Yclon: Ultrafast clustering of B cell clones from high-throughput immunoglobulin repertoire sequencing data

João Gervásio, Alice Ferreira, Liza F. Felicori

Laboratory of Synthetic Biology and Biomimetics, Departamento de Bioquímica e Imunologia, Universidade Federal de Minas Gerais, Belo Horizonte, Minas Gerais, 31270-901, Brazil.

Abstract

Motivation: Next-generation sequencing technologies revolutionize our understanding of immunoglobulin (Ig) profile in different immune states. Clonotyping (grouping Ig sequences into B cell clones) allows the investigation of the diversity of repertoires and how they change upon antigen exposure. Despite its importance, there is no consensus on the best method for clonotyping, and the methods developed for that are computationally intractable for large sequencing datasets. This is the case of Change-O, which compares sequences through hamming distance and uses hierarchical clustering for this task.

Results: We propose implementing an approach to identify B cell clones from Ig repertoire data, named Yclon, that makes an alignment-free comparison of the sequences, focusing on reducing the runtime and computer memory usage. Overall, we find that a hierarchical clustering approach grouped Ig sequences into B cell clones similarly to Change-O. However, we observed that Yclon was around 35 times faster and even was able to process larger than 2 million sequences Ig repertoire, which is a critical part of repertoire studies and enables understanding antibody repertoire structure and affinity maturation.

Availability and implementation: YClon is written in Python3 and is available on GitHub (https://github.com/jao321/YClon.git)

Contact: joaodgervasio@gmail.com, liza@icb.ufmg.br

1. Introduction

Using "next-generation sequencing" in studies of Adaptive Immune Receptor Repertoire Sequencing (AIRR-Seq) for antibody profiling enables us to produce only snapshots of the repertoire diversity. Although humans can produce as many as 10^{18} distinct antibodies, those advanced techniques help to better understand the immune system (Briney *et al.*, 2019). Besides that, incredible advances in the field make it possible for scientists to assemble as many as millions to billions of distinct antibody sequences (DeWitt *et al.*, 2016; Briney *et al.*, 2019).

However, identifying antibody sequences that belong to the same B-cell clone in these data remains a challenge (Gupta et al., 2018). A B-cell clone or clonotype is defined as a group of antibody sequences derived from a common progenitor B cell (Hershberg et al., 2015).

Clone size reflects on repertoire diversity, which is related to age (Davydov *et al.*, 2018), response to vaccination or infection (Khavrutskii *et al.*, 2017), allergy (Wu *et al.*, 2014), cancer (Zhang *et al.*, 2019), among several other conditions. It is also worth mentioning that antibodies from the same clonotype often bind to the same epitope.

Even though B-cell clone identification is crucial, there are no consensus methods to do it. Some approaches consider the same clone as antibody sequences sharing the same V and J genes while having likeness in the third complementary-determining region (CDR3), both in length and similarity, with thresholds normally ranging from 80% to 100%, which can be achieved using hierarchical clusterization (Teng and Papavasiliou, 2007; Elhanati *et al.*, 2015; Hershberg and Luning Prak, 2015; Gupta *et al.*, 2017). In general, there is a consensus that the V and J genes have to be shared by antibodies to consider them as part of the same clonotype. The main difference in the clonotyping methods is the way and the cut-offs used to group the antibodies CDR3. Some approaches like the one containing in Change-O and SCOPer assigns clones based on the hamming distance of CDR3 or junction, then performs a hierarchical clustering (Gupta *et al.*, 2015, Nouri *et* al., 2018) among other strategies (Chen *et al.*, 2010; Gupta *et al.*, 2015; Nouri and Kleinstein, 2020; Aouinti *et al.*, 2015). Others, like IMGT, consider that the entirety of the junction between the VD and DJ genes to clusterize B-cell clones should be equal. The problem with most of the above approaches is the time to process high-throughput data or even the unavailability to process more than 1 million sequences such as IMGT. Lindenbaum *et* al., 2020 [15] proposed a method to group sequences into clonotypes, using an alignment-free method, vectorizing a sequence of 150 nucleotides starting from the latter part of the V gene to the 3' end into k-mers of size 4, which should be fast, but they do not account if the sequences are annotated as the sharing the same V and J genes. Therefore the clones assigned by then do not fit the definition of clone hereby established.

