# Assessing the Consistency of Computational Tools for Quantifying Key Antibody Repertoire Metrics


JIANCHE LIU[1], Helen Huang[2,3,4], Haripriya Vaidehi Narayanan[2,3], Alexander Hoffmann[2,3]

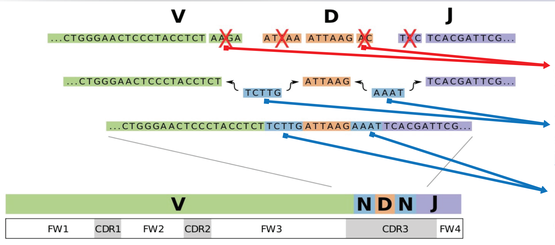
[1] BIG Summer Program, Institute for Quantitative and Computational Biosciences, UCLA
[2] Department of Microbiology, Immunology, and Molecular Genetics (MIMG), UCLA
[3] Signaling Systems Laboratory, Institute for Quantitative and Computational Biosciences, UCLA
[4] Bioinformatics Interdepartmental PhD Program, UCLA


## Abstract


B-cell undergoes somatic recombination and hypermutation to construct diverse B-cell receptors (BCRs) to bind different antigens. The diversity in these BCR sequences poses challenges in drawing biologically meaningful conclusions, calling for effective computational software. In the past decade, several tools have been developed to assign germline genes, determine complementarity-determining region 3 (CDR3), and characterize sequence mutation frequency and selection landscape. However, no work has been done to benchmark the performance of these tools and guide the selection of software. Here, we implemented a few of the highly cited software (Change-O, MixCR, and Partis) to compare the BCR repertoire between NFκB mutant and wide-type mice. We found that these software packages showed consistency in summarizing clonal diversity and CDR3 length distribution but diverged in quantifying mutation frequency and selection pressure. Our results demonstrate the value of comparing software using real data, and provide insights into software selection in BCR repertoire analysis.


## Background

### 1. B-cell undergoes somatic recombination and hypermutation to construct diverse antibodies



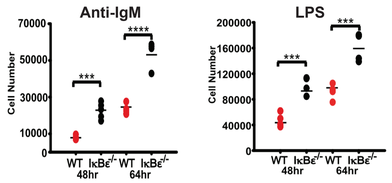

Recombination

1. B cell randomly selects V, D, and J gene segments from the gene pools in the body.
2. B cell randomly deletes nucleotides at both ends of the V-D and D-J junction regions.
3. B cell randomly inserts nucleotides in the same junction regions as the deletions (N = non-templated insertion).

