ICING: Large-scale Inference of Immunoglobulin Clonotypes

Federico Tomasi[1⋆][0000−0002−8718−3844],
Margherita Squillario[1][0000−0002−6612−3383],
Alessandro Verri[1][0000−0001−9777−9986],
Davide Bagnara[23⋆⋆][0000−0001−7889−8103], and
Annalisa Barla[1⋆⋆][0000−0002−3436−035X]

[1] Department of Informatics, Bioengineering, Robotics and System Engineering (DIBRIS), Università degli studi di Genova, Genoa, I-16146, Italy
[2] Department of Experimental Medicine (DIMES), Università degli studi di Genova, Genoa, I-16132, Italy
[3] The Feinstein Institute for Medical Research, North Shore-LIJ Health System, 350 Community Drive, Manhasset, NY 11030, USA

Abstract. Immunoglobulin (IG) clonotype identification is a fundamental open question in modern immunology. An accurate description of the IG repertoire is crucial to understand the variety within the immune system of an individual, potentially shedding light on the pathogenetic process. Intrinsic IG heterogeneity makes clonotype inference an extremely challenging task, both from a computational and a biological point of view. Here we present ICING, a framework that allows to reconstruct clonal families also in case of highly mutated sequences. ICING has a modular structure, and it is designed to be used with large next generation sequencing (NGS) datasets, a technology which allows the characterisation of large-scale IG repertoires. We extensively validated the framework with clustering performance metrics on the results in a simulated case. ICING is implemented in Python, and it is publicly available under FreeBSD licence at `https://github.com/slipguru/icing`.

Keywords: Clonotype identification · Immunoglobulin · NGS data · Cluster analysis

1 Scientific Background

The identification of immunoglobulin (IG) clonotypes is a key question in modern immunology. A clonotype is a particular combination of IGs generated by a single plasma cell clone, which is a population of cells all derived from a single progenitor cell (germline). The ability to infer clonotypes is crucial as it allows to understand how much diversity an individual has in its immune repertoire and to

⋆ Corresponding author: `federico.tomasi@dibris.unige.it`
⋆⋆ These authors contributed equally to this work.

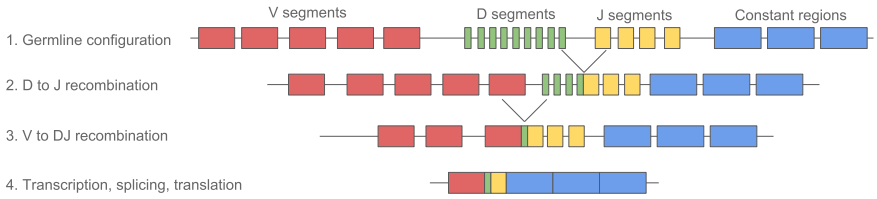

Fig. 1. IG recombination. Starting from V(D)J gene segments, one of each type is selected to produce the IG sequence. When joining two segments, some insertions and deletions (*indels*) may occur. A constant region is appended to the IG sequence after the recombination.

study immune response through B-cell clonal amplification and diversification. Indeed, understanding the variety within the immune system of an individual may potentially shed light on pathogenetic processes. In healthy individuals the repertoire is expected to be extremely diverse, to guarantee the ability to respond to a wide range of antigens (*e.g.* bacteria, viruses). The diversity of the B-cell repertoire is due to the gene recombination process, where, by random selection, one for each V, D and J genes are joined together, with a simultaneous trimming and addition of random nucleotides (Figure 1). The resulting bridging segment between V and J genes, called complementarity determining region 3 (CDR3), is the most variable and therefore important for the antigen binding [11]. Before encountering an antigen, B-cells have zero (or few) somatic mutations. Without considering mutations, the overall repertoire diversity usually comprises 10^7 to 10^8 clonotypes, with lower bounds of diversity of 10^5 and potentially as high as 10^{11} unique molecules in a single individual [4]. After the immune response, they undergo clonal amplification and somatic hypermutation, to increase the binding affinity to the antigen [8]. The potential frequency of somatic hypermutation, which can be at least 10^5-10^6 fold greater than the normal rate of mutation across the genome [9], may generate many orders of magnitude more diversity in the B-cell receptor repertoire than the 10^{11} unique molecules per individual. Therefore, intrinsic data heterogeneity makes IG clonotyping an extremely difficult task.