As previously mentioned, it is of absolute necessity to have a method capable of cluster B-cell clones optimizing memory usage to deal with larger repertoires. The two most relevant repertoire databases as of today are OAS and iReceptor gathering 1,471,364,639(organized in 12,042 distinct files) and 803,662,891(organized in 1,308 distinct files) human IGH sequences, respectively, on January 17th, 2022 (Corrie *et* al., 2018; Kovaltsuk *et* al., 2018). For those repertoires that have more than a million sequences, OAS has 359 repertoires, and iReceptor has

206. This shows the need for a tool that clonotypes quickly and efficiently while respecting the definition of clone.
Because of this, we propose YClon, a tool that rapidly processes repertoires with more than 2 million sequences in less than one hour, considering V and J genes and similarity in the CDR3.

2. Material and Methods

2.1 Yclon Implementation

YClon is a script written in Python3 and receives as input, on its default setting, ".tsv" files formatted in the MiAIRR data standard. The file must contain at least four columns, and they should be assigned as "sequence_id" with a unique identifier for the sequence, "v_call" with the V gene annotation, "j_call" with the J gene annotation, and "cdr3" containing the sequence of the CDR3. In order to clonotype the sequences YClon goes through 5 steps:

I. Subsetting: The first step is to subset the data sharing the same V and J gene segments and CDR3 length.
II. Vectorizing CDR3 sequence: CDR3 sequences of the data subsetted were deconstructed in n-grams of size 3 with a sliding window of 1 nucleotide.
III. Comparing vectors: Afterwards, a similarity matrix is built by comparing every vector, representing the sequences, using cosine similarity (Sidorov *et al.*, 2014).
IV. Clustering: The resulting square distance matrix was the input for a hierarchical agglomerative method. It uses a scikit-learn (sklearn) library, with a default threshold distance of 0.09 (Pedregosa *et al.*, 2011).
V. Writing output file: The last step consists of writing an output file containing all assigned clone sequences into the "clone_id" column. If the file already has a "clone_id" column, the new column is called "clone_id_yclon". The output file will be named by adding "Yclon_clonotyped" to the input file name.

To evaluate the tool's functionality, we ran YClon with 3 antibody repertoire datasets of different sizes (Table 1). The smaller repertoire (Ghraichy *et al.*, 2020) was re-annotated using IMGT/HighV-Quest (Aouinti *et al.*, 2015), while the two others were downloaded already annotated (Chang *et al.*, 2016; Briney *et al.*, 2019) from iReceptor (Corrie et al. 2018) and OAS respectively. We also clonotyped the same repertoires using Change-O and SCOPer to compare the results with YClon.

2.2 CDR3 intraclonal similarity

In order to assess intraclonal CDR3 similarity of the clones generated by Yclon, Change-O, and SCOPer we aligned the CDR3 sequences of the clones with every other sequence in that clone, checking their similarity. We did that for every clone in each repertoire for all the tools tested. We did so by comparing the CDR3 sequences inside each clone using the python library

“distance”, treating the sequences as if they were aligned since all of them have the same length, and used the function “hamming()” to compute the proportion of sites that differ between each pair of sequences. We calculated the average sequence similarity inside each clone for each dataset and method.

2.3 Specificity, sensitivity and PPV

We also tested the tool against grounded true data using simulated repertoire. To do so, we ran YClon using 25 simulated datasets from Nouri *et* al., 2020, and other 74 simulated repertoires from Lindenbaum *et* al., 2020.

Then it was possible to calculate specificity by dividing the number of sequences that were assigned as not being in a clone and truly not being in it (TN) by the sum of TN with the number of sequences assigned as being in that clone but in fact not being in it (FP).

The calculation of sensitivity was conducted by dividing the number of sequences assigned to a clone and truly being in it (TP) by the sum of TP with the number of sequences assigned to a different clone, but actually being part of the analyzed clone (FN).

Finally, to calculate the positive predictive value (PPV) we divided TP by the sum of TP with FP, adjusting the weight of the PPV in the whole repertoire.

