A Dataset of Sequences with Manually Curated V(D)J Designations

Mikaël Salson[1☯], Aurélie Caillault[2], Marc Duez[4], Yann Ferret[2], Alice Fievet[7], Michaela Kotrova[8], Florian Thonier[6], Patrick Villarese[6], Stephanie Wakeman[5], Gary Wright[3], Mathieu Giraud[1☯]

1 CRIStAL (UMR 9189 CNRS, Université de Lille) and Inria Lille, France
2 CHRU Lille, Department of Hematology, Biology and Pathology Center, Lille, France
3 Great Ormond Street Hospital, London, UK
4 School of Social and Community Medicine, University of Bristol, Bristol, UK
5 Bristol Genetics Laboratory, Southmead Hospital, North Bristol NHS Trust, UK
6 Inserm, Hôpital Necker – Enfants Malades, Paris, France
7 CHU Rennes, Service d'hématologie – Rennes, France
8 Charles University and University Hospital Motol – Prague, Czech Republic

☯ These authors contributed equally to this work. contact@vidjil.org

Abstract. Repertoire Sequencing (RepSeq) studies analyze and quantify clones from lymphocyte sequences with V(D)J recombinations. Mapping a sequence to V(D)J genes is a fundamental operation in immunoinformatics. This operation should be of highest quality possible. We present here a collection of 238 test sequences coming from patient clones, covering all human TR and Ig loci, with manually curated V(D)J designations. These designations were checked by hand, possibly with the help of some bioinformatics tools. Tests may range from very easy cases with unambiguous V(D)J designations to borderline or difficult cases, including incomplete or unusual recombinations or translocations. This collection of sequences, distributed as open-source data, may help to benchmark or to improve the robustness of any software doing immune repertoire sequencing (RepSeq) analysis.

1 Introduction

V(D)J recombinations and Repertoire Sequencing (RepSeq). The lymphocytes play a key role in the adaptive immunological system, enabling specific immune response against a wide range of infections. The diversity of the lymphocytes mainly comes from the V(D)J recombinations. These recombinations influence the production of antibodies (Ig) and antigen receptors (TR) [1, 2]. V(D)J recombinations occur in B-cell heavy chains (IgH, see Figure 1) and T-cell β and δ chains (TRβ and δ), whereas VJ recombinations occur in B-cell light chains κ (Igκ) and λ (Igλ), and T-cell α and γ chains (TRα and γ).

High-throughput sequencing (HTS) now enables the deep sequencing of a lymphoid population: Repertoire Sequencing (RepSeq) studies are based on thousands or billions of reads with V(D)J recombinations coming from lymphocytes. The RepSeq studies usually analyze and quantify the *clones* that share a same V(D)J recombination. These clones may come from some immune response, but also from a pathology. Indeed, V(D)J recombinations are useful markers of lymphocyte populations. In leukemia, they are used to quantify the minimal residual disease (MRD) during patient follow-up [3].

RepSeq software. Dedicated Repertoire Sequencing (RepSeq) methods and software [4] have to take into account the specificity of V(D)J recombinations to correctly handle small recombinations, somatic hypermutations, and short insertions.

For more than 20 years, IMGT developed many tools for the analysis of sequences with V(D)J recombinations [5–8]. Since the first Repertoire Sequencing studies in 2009, new software able to deal with up to millions of sequences have appeared, including [9], Decombinator [10], IgBlast [11], IMSEQ [12]. miTCR [13], MiXCR [14], TCRklass [15] and Vidjil [16]. At the heart of these algorithms is optimized comparison of the reads against germline databases, such as the genes standardized in IMGT/GENE-DB [17], using string matching techniques such as optimized dynamic programming or statistical models.

Some other programs enable to further analyze or visualize the whole lymphocyte population, in particular by computing various statistics. These programs possibly post-process some of the former software outputs, and include ARResT [18], IgGalaxy [19], ImmunExplorer [20], tcR [21], VDJviz or the Vidjil web application [23].

The goal of these analysis and visualization software is mostly to analyze reads from a RepSeq dataset and to gather them into *clones* that should represent a biological clone. Several definitions of a clone are used: One may primarily look at the V, possibly D, and J gene/alleles names, or the amino acid sequence or the nucleic sequence, either on the CDR3 or on some larger sequence. One may also tolerate none or a few mutations between reads assigned to a same clone. These various definitions may depend on the RepSeq study. For instance, for IgH recombinations, one may would like distinguish sequences with different hypermutation patterns. For a clinical application, acute (ALL) and chronic (CLL) lymphoblastic leukemias have clones with different patterns of mutations that could be handled by specific clone gathering methods.

