junction)		
V-GENE and allele	Homsap IGHV3-23*01 F, or Homsap IGHV3-23D*01 F	score = 1345	identity = 96,53% (278/288 nt)
J-GENE and allele	Homsap IGHJ4*02 F	score = 177	identity = 85,42% (41/48 nt)
D-GENE and allele by IMGT/JunctionAnalysis	Homsap IGHD2-21*01 F	D-REGION is in reading frame 3	
FR-IMGT lengths, CDR-IMGT lengths and AA JUNCTION	[25.17.38.11]	[8.8.9]	CAKPFPYFDYW

V-DOMAIN: 2 AJ006113_K (associated V-DOMAIN: 1 AJ006113_H)

Sequence compared with the human IG set from the IMGT reference directory

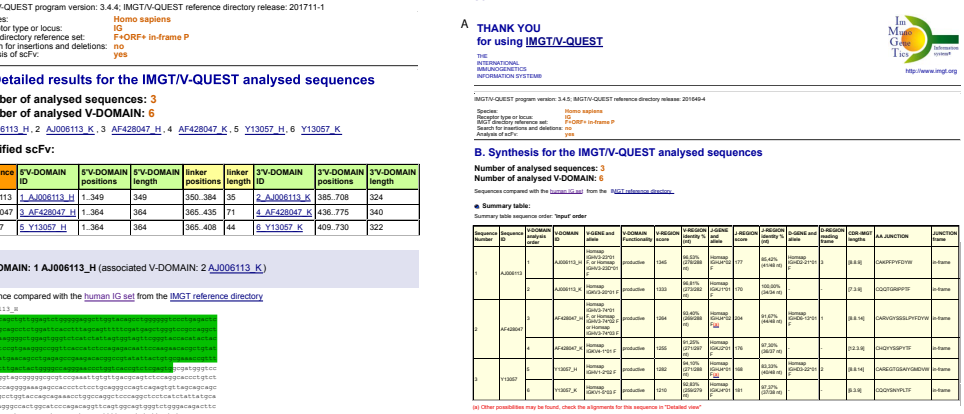
>AJ006113_K



Result summary:	Productive IGH rearranged sequence: (no stop codon and in-frame junction)		
V-GENE and allele	Homsap IGKV3-20*01 F	score = 1333	identity = 96,81% (273/282 nt)
J-GENE and allele	Homsap IGKJ1*01 F	score = 170	identity = 100,00% (34/34 nt)
FR-IMGT lengths, CDR-IMGT lengths and AA JUNCTION	[26.17.36.10]	[7.3.9]	CQQTGRIPPTF

3 IMGT/V-QUEST Synthesis view results for scFv. (A) Three scFv were analyzed with the option "Analysis of single chain Fragment variable (scFv)". The Summary table includes, for each sequence identified as a scFv, 2 lines corresponding to the two V-DOMAIN. Each V-DOMAIN is identified by a number (column 3, V-DOMAIN analysis order in the submitted set) and its ID (column 4, sequence ID followed by an underscore and a capital letter for the locus as identified by IMGT/V-QUEST (e.g., H for IGH, K for IGK). (B) Results of IMGT/JunctionAnalysis for the VH domain of the 3 scFv.

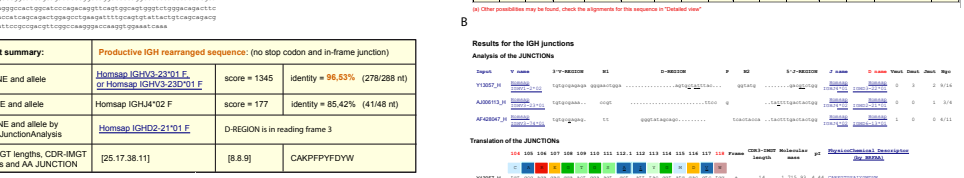
A



B

Results for the IGH junctions

Analysis of the JUNCTIONS



Translation of the JUNCTIONS

Conclusion

The new functionality "Analysis of single chain Fragment variable (scFv)" provides the identification and full characterization of the two V-DOMAIN of full-length scFv by IMGT/V-QUEST online or, for NGS, by IMGT/HighV-QUEST. This functionality for scFv sequence analysis is generic for IG and TR, and to our knowledge, is proposed by IMGT online tools, only.
This functionality was used to analyse more than 450,000 scFv sequences from a combinatorial phage library. The sequencing reads of about 1000 bp were obtained 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 fully characterized in 89 % of the ~348,000 reads filtered for their sequencing quality and 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.

Perspectives

The need for the analysis of NGS sequences containing two V domains from IG or TR expressed repertoires is also rapidly rising with novel methodological advances, as illustrated by single-cell sequencing of paired chains, paired recovery of transcripts and concatenation per single cell, or capture strategies.
As IMGT/HighV-QUEST is generic for IG and TR, the functionality for the 'Analysis of single chain Fragment variable (scFv)' can be used, without any change, for the characterization of the two V domains of various NGS single chains (IG or TR) which mimic the V domain pairing of the natural antigen receptor binding sites. It is expected that this will facilitate the identification of novel paratopes in infections, cancers, autoimmune diseases or neurodegenerative diseases.

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 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