The specificity, sensitivity and PPV values for each clone were summed and then divided by the number of total clones.

$$Specificity_i = TN/(TN + FP)$$

$$Specificity_{wholerepertoire} = \frac{\sum_{i=1}^{numberofclones} specificity_i}{TotalNumberOfClones}$$

$$Sensitivity_i = [TP/(TP + FN)]$$

$$Sensitivity_{wholerepertoire} = \frac{\sum_{i=1}^{numberofclones} sensitivity_i}{TotalNumberOfClones}$$

$$PPV_i = TP/(TP + FP)$$

$$PPV_{wholerepertoire} = \frac{\sum_{i=1}^{numberofclones} PPV_i}{TotalNumberOfClones}$$

2.4 Performance benchmark

To have an understanding on how faster our approach is, when compared to Change-O and SCOPer we clonotyped the three real data repertoires (Table 1) with those tools in a Desktop with an Intel(R) Core(TM) i5-4590 CPU @ 3.30GHz, with 16 Gb of RAM, running ubuntu 20.04 and compared the time each software took to complete the task.

We made a speed test by resampling the largest simulated repertoire from the largest repertoire (coming from the 25 simulated repertoires), which had a little more than 870,000 sequences. We tested the speed by increasing every resample by 10,000 sequences.

3. **Results and Discussion**

The efficacy of YClon was tested using two different datasets of simulated repertoire. We observed that in a similar way to Change-O and SCOPer, YClon is highly specific (around 100% for both sets of simulated repertoire) and even presents a higher PPV when compared to SCOPer. While YClon achieved PPV values between 99.4% and 100%, SCOPer was between 94% and 98.7%, and Change-O was between 99.4% and 99.9% (Figure 1).

This indicates that most sequences assigned together are indeed part of the same clones. There are differences in the sensitivity values depending on the analyzed simulated dataset. Those 74 simulated antibodies repertoires from Lindenbaum *et* al., 2020 showed high sensitivity, more than 97%, whereas those 25 from Nouri *et* al.,2020 ranged from 72% to 92% (Figure 1) when using Yclon. The difference between the repertoires with distinct sensitivity values was that those with lower sensitivity had a higher number of large clones. Also, the other two methods struggled to be consistent with these different simulated repertoires.

When comparing the similarity of the CDR3 sequences, via alignment, within the clones, YClon grouped sequences with more than 90% similarity in a similar way that Change-O and SCOPer did (Figure 2).

YClon was faster than Change-O and SCOPer when clonotyping the 3 benchmark repertoires. YClon was 130 times faster than Change-O, and 53 times faster than SCOPer to clonotype the small repertoire, 77 times faster than Change-O and 262 times faster than SCOPer for the medium repertoire. The computer used to run those tests could not run the larger repertoire, containing more than 2 million sequences, with neither Change-O nor SCOPer, whereas it took 15 minutes and 44 seconds with YClon (Table 2).

Lastly, the speed test with resampled data (Figure 3) shows that YClon time to handle larger repertoires increases in an near to linear manner.

4. **Conclusion**

In the latter years, the need to analyze large repertoire datasets with more than 1 million sequences has been increasing. iReceptor Gateway integrates more than 1,433 antibody repertoires, 206 of them with more than 1 million sequences, and OAS has more than 359 of those. YClon is a tool that makes clonal assignment faster, reliable, and possible in an ordinary computer since this step is a crucial process in understanding antibody repertoire data.

Conflict of Interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Fundings
Coordenação de Aperfeiçoamento de Pessoal de Nível Superior – Brazil (CAPES) [grant numbers 88887.506611/2020-00, 88887.504420/2020-00 and 935/19 (COFECUB)]; Fundação de Amparo a Pesquisa de Minas Gerais (FAPEMIG) [grant numbers PPM-00615-18, Rede Mineira de Imunobiologicos grant #REDE-00140-16]; Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq) [Pq to LFF]; National Institutes of Health (NIH) [grant number 1R01AI143552-02]; Pro-Reitoria de Pesquisa da Universidade Federal de Minas Gerais