2 ICING

To tackle the problem of IG clonotyping inference, we developed ICING (Inferring Clonotypes of ImmuNoGlobulins), a Python library publicly available at `https://github.com/slipguru/icing`. The method aims at grouping IGs into clonal families, whose members derive from the same germline ancestor. Input and output data have the same format used by the Change-O suite, hence ICING is easily integrable in the usual pRESTO/Change-O pipeline [14, 5]. In particular, data should be in the format produced by Change-O, that is, IGs should be represented via their V gene calls and CDR3 amminoacidic (or nucleotidic) sequence. Also, an indication of the mutation level of the sequence

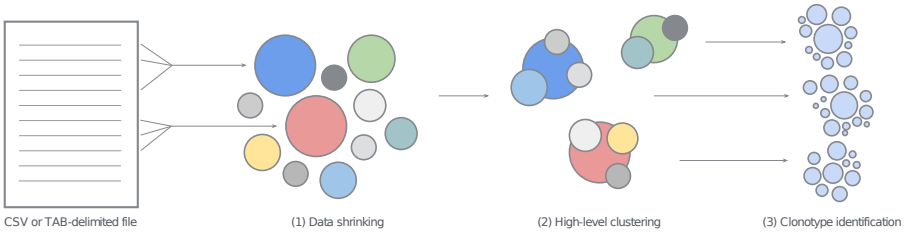

Fig. 2. ICING pipeline. Starting from a CSV or TAB-delimited file, the first step consists in grouping together sequences based on their V gene calls and CDR3 identity (data shrinking step). An high-level clustering is done on CDR3 lengths to reduce the computational workload of the third and final phase, which involves a clustering step on each of the previously found groups to obtain fine-grained IG clonotypes.

with respect to reference should be present, to allow for the final steps of the pipeline (Section 3.3).

ICING is designed to be used with a large number of data, for example coming from NGS technologies, composed of more than 10^6 sequences. The method is implemented in Python, exploiting separate processes on multi-core machines for almost each step of three sequential phases: *(i)* data shrinking, *(ii)* high-level grouping and *(iii)* fine-grained clonotype identification (Figure 2).

3 Materials and Methods

3.1 Synthetic Data Generation

We used *partis* [10] to generate synthetic datasets, which are characterised by an increasing number of IGs and clones, 0.05 frequency of insertions and deletions (*indels*) of maximum length of 6 nucleotides on the CDR3 sequence, and different degrees of V gene sequence mutation level. Table 1 presents an overview of the datasets.

3.2 Preprocessing

The datasets were submitted to IMGT/HighV-QUEST [1] for V(D)J genes inference, then preprocessed by a Change-O feature [5]. The outcome is a single TAB-delimited file containing the information about IGs and their metadata, such as the identification of V(D)J sequences (*i.e.*, V(D)J gene calls), V gene sequence mutation level and identification of CDR3 sequence, to be used as input to the pipeline.

3.3 Clonotype Identification

The clonotype identification step is divided into three parts.

Table 1. Datasets overview. For reference, the total number of functional gene segments for the V/D/J regions of heavy chains in the human genome are 65/27/6 [7].