Motivation and Contents. Even if the V(D)J designation is not the only key to gather clones, it is the common way to describe the haematopoiesis of some clone and to enable the computation of statistics of the whole repertoire.

Mapping sequences onto V(D)J genes was done before RepSeq studies: In particular, the clinical onco-haematological practice in the diagnosis of leukemia usually requires to identify some clones with their V(D)J designation. This operation may be renewed at the emergence of new clones during the follow-up. For that purpose, many labs use a combination of IMGT/V-QUEST and manual analysis. This mapping operation should be of highest possible quality as it has influence on the diagnosis and possibly on the treatments. It may influence the design of patient specific primers for the MRD in ALL. The V gene mutation status is a prognosis factor in CLL [24].

```
                                      ++              **
IGHV3-48*01 ...TGTGTATTACTGTGCGAGAGA
clone       ...TGTGTATTACTGTGCGAGAGAAAATAGTGGCTACGATTTGACTACTGGGGCCAGGG...
IGHD5-12*01                       gtggatATAGTGGCTACGATTac
                                  123456             4321
IGHJ4*02                                        actacTTTGACTACTGGGGCCAGGG...
                                                1234567
```

Fig 1. Example of VDJ recombination on the IgH locus, from a leukaemic patient from the Lille hospital. The `clone` sequence can be described as `IGHV3-48 0/AA/6 IGHD5-12 3//6 IGHJ4`;

V/D recombination : `0/AA/6`. The V gene `IGHV3-48` was kept without trimming. Only the end of this 296 bp gene is displayed here. The D gene, `IGHD5-12`, was trimmed on its 6 first nucleotides (`gtggat`). Two nucleotides `AA`, marked with `++`, were added between V and D genes.

D/J recombination: `2//7`, `3//6` or `4//5`. In the IgH locus, the D/J recombination is actually the first one, occuring before the V/D recombination. The D gene, `IGHD5-12`, was recombined with the 48 bp J gene `IGHJ4`. Considering only the sequence, one does not know here what was exactly this recombination. The nucleotides `TT`, marked with `**`, are perfectly aligned with both the end of the D gene and the start of the J gene: There are thus several interpretations on the number of deletions at the end of D and at the start of J; the D/J recombination can be viewed as 2//7, 3//6, or 4//5. There are probably not inserted nucleotides here.

An important question for the users and developers of RepSeq software is *how to assess the quality of their V(D)J mapping*. We present here a collection of test sequences coming from patient clones, covering all human TR and Ig locus, with manually curated V(D)J designations. These designations were checked by hand, possibly with the help of some bioinformatics tools. Tests may range from very easy cases with unambiguous V(D)J designations to borderline or difficult cases, including incomplete or unusual recombinations or translocations. This collection of sequences is still growing and distributed as open-source data. It can be viewed as a quality control of bioinformatics analysis and may help to benchmark or improve the robustness of any software doing RepSeq analysis. The following paragraphs detail the format to encode the designations (Section 2), present the collection of sequences and discuss the usages and the perspectives of this dataset (Section 3).

2 A format for V(D)J designations

A `.curated-vdj.fa` file is (almost) a Fasta file, containing one or several sequences with their V(D)J designation and possibly other information (Figure 2).