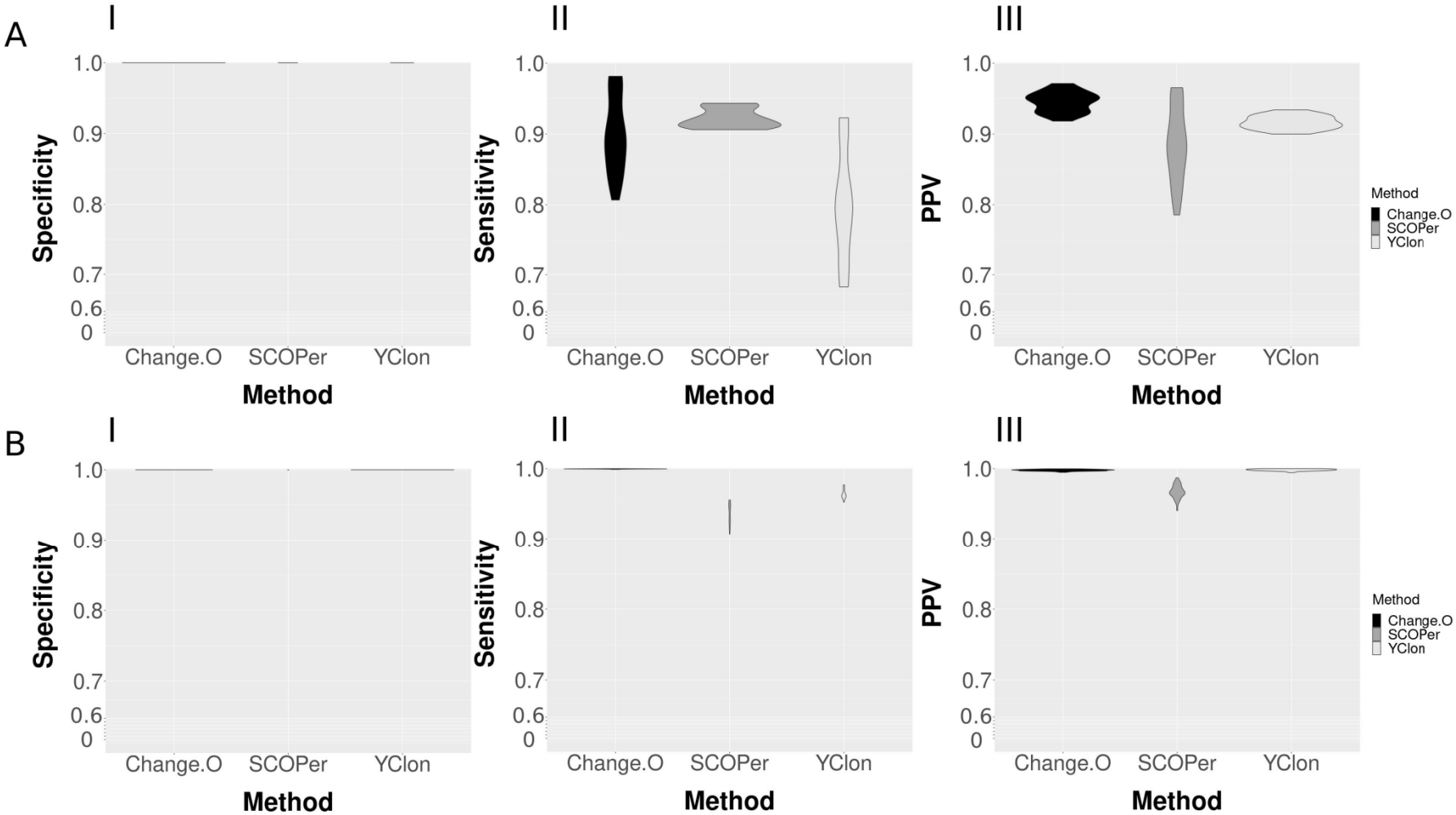

Figure 1: Simulated repertoire clone identification metrics. Violin plot of the values for specificity (I), sensitivity (II), and PPV (III) in clone identification regarding the 25 simulated repertoires from Nouri *et* al., 2020 (A) or the 74 simulated repertoires from Lindebaum *et* al., 2020 (B). Values of specificity (I) indicate the capacity of identifying sequences that are not part of a clone. Values of sensitivity (II) indicate the capacity of identifying sequences that are part of a clone. PPV values (III) indicate the probability that sequences in a clone are indeed part of it. This analysis resulted from 3 different clonotyping methods: Yclon, here developed, Change-O and SCOPer.