dataset	sequences	clonotypes	avg seqs/clone	unique V genes	unique D genes	unique J genes	mean (std) of V gene mutation
D1	9233	77	92.35	35	24	6	9.59 (4.64)
D2	17825	74	185.09	38	24	6	8.64 (4.46)
D3	37897	77	396.43	34	25	6	9.04 (4.51)
D4	47764	389	99.08	56	25	6	8.63 (4.30)
D5	102336	388	209.44	58	25	6	8.41 (4.70)
D6	205986	379	428.44	56	25	6	9.56 (4.46)
D7	162713	1168	109.66	58	25	6	8.72 (4.67)
D8	301978	1180	206.22	58	25	6	9.15 (4.73)
D9	589680	1185	400.26	58	25	6	8.94 (4.65)
D10	291076	2282	96.29	58	25	6	8.84 (4.46)
D11	568799	2317	187.76	58	25	6	9.12 (4.76)
D12	1208110	2358	404.30	58	25	6	9.11 (4.77)

Data Shrinking. Input data are grouped based on V gene calls (exact correspondence) and CDR3 identity (completely overlapping sequence). This allows to reduce the computational workload of next clustering steps. To each group is assigned a weight, equal to the cardinality of the group.

High-level Group Inference. This phase involves a clustering step on CDR3 lengths of previously identified groups. The outcome, which consists of high-level groups of IGs to be refined afterwards, contains IG sequences having comparable CDR3 lengths. This is done using MiniBatchKMeans clustering algorithm [12], which is computationally efficient and, more importantly, may group together very similar clusters.

Fine-grained Group Inference. Each high-level group extracted before is then subdivided based on the actual IG distance. The distance between IGs is computed taking into account V gene calls and CDR3 sequences. In particular, the distance between two IGs is lower than infinity if and only if they have at least one V gene call in common. In such case, their actual distance is computed using a sequence distance method on their CDR3 sequences. In particular, the method implements a generic normalised distance measure based on a particular model matrix $\mathcal{M}$. Let $\|\mathcal{M}\|_{\max} = \max_{i,j} |\mathcal{M}_{ij}|$. For two sequences s and t of equal length ℓ, we defined their distance $\mathcal{D}(s,t)$ as follows:

$$\mathcal{D}(s,t) = \frac{1}{\ell \cdot \|\mathcal{M}\|_{\max}} \sum_{i=1}^{\ell} \mathcal{M}(s^i, t^i). \tag{1}$$

The choice of a specific model depends on the type of data under analysis. When $\mathcal{M} = \mathcal{H}$, where $\mathcal{H}(x,y) = 0$ if $x = y$ and 1 otherwise, the model assumes the form of a normalised Hamming distance [6].

Such distance measure allows seamless integration of different nucleotidic and amminoacidic models. ICING includes Hamming and its weighted variants,

such as HS1F [16]. The models are defined between sequences of equal length. The method allows also the comparison of sequences with different lengths, by tuning a *tolerance* parameter. In such case, a standard alignment step between two sequences of different lengths may be performed before the computation of their distance, using the Smith-Waterman algorithm for sequence alignment [13].

IG sequences are characterised by an high level of mutation. Therefore, a correction function based on V gene sequence mutation level may be used to reduce distances between two IGs if mutated. This procedure encodes the uncertainty of the distance measure when dealing with highly mutated data, allowing for a more robust measure. We note that this is a step which is strongly depends on the data at hand. In our experiments, we corrected the distances between two IGs by multiplying $\mathcal{D}(s,t)$ with ν_{st}, where $\nu_{st} = 1 - \frac{m_s - m_t}{2}$, with m_s and m_t are the mutation levels of the sequences s and t, respectively.

After the design of such distance metric, fine-grained groups (*i.e.*, final clonotypes) are extracted using the DBSCAN clustering algorithm [2], which only require the parameter ϵ for the neighbourhood search of spatial distances. On top of an appropriate index structure, the algorithm can run in $O(n \log n)$ and it only needs linear memory, allowing the analysis of large-scale data.

3.4 Performance Assessment

For synthetic datasets the information about IG clonotypes is known, and it is used as ground truth. In order to evaluate clustering performance of the method, we used standard metrics such as homogeneity (HOM), completeness (COM) and V-measure (VSC), mutual information based scores, namely Adjusted Mutual Information (AMI) and Normalized Mutual Information (NMI), Adjusted Rand Index (ARI), and Fowlkes-Mallows score (FMI) [3, 15]. Such measures are bound by $[0, 1]$, and no assumption is made on the cluster structure. Moreover, AMI, ARI and FMI are adjusted against chance, which is an important feature when evaluating a clustering performance in presence of a large number of clusters. Therefore, random (uniform) label assignments have scores close to 0 for measures normalised against chance.