Hypermutation

4. B-cell undergoes somatic hypermutation (SHM), generating diversified B-cell receptors (BCRs).

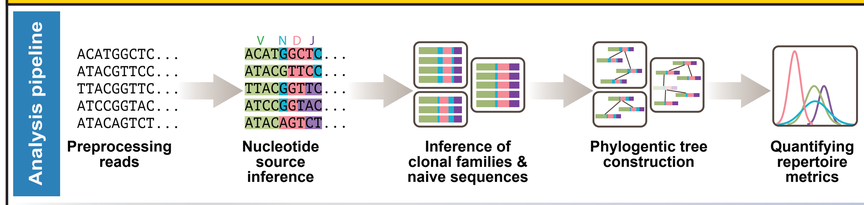

### 2. IκBε deficiency in B cells results in increased stimulus-responsive proliferation and survival [1]



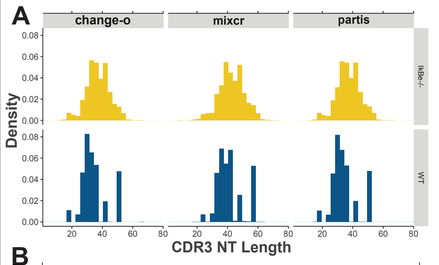

### 3. Many computational tools were developed to analyze BCR repertoire

| | | pRESTO [9] | MixCR [2] | Change-O [4] | Partis [8] | IGoR [6] | SHazaM [3] |
|---|---|---|---|---|---|---|---|
| Summary | | Perform Raw sequence processing prior to alignment | A compositive immune reportoire processing pipeline with multiple preset for commercial sequencing kit | A collection of tools performing V(D)J alignment and clonal clustering | A pipeline performing sequence annotation, simulation, clonal clustering and mutation profilling | Evaluate all possible recombination scenarios for the read based on given or learned model with its probabilities | An R package for mutation and selection quantification |
| BASIC PROCESSING | Preprocessing | • Single-end<br>• Paired-ends<br>• Input: fastq or fasta format<br>• Able to perform UMI based correction | • Single-end<br>• Paired-ends<br>• Input: fastq or fasta format<br>• Can do amplification error correctionbe chosen based on the sequencing methods | -- | -- | -- | -- |
| | V(D)J Alignment | -- | • VDJ+C<br>• Output: .vdjca / AIRR-tsv | • VDJ alignment based on IgBlast<br>• Output: AIRR-tsv | • VDJ alignment based on hidden-markov model<br>• Output: .yaml / Able for AIRR-tsv | • VDJ alignment<br>• The learning of the new model is based on pygor<br>• Output: series of .csv file /grouped to AIRR-tsv in pygor | |
| | Clonal Grouping | -- | • Grouped by clonal sequence (CDR3 region by default)<br>• Input: .vdjca<br>• Output: .clns / Able for AIRR-tsv | • Grouped the clonal sequence by hamming distance on CDR3 region (AA or nucleotides)<br>• Output: .tab/ Able for AIRR-tsv | • First find the most likely germline sequence (common ancester) then group each cluster by hamming distance<br>• Output: .yaml / Able for AIRR-tsv | -- | -- |
| REPERTOIRE CHARACTERIZATION | Diversity | -- | • Can be estimated based on clonal AIRR-tsv | • Can be estimated based on clonal AIRR-tsv | • Can be estimated based on clonal AIRR-tsv | -- | -- |
| | Mutation Analysis | -- | • Can be done by SHazaM based on clonal AIRR-tsv | • Can be done by SHazaM based on clonal AIRR-tsv | • Can be done by SHazaM based on clonal AIRR-tsv<br>• Have bulid-in function to plot out the SHM | -- | • Input: AIRR-tsv<br>• Quantification of mutational load |
| | Selection Estimation | -- | • Can be done by SHazaM based on clonal AIRR-tsv | • Can be done by SHazaM based on clonal AIRR-tsv | • Can be done by SHazaM based on clonal AIRR-tsv | -- | |

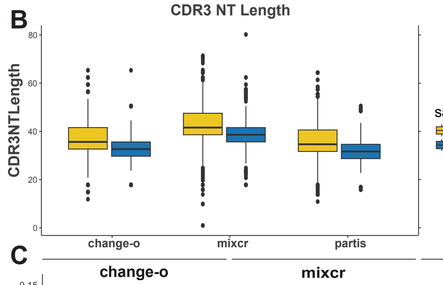
## Methods & Results




### 1. Consistent estimates of CDR length and the rank abundance curve indicate greater clonal diversity in IκBε-/- mice

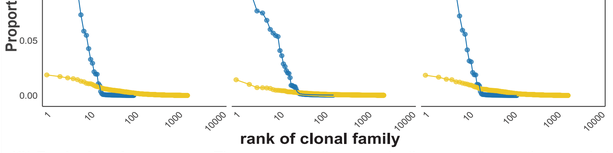



- The distribution of CDR3 length is associated with the polymorphism of the clone: a CDR3 length distribution that is "skewed" or "perturbed" may indicate oligoclonality whereas the distribution similar to a Gaussian-like distribution is associated with being polyclonal and more diverse [7].
- CDR3 length is associated with a B-cell's ability to recognize diverse antigens. Long CDR3 has been associated with auto-reactivity and polyreactivity in previous study [5].

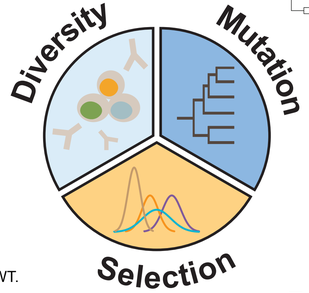



**(A) CDR3 nucleotide length distribution.** All computational tools shows that the CDR3 distribution in IκBε-/- have more tendency to be Gaussian-like distribution than WT, indicating a more diversified repertoire in IκBε-/-.

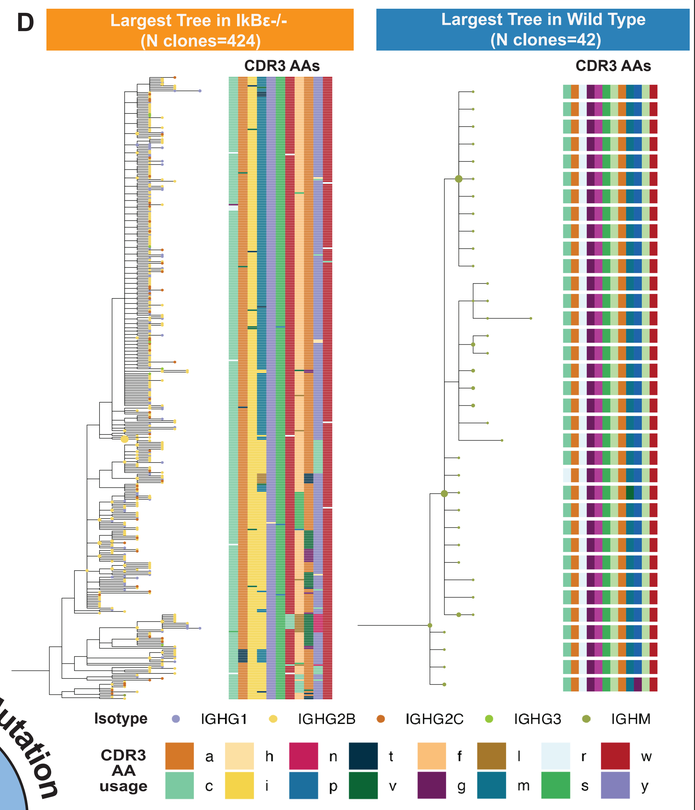
**(B) CDR3 nucleotide length box plot.** Longer CDR3 length in IκBε-/- sample was conserved among different softwares, suggesting auto-reactivity or polyreactivity.