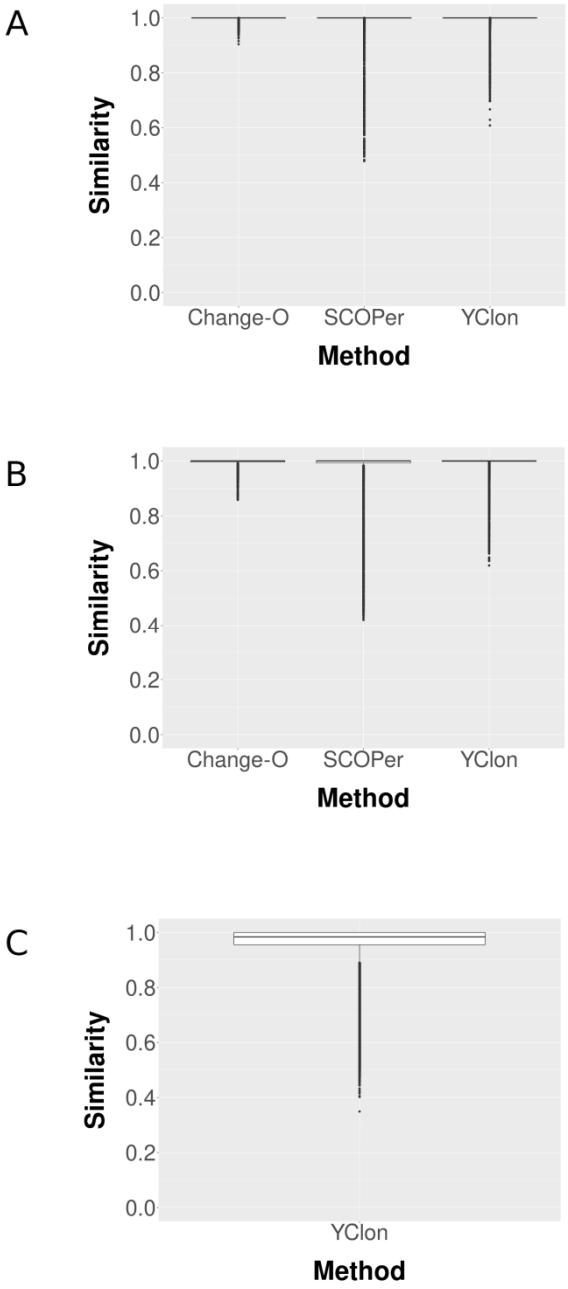

Figure 2: Sequence similarity inside each clone using three different size repertoires. Boxes encompass the 25th-100th percentiles of all data; whiskers enclose the 25th percentiles, and circles represent all outliers. If not counting the outliers, every method gathers sequences in a very similar way inside the clones. This analysis resulted from 3 different clonotyping methods: Yclon, here developed, Change-O and SCOPer.