3.5 Computing Architecture

Experiments were performed using a computing machine equipped with two Intel® Xeon® CPUs E5-2630 v3 (2.4 GHz, 8 cores each) and 128 GB of RAM[4].

4 Results

4.1 Performance Evaluation

We evaluated the method performance on the datasets shown in Table 1. In particular, Table 2 shows the clustering scores (Section 3.4) for datasets D1–3,

[4] This is not representative of the amount of computational resources required by the method.

Table 2. Comparison of performance metrics between various ICING configuration on synthetic datasets. Columns are: ϵ (the DBSCAN parameter for neighbourhood selection), *tolerance* (tolerance parameter on CDR3 length), *correction* (Y for a correction based on the mutation level of V gene segments, N for no correction), followed by the clustering measures as described in Section 3.4. For each dataset, results are ordered by a decreasing FMI, which is the most strict of the measures for its properties.

				no chance normalisation				*chance normalisation*		
dataset	ϵ	tolerance	correction	HOM	COM	VSC	NMI	AMI	ARI	FMI
	0.2	0	Y	0.91	0.94	0.92	0.92	0.90	0.86	0.87
	0.2	6	Y	0.90	0.94	0.92	0.92	0.89	0.86	0.86
D1	0.2	3	Y	0.87	0.94	0.90	0.90	0.86	0.76	0.78
	0.2	6	N	0.87	0.94	0.90	0.90	0.86	0.75	0.77
	0.2	0	N	0.86	0.94	0.90	0.90	0.85	0.75	0.77
	0.2	0	Y	0.93	0.93	0.93	0.93	0.93	0.90	0.91
	0.2	6	Y	0.93	0.93	0.93	0.93	0.93	0.90	0.91
D2	0.2	3	Y	0.93	0.93	0.93	0.93	0.93	0.90	0.90
	0.2	3	N	0.92	0.93	0.92	0.92	0.91	0.88	0.88
	0.2	0	N	0.91	0.93	0.92	0.92	0.91	0.87	0.88
	0.2	0	Y	0.94	0.93	0.93	0.93	0.92	0.92	0.92
	0.2	3	Y	0.94	0.92	0.93	0.93	0.92	0.92	0.92
D3	0.2	0	N	0.92	0.93	0.92	0.92	0.91	0.89	0.89
	0.2	6	Y	0.92	0.93	0.93	0.93	0.92	0.88	0.88
	0.2	6	N	0.92	0.93	0.92	0.92	0.91	0.87	0.87

obtained using different ICING configurations. The metric used for CDR3 sequence distance computation is the Hamming metric. The other parameters we investigated involve the neighbourhood selection radius of the DBSCAN clustering algorithm (restricted to 0.2 or 0.6), the tolerance of the difference in CDR3 sequence lengths (0, 3 or up to 6 allowed insertions or deletions), and the optional distance correction based on the V gene segment mutation level. Table 2 is ordered based on a decreasing FMI score, which, for its properties, it is the most strict of the clustering measures described in Section 3.4. The highest scores (close to 1) for each of the three datasets are associated to similar ICING configurations, in which the neighbourhood selection of the DBSCAN clustering algorithm is restricted to 0.2, the tolerance of the difference in sequence lengths is 0 (*i.e.*, no alignment between CDR3s needed to be done), and sequence distances are corrected based on the V gene segment mutation level. Particularly for dataset D1, the distance correction is shown to be a critical step to reliably identify IG clonotypes, as confirmed by high ARI, AMI and FMI scores (chance-corrected clustering measures). Notably, for D2 and D3 datasets, the correction gives better results when associated to a tolerance parameter of 0 or 6 nucleotides for CDR3 sequences.