Encoding the genes and the N-regions

The Fasta header of each sequence, beginning by `>`, gives the V(D)J designation of the underlying sequence, such as in `IGKV1-5*03 9/CTAC/1 IGKJ1*01 [IGK]`.

```
>IGKV1-5*03 9/4/1 IGKJ1*01  [IGK]
TATTAATAACAACTTGGCCTGGTATCAGGAGAAGCCAGGGAAAGCCCCTAAGGTCCTG
ATCTATAAGGCGTCTAGTTTAGAAAGTGGGGTCCCATCAAGGTTCAGCGGCAGTGGAT
CTGGGACAGAATTCACTCTCACCATCAGCAGCCTGCAGCCTGATGATTTTGCAACCTA
TTACTGCCAACAATATAATAGACTTTGGACGTTCGGCCAAGGGACCAAGGTGGAAGTC
AAACGAACTGTGGCTGCACCATCT

# Patient 0122
>TRGV5*01 4/AG/5 TRGJP2*01  [TRG]   # 1st clone
GGAAGGCCCCACAGCGTCTTCTGTACTATGACGTCTCCAACTCAAAGGATGTGTTGGAA
TCAGGACTCAGTCCAGGAAAGTATTATACTCATACACCCAGGAGGTGGAGCTGGATATT
GATACTACGAAATCTAATTGAAAATGATTCTGGGGTCTATTACTGTGCCACCTGGGA
ag
AGTGATTGGATCAAGACGTTTGCAAAAGGGACTAGGCTCATAGTAACTTCGCCTGGTAA

>TRGV11 TRGJ1  [TRG]               # 2nd clone
TCTTCCACTTCCACTTtgAAAATAAAGTTCTTAGAGAAAGAAGATGAGGTGGTGTACC
aCTGTGCCTGCtaGTCACCTCATCGAATTATTATAAGA
```

Fig 2. Some reference designations. The designation can include either full gene name or alleles. The N-region can be either fully specified, or restricted to the number of deleted and inserted nucleotides or simply omitted.

- Gene names are taken as specified in IMGT/GENE-DB [17]. They can be either fully qualified, with their alleles (`IGKJ1*01`) or without (`IGKJ1`). When there is no allele given, the number of deletions for the N-regions should refer to the `*01` allele. Special names are also accepted, such as `Intron` or `KDE`.
- N-regions are given in three components: 1) number of nucleotides deleted at the right (3') of the left segment, 2) insertion, 3) number of nucleotides deleted at the left (5') of the right segment. The insertion can be specified either by the full sequence of nucleotides (`9/CCCTGG/1`), or by only the number of inserted nucleotides (`9/6/1`).
- N-regions are optional : `>IGKV1-5 IGKJ1` can be given instead of `>IGKV1-5 9/4/1 IGKJ1`

The designation can thus be very short, such as in `>TRGV2 TRGJP1`. However, it is advised to put as much information as possible in the designation, such as in `>TRGV2*01 9/CCCTGG/1 TRGJP1*01`. Such complete designations will give more extensive tests.

VDJ recombinations can be encoded, such as

- `>IGHV3-74*02 7/CCGCGGT/6 IGHD3-9*01 4/CTTCGAACA/7 IGHJ4*02`

Incomplete or unusual recombinations can also be specified, such as

- `>IGHD7-27*01 10/CATTA/0 IGHJ3*02`
- `>TRDV1*01 TRDD2*01 TRAJ29*01`
- `>TRDD2 18/6/0 TRDD3 5/5/0 TRDJ1`
- `>Intron 2/0/9 KDE`

Very special cases should be explained by comments in plain English. Those comments are introduced by a hash (#).

Encoding the locus

The end of the header may also contain information on the locus, between brackets, leading to additional tests. This also allows to specify only `>[TRG]` for a sequence that should be recognized as TRG even if it is difficult to choose a precise VJ designation.

Human locus should be encoded by `[IGH]`, `[IGK]`, `[IGL]`, `[TRA]`, `[TRB]`, `[TRG]`, `[TRD]`. Incomplete or unusual recombinations can be encoded with an additional `+` character, such as in `[IGH+]` or `[TRD+]`. Mixed TRα/TRδ recombinations can be encoded with `[TRA+D]`.

Other special cases, such as translocations involving BCL1 or BCL2, should be written now as comments after a `#` character.

Encoding the JUNCTION/CDR3 information

JUNCTION or CDR3 information can be optionally encoded, using curly braces:

```
>TRGV10*02 5/AGAC/3 TRGJP1*01  [TRG]  {CAAWRPTGWFKIF}
AAGTCCGTAGAGAAAGAAGACATGGCCGTTTACTACTGTGCTGCGTGGAGACCCACTGGT
TGGTTCAAGATATTTGCTGAAGGGACTAAGC
```