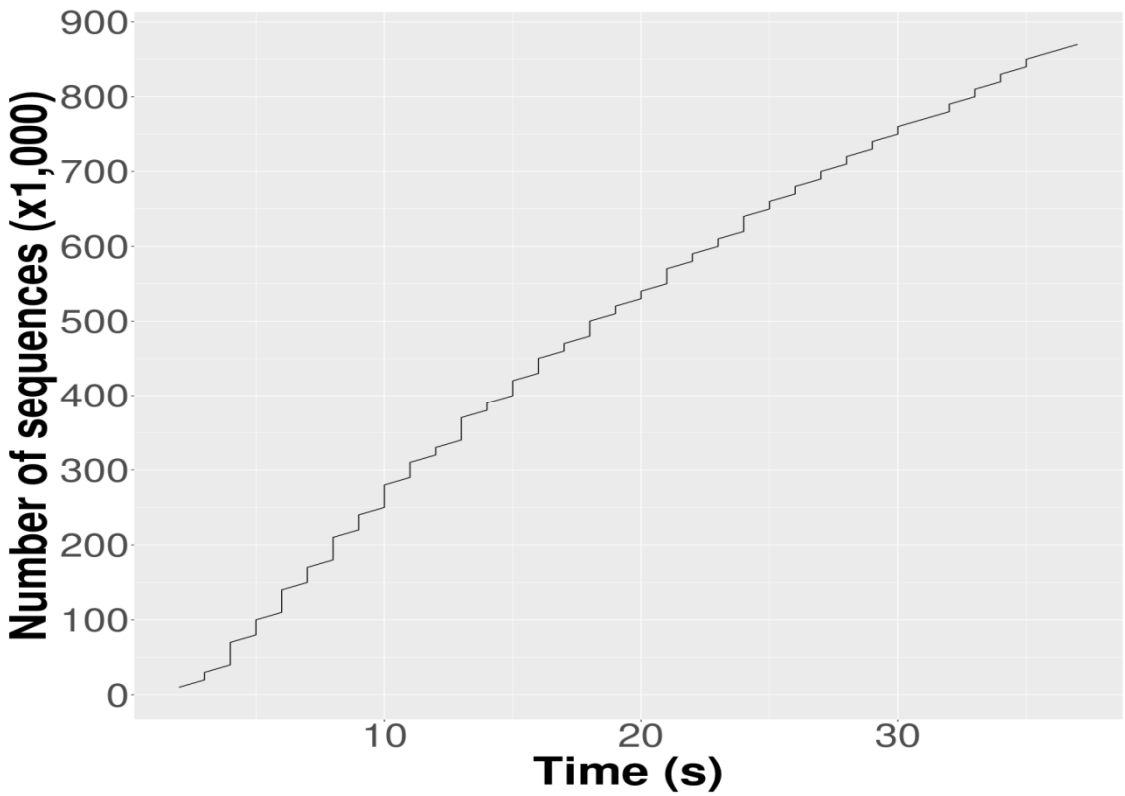

Figure 3: Speed test. The speed test was made by resampling a large repertoire, increasing the resulting simulated repertoire in 10,000 sequences every round. The time it takes to process the repertoires increase

almost linearly, with a caviat that in some instances it takes the same time for repertoires with different sizes.

Table 1: Comparison between Yclon, SCOPer and Change-O regarding the number of clones, the most abundant ones and the time of execution

	Dataset 1			Dataset 2			Dataset 3
	YClon	SCOPer	Change-O	YClon	SCOPer	Change-O	YClon
Number of sequences	**82,928**			**365,707**			**2,081,722**
Total number of clones	24,127	22,590	24,031	73,741	23,236	72,446	1,610,373
Clones with more than 1 sequence	15,833	14,477	15,648	35,537	15,095	35,474	203,766
Benchmark time	7s	6min13s	15m12s	50s	4h51m	1h11m 20s	15min44s
Number of clones that represents the most abundant 5% sequences	35	23	33	2	2	2	946
Number of clones that represents the most abundant 15% sequences	376	162	327	22	5	19	32,836
Number of clones that represents the most abundant 25% sequences	1,180	719	1,190	305	18	276	126,468

Supplementary Table S1 - Details of the datasets used in this study

	Dataset 1	**Dataset 2**	**Dataset3**
Number of sequences	82,966	365,370	2,164,496
Database cointaing Sequences	OAS	iReceptor	OAS
Study ID in iReceptor	-	PRJNA628125	SRR8283827
Subject	C.037.1	7450	316188
Sample	-	M369-S001	
Study	Ghraichy *et* al., 2020	Nielsen *et* al., 2020	Briney *et* al., 2019
Sex	Male	Female	-
Age	4	73	30
Condition	Healthy	Covid-19	Healthy