Table 3. ICING results on synthetic datasets, using the best parameters as selected in Table 2 (ϵ: 0.2, *tolerance*: 0, *correction*: Y). For each datasets, clustering measures are reported as described in Section 3.4.

		no chance normalisation				*chance normalisation*		
dataset	sequences	HOM	COM	VSC	NMI	AMI	ARI	FMI
D4	47764	0.90	0.95	0.93	0.93	0.88	0.79	0.80
D5	102336	0.94	0.95	0.94	0.94	0.93	0.89	0.89
D6	205986	0.94	0.95	0.94	0.94	0.94	0.89	0.89
D7	162713	0.93	0.96	0.94	0.94	0.91	0.84	0.84
D8	301978	0.93	0.95	0.94	0.94	0.92	0.86	0.86
D9	589680	0.93	0.96	0.95	0.95	0.92	0.88	0.87
D10	291076	0.94	0.95	0.95	0.96	0.92	0.87	0.86
D11	568799	0.93	0.95	0.94	0.96	0.91	0.89	0.88
D12	1208110	0.95	0.94	0.95	0.95	0.90	0.88	0.90

The best parameters selected on datasets D1–3 were used to evaluate the results on the remaining datasets of Table 1. The results presented in Table 3 show that ICING is capable to achieve high performance, which means a reliable IG sequence clonotyping, even with an increasing number of sequences. Also, the method is stable across datasets with different sizes.

4.2 Expected Clonotypes

Figure 3 shows the number of clonotypes found by ICING compared to the expected clonotypes (*ground truth*). Inferred clonotypes are very close to the ground truth disregarding the size of the datasets. This result, together with the high clustering performance achieved by our method (Table 2 and Table 3), makes ICING a reliable framework for IG clonotype identification in real contexts, where real clonotypes are not known.

5 Conclusion

Our results show ICING to be capable of successfully identifying IG clonotypes, using synthetic data comprising highly mutated sequences, different V(D)J recombination events and *indels* on CDR3 sequences. Due to the intrinsic difficulty of validating the method on real data (where the ground truth is not known), we chose to include only the results obtained on synthetic data, where the method can be validated in relation to the ground truth.

ICING has a modular structure which allows to combine different features. In particular, the clonotype identification step has the potential to include Hamming or other arbitrary nucleotidic or amminoacidic models to compute sequence

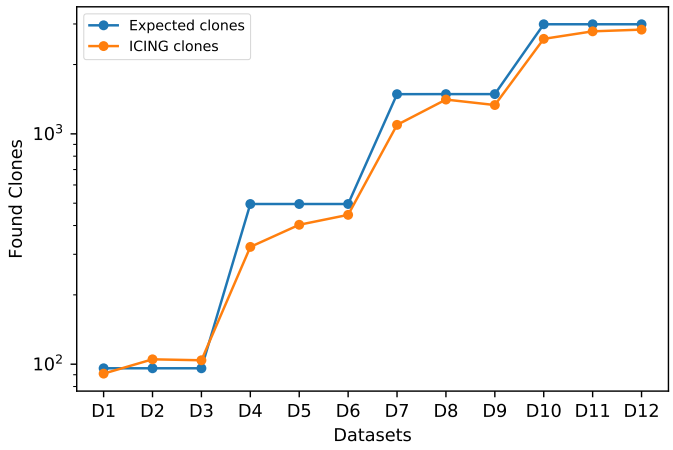

Fig. 3. Comparison between ICING clusters and expected clonotypes on synthetic datasets. For each dataset (x-axis), the number of clonotypes found by ICING is compared with the expected clonotypes (y-axis), *i.e.*, the *ground truth*. For datasets D1–3, only the best results based on FMI score (Table 2) are included.

distances, arbitrary CDR3 length tolerance or V gene sequence mutation-based correction, which is an original contribution of this framework. ICING is scalable with the number of input sequences, allowing for the analysis of large-scale datasets composed of more than 10^6 sequences, which is a typical use-case when dealing with NGS data. To achieve scalability, ICING is based on a novel methodology which exploits the DBSCAN clustering algorithm, on top of an appropriate index structure. In particular, we were not able to compare our pipeline with plain Change-O which, since it is based on hierarchical clustering, has memory complexity of $O(n^2)$, thus infeasible for large datasets. However, we were able to analyse arbitrarily large datasets by exploiting all of the steps shown in Section 3.3, which turned out to be fundamental in our analysis.