**(C) Rank-abundance curve.** The more even the curve is, the more diverse the sample is. IκBε-/- repertoire shows more even distribution in rank-abundance curve, suggesting higher diversity than WT.




### 2. Increased mutation rate is consistently observed in IκBε-/- over WT, but softwares diverge in their numerical quantification




**(D) Phylogenetic tree constructed based on somatic hypermutation (SHM) distance from germline by MixCR.** The largest phylogetic tree were visualized with their CDR3 amino acid sequence in each clone. IκBε-/- sample have more complecated mutational landsacape than WT, with more mutation distance and more clone in the SHM tree.

**(E) Boxplot of mutation frequency from different software.** Generally, more mutation frequency in IκBε-/- sample was observed among different softwares although different software diverged in quantifying the exact value of mutation frequency.

### 3. Computational tools yielded consistency in identifying positive selection in CDR and negative selection in FWR, but diverged in quantifying the selection

- **Estimating Selection Strength Σ:**

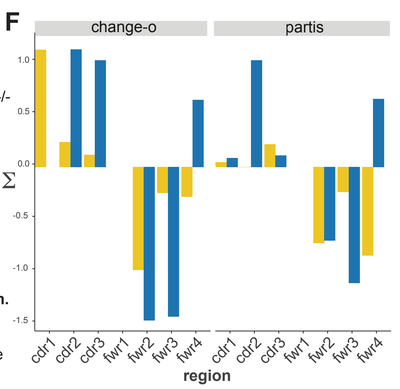
  Selection is estimated by the observed frequency of replacement and silent mutations normalized by their expectated frequency based on a targeting model [10].

$$\sum = log\frac{\pi/(1-\pi)}{\hat{\pi}/(1-\hat{\pi})}$$

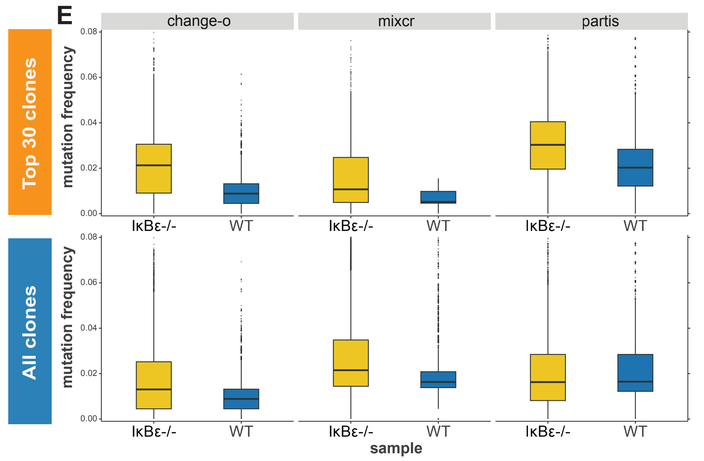
  - Σ>0: Positive selection
  - Σ<0: Negative selection

$$\hat{\pi} = \frac{\hat{R}}{\hat{R}+\hat{S}} \qquad \pi = \frac{R}{R+S}$$

  - *R*: the number of replacement mutations
  - *S*: the number of silent (i.e. synonymous) mutations.




**(F) Change-O and partis diverge in selection strength estimation.** The clone of change-o showed less selection strength in IκBε-/- sample among all BCR region, whereas partis's clone didn't. The positive selection in CDR region and negative selection in FWR were conserved between the softwares.




## Conclusion

- **Antibody profiling is a dynamic research field -- multiple software have been developed.**
- **Typical repertoire metrics include CDR3 characteristic, clonal diversity, mutation rate and selection pressure.**
- **Change-O and MixCR have a completed BCR profilling pipeline and is more user-friendly.**
- **Different software showed consistency in concluding clonal diversity and CDR3 length distribution in the antibody repertoire.**
- **Different softwares diverged in the exact numerical quantification of mutation rate and selection strength in antibody repertoire, with consisency in drawing qualitative conclusion between IκBε-/- sample and WT.**

## References


1. Alves, B.N. et al. (2014), *The Journal of Immunology.*
2. Bolotin, D.A. et al. (2015), *Nature Methods.*
3. Cui, A. et al. (2016), *Journal of Immunology.*
4. Gupta, N.T. et al. (2015), *Bioinformatics.*
5. Hou, X.-L. et al. (2016), *Genes and Immunity.*
6. Marcou, Q., Mora, T. and Walczak, A.M. (2018), *Nature Communications.*
7. Miqueu