References

1. Briney,B. *et al.* (2019) Commonality despite exceptional diversity in the baseline human antibody repertoire. *Nature*, **566**, 393–397.
2. DeWitt,W.S. *et al.* (2016) A Public Database of Memory and Naive B-Cell Receptor Sequences. *PLOS ONE*, **11**, e0160853.
3. Gupta,N.T. *et al.* (2017) Hierarchical Clustering Can Identify B Cell Clones with High Confidence in Ig Repertoire Sequencing Data. *J. Immunol.*, **198**, 2489–2499.
4. Hershberg,U. and Luning Prak,E.T. (2015) The analysis of clonal expansions in normal and autoimmune B cell repertoires. *Philos. Trans. R. Soc. B Biol. Sci.*, **370**, 20140239.
5. Davydov,A.N. *et al.* (2018) Comparative Analysis of B-Cell Receptor Repertoires Induced by Live Yellow Fever Vaccine in Young and Middle-Age Donors. *Front. Immunol.*, **9**.
6. Khavrutskii,I.V. *et al.* (2017) Quantitative Analysis of Repertoire-Scale Immunoglobulin Properties in Vaccine-Induced B-Cell Responses. *Front. Immunol.*, **8**.
7. Wu,Y.-C.B. *et al.* (2014) Influence of seasonal exposure to grass pollen on local and peripheral blood IgE repertoires in patients with allergic rhinitis. *J. Allergy Clin. Immunol.*, **134**, 604–612.
8. Zhang,J. *et al.* (2019) Immune receptor repertoires in pediatric and adult acute myeloid leukemia. *Genome Med.*, **11**, 73.
9. Teng,G. and Papavasiliou,F.N. (2007) Immunoglobulin Somatic Hypermutation. *Annu. Rev. Genet.*, **41**, 107–120.
10. Elhanati,Y. *et al.* (2015) Inferring processes underlying B-cell repertoire diversity. *Philos. Trans. R. Soc. B Biol. Sci.*, **370**, 20140243.
11. Gupta,N.T. *et al.* (2017) Hierarchical Clustering Can Identify B Cell Clones with High Confidence in Ig Repertoire Sequencing Data. *J. Immunol.*, **198**, 2489–2499.
12. de Wildt,R.M. *et al.* (1999) Somatic insertions and deletions shape the human antibody repertoire. *J. Mol. Biol.*, **294**, 701–710.
13. Gupta,N.T. *et al.* (2015) Change-O: a toolkit for analyzing large-scale B cell immunoglobulin repertoire sequencing data. *Bioinformatics*, **31**, 3356–3358.
14. Nouri N, Kleinstein S (2018). “A spectral clustering-based method for identifying clones from high-throughput B cell repertoire sequencing data.” Bioinformatics, i341-i349.
15. Lindenbaum, O. *et* al. (2021). Alignment free identification of clones in B cell receptor repertoires. *NAR*, **49**, (4),, e21,
16. Corrie , B.D. *et* al. (2018) iReceptor: A platform for querying and analyzing antibody/B-cell and T-cell receptor repertoire data across federated repositories. *Immunol Rev*.,248,24-41.
17. Kovaltsuk, A. *et* al. (2018) Observed Antibody Space: A Resource for Data Mining Next-Generation Sequencing of Antibody Repertoires. *J. Immunol.*, **201** (8), 2502-2509
18. Chen,Z. *et al.* (2010) Clustering-based identification of clonally-related immunoglobulin gene sequence sets. *Immunome Res.*, **6**, S4.
19. Nouri,N. and Kleinstein,S.H. (2020) Somatic hypermutation analysis for improved identification of B cell clonal families from next-generation sequencing data. *PLOS Comput. Biol.*, **16**, e1007977.
20. Aouinti,S. *et al.* (2015) 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*, **10**, e0142353.

21. Beel,J. *et al.* (2016) Research-paper recommender systems: a literature survey. *Int. J. Digit. Libr.*, **17**, 305–338.
22. Cong,Y. *et al.* (2016) A novel alignment-free method for detection of lateral genetic transfer based on TF-IDF. *Sci. Rep.*, **6**, 30308.
23. Sidorov,G. *et al.* (2014) Soft Similarity and Soft Cosine Measure: Similarity of Features in Vector Space Model. *Comput. Sist.*, **18**.
24. Campello,R.J.G.B. *et al.* (2013) Density-Based Clustering Based on Hierarchical Density Estimates. In, Pei,J. *et al.* (eds), *Advances in Knowledge Discovery and Data Mining*, Lecture Notes in Computer Science. Springer, Berlin, Heidelberg, pp. 160–172.
25. Pedregosa,F. *et al.* (2011) Scikit-learn: Machine Learning in Python. *J. Mach. Learn. Res.*, **12**, 2825–2830.
26. Ghraichy,M. *et al.* (2020) Maturation of the Human Immunoglobulin Heavy Chain Repertoire With Age. *Front. Immunol.*, **11**.
27. Chang,Y.-H. *et al.* (2016) Network Signatures of IgG Immune Repertoires in Hepatitis B Associated Chronic Infection and Vaccination Responses. *Sci. Rep.*, **6**, 26556.
28. Nielsen,S.C.A. *et al.* (2020) Human B Cell Clonal Expansion and Convergent Antibody Responses to SARS-CoV-2. *Cell Host Microbe*, **28**, 516-525.e5.