Encoding ambiguous or alternate designations

On some sequences, several V(D)J designations may be equally acceptable. These alternate choices can be encoded as `(choice1, choice2)`. For difficult cases, it is advised to further leave a comment in plain English:

```
# The D/J junction can be seen as 2//7, 3//6, or 4//5
>IGHV3-48*01 0/AA/6 IGHD5-12*01 (2//7, 3//6, 4//5) IGHJ4*02  [IGH]
ATGAACAGCCTGAGAGCCGAGGACACGGCTGTGTATTACTGTGCGAGAGAAAATAGTG
GCTACGATTTGACTAC
TGGGGCCAGGGAACCCTGGTCACCGTCTCCTCAGTT

# TRGJ1*01 or TRGJ1*02
>TRGV5*01 (TRGJ1*01, TRGJ1*02)  [TRG]
...
```

3 The dataset: availability and perspectives

The designations are available at `http://vidjil.org/curated-vdj/`. As the dataset will evolve and improve, there will be additional versions. Therefore a dataset should always be referred with its version number. Table 1 shows statistics on the annotations in the dataset.

This dataset was initially conceived as a test suite for the Vidjil algorithm. It includes some noteworthy sequences taken from the 125 sequences used in the evaluation shown in [25], comparing Vidjil with IgBlast and IMGT/V-QUEST designations. This is also the reason why we do not provide yet a comparison between several software: including

Locus/target	Sequences
IgH	105, including 22 Dh-Jh
Igκ	26, including 8 Intron-KDE
Igλ	1
TRα	22, including 22 mixed TRα/TRδ
TRβ	33, including 6 Dβ-Jβ
TRγ	18
TRδ	34, including 16 Dδ2/Dδ3

Table 1. Reference curated designations, version 2016.09. The 238 designations cover all the human loci. However some locus have few curated designations yet.

Vidjil would not have been fair since we can correct bugs with those sequences as soon as we detect them (and excluding it from the comparison would not have been fair too). The dataset was progressively extended by people used to manually checking V(D)J recombinations. We notably add alternate designations on existing sequences (currently 12 alternate designations). This dataset, with today 238 sequences, is still an ongoing work and it welcomes any quality contributions. We will also try to propose scripts to check the outputs of annotation software against the dataset.

The more sequences the dataset will contain, the more robust software will be as they will have hundreds or thousands of real-life tests to check their designations. We do not want to favour quantity over quality: It is important that the designations are of high quality so that this dataset could be considered as a gold standard. Some loci have very few manually curated designations as they are of little interest for clinical applications, or because they are more difficult to sequence. However we would like to have a high coverage of each human locus. We welcome any contribution that would help up improving that coverage.

The designations should be checked by at least two curators or software. And any ambiguity must be stated in the file either with the choice syntax or possibly in plain English. Sequences that were not checked independently by two curators are tagged with “TODO” (at the end of the sequence). There are currently 14 “TODO” tags in sequences that should be further checked.

As we wanted to keep the format simple, it has some limitations: we cannot encode yet hypermutations and we rely only on IMGT/GENE-DB annotations. The format could evolve in the future and we are open to any suggestion to overcome those limitations.

Finally it would be very helpful to crowdsource the V(D)J designation in order to create a massive dataset of carefully curated sequences. This could happen either through an application used by experts or even through a serious game.

Acknowledgments

We thank the reviewers for their helpful comments, as well as the EuroClonality-NGS consortium for productive discussions.

References

1. Tonegawa S. Somatic generation of antibody diversity. Nature. 1983;302(5909):575–581.

2. Market E, Papavasiliou FN. V(D)J recombination and the evolution of the adaptive immune system. PLoS Biology. 2003;1(1):E16.

3. Cavé H, van der Werff Ten Bosch J, Suciu S, Guidal C, Waterkeyn C, Otten J, et al. Clinical significance of minimal residual disease in childhood acute lymphoblastic leukemia. New England Journal of Medicine. 1998;339(9):591–598.

4. Benichou J, Ben-Hamo R, Louzoun Y, Efroni S. Rep-Seq: uncovering the immunological repertoire through next-generation sequencing. Immunology. 2012;135(3):183–91.

5. Yousfi Monod M, Giudicelli V, Chaume D, Lefranc MP. 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 Suppl 1:i379–85.

6. Brochet X, Lefranc MP, 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 Research. 2008;36(Web Server issue):W503–8.

7. Lefranc MP. IMGT, the International ImMunoGeneTics Information System. Cold Spring Harbor Protocols. 2011;2011(6):pdb–top112. doi:10.1101/pdb.top115.

8. Alamyar E, Giudicelli V, Li S, Duroux P, Lefranc MP. 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 Research. 2012;8(1).