ICING is easily integrable in the usual pRESTO/Change-O pipeline for IG analysis and it is ready to be used in real scenarios. In presence of sequences with low rate of recombination and mutation (*i.e.*, as in the case of non-healthy patients), we expect the data shrinking step (Section 3.3) to be highly beneficial for reducing the complexity of the algorithm, which is proportional to the number of unique CDR3 sequences and V gene calls in the dataset.

Acknowledgments

The authors would like to thank the reviewers for their helpful and constructive comments that greatly contributed to improve the final version of the paper. DB thanks *Fondazione Umberto Veronesi* for the support.

References

1. Eltaf Alamyar et al. 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*, 8(1):26, 2012.
2. Martin Ester, Hans-Peter Kriegel, Jörg Sander, Xiaowei Xu, et al. A density-based algorithm for discovering clusters in large spatial databases with noise. In *Kdd*, volume 96, pages 226–231, 1996.
3. Edward B Fowlkes and Colin L Mallows. A method for comparing two hierarchical clusterings. *Journal of the American statistical association*, 78(383):553–569, 1983.
4. Jacob Glanville et al. Precise determination of the diversity of a combinatorial antibody library gives insight into the human immunoglobulin repertoire. *Proceedings of the National Academy of Sciences*, 106(48):20216–20221, 2009.
5. Namita T Gupta et al. Change-O: a toolkit for analyzing large-scale B cell immunoglobulin repertoire sequencing data. *Bioinformatics*, 31(20):3356–3358, 2015.
6. Richard W Hamming. Error detecting and error correcting codes. *Bell Labs Technical Journal*, 29(2):147–160, 1950.
7. Charles A Janeway, Paul Travers, Mark Walport, and Mark J Shlomchik. *Immunobiology: the immune system in health and disease*, volume 1. Current Biology Singapore, 1997.
8. Steven H Kleinstein, Yoram Louzoun, and Mark J Shlomchik. Estimating hypermutation rates from clonal tree data. *The Journal of Immunology*, 171(9):4639–4649, 2003.
9. Mihaela L Oprea. *Antibody repertoires and pathogen recognition: the role of germline diversity and somatic hypermutation*. PhD thesis, Citeseer, 1999.
10. Duncan K Ralph and Frederick A Matsen IV. Consistency of VDJ rearrangement and substitution parameters enables accurate B cell receptor sequence annotation. *PLoS Comput Biol*, 12(1):e1004409, 2016.
11. Edwin P Rock et al. CDR3 length in antigen-specific immune receptors. *The Journal of experimental medicine*, 179(1):323–328, 1994.
12. David Sculley. Web-scale k-means clustering. In *Proceedings of the 19th international conference on World wide web*, pages 1177–1178. ACM, 2010.
13. Temple F Smith and Michael S Waterman. Identification of common molecular subsequences. *Journal of molecular biology*, 147(1):195–197, 1981.
14. Jason A Vander Heiden et al. pRESTO: a toolkit for processing high-throughput sequencing raw reads of lymphocyte receptor repertoires. *Bioinformatics*, page btu138, 2014.
15. Nguyen Xuan Vinh, Julien Epps, and James Bailey. Information theoretic measures for clusterings comparison: is a correction for chance necessary? In *Proceedings of the 26th Annual International Conference on Machine Learning*, pages 1073–1080. ACM, 2009.
16. Gur Yaari et al. Models of Somatic Hypermutation Targeting and Substitution Based on Synonymous Mutations from High-Throughput Immunoglobulin Sequencing Data. *Frontiers in Immunology*, 4, 2013.