9. Arnaout R, Lee W, Cahill P, Honan T, Sparrow T, Weiand M, et al. High-Resolution Description of Antibody Heavy-Chain Repertoires in Humans. PLoS ONE. 2011;6(8):e22365.

10. Thomas N, Heather J, Ndifon W, Shawe-Taylor J, Chain B. Decombinator: a tool for fast, efficient gene assignment in T-cell receptor sequences using a finite state machine. Bioinformatics. 2013;29(5):542–550.

11. Ye J, Ma N, Madden TL, Ostell JM. IgBLAST: an immunoglobulin variable domain sequence analysis tool. Nucleic Acids Research. 2013;41:W34–W40.

12. Kuchenbecker L, Nienen M, Hecht J, Neumann AU, Babel N, Reinert K, et al. IMSEQ – a fast and error aware approach to immunogenetic sequence analysis. Bioinformatics. 2015;31(18):btv309. doi:10.1093/bioinformatics/btv309.

13. Bolotin DA, Shugay M, Mamedov IZ, Ekaterina V Putintseva MAT, Zvyagin IV, Britanova OV, et al. MiTCR: software for T-cell receptor sequencing data analysis. Nature Methods. 2013;10:813–814. doi:10.1038/nmeth.2555.

14. Bolotin DA, Poslavsky S, Mitrophanov I, Shugay M, Mamedov IZ, Putintseva EV, et al. MiXCR: software for comprehensive adaptive immunity profiling. Nature Methods. 2015;12(5):380–381. doi:10.1038/nmeth.3364.

15. Yang X, Liu D, Lv N, Zhao F, Liu F, Zou J, et al. TCRklass: A New K-String-Based Algorithm for Human and Mouse TCR Repertoire Characterization. Journal of Immunology. 2014;194(1). doi:10.4049/jimmunol.1400711.

16. Giraud M, Salson M, Duez M, Villenet C, Quief S, Caillault A, et al. Fast multiclonal clusterization of V(D)J recombinations from high-throughput sequencing. BMC Genomics. 2014;15(1):409. doi:10.1186/1471-2164-15-409.

17. Giudicelli V, Chaume D, Lefranc MP. IMGT/GENE-DB: a comprehensive database for human and mouse immunoglobulin and T cell receptor genes. Nucleic Acids Research. 2005;33(S1):D256–D261.

18. Bystry V, Darzentas N, al. ARReST – Antigen Receptors Research Tool;. `http://tools.bat.infspire.org/arrest`.

19. Moorhouse MJ, van Zessen D, IJspeert H, Hiltemann S, Horsman S, van der Spek PJ, et al. ImmunoGlobulin galaxy (IGGalaxy) for simple determination and quantitation of immunoglobulin heavy chain rearrangements from NGS. BMC Immunology. 2014;15(1):1.

20. Schaller S, Weinberger J, Jimenez-Heredia R, Danzer M, Oberbauer R, Gabriel C, et al. ImmunExplorer (IMEX): a software framework for diversity and clonality analyses of immunoglobulins and T cell receptors on the basis of IMGT/HighV-QUEST preprocessed NGS data. BMC Bioinformatics. 2015;16(1):252. doi:10.1186/s12859-015-0687-9.

21. Nazarov VI, Pogorelyy MV, Komech EA, Zvyagin IV, Bolotin DA, Shugay M, et al. tcR: an R package for T cell receptor repertoire advanced data analysis. BMC bioinformatics. 2015;16(1):175.

22. Dmitry B, al. VDJviz;. `http://vdjviz.milaboratory.com`.

23. Duez M, Giraud M, Herbert R, Rocher T, Salson M, Thonier F. Vidjil: High-throughput analysis of immune repertoire. submitted;.

24. Damle RN, Wasil T, Fais F, Ghiotto F, Valetto A, Allen SL, et al. Ig V gene mutation status and CD38 expression as novel prognostic indicators in chronic lymphocytic leukemia. Blood. 1999;94(6):1840–1847.

25. Ferret Y, Caillault A, Sebda S, Duez M, Grardel N, Duployez N, et al. Multi-loci Diagnosis of Acute Lymphoblastic Leukemia with High-Throughput Sequencing and Bioinformatics Analysis. British Journal of Haematology. 2016;173(3):413–420. doi:10.1111/bjh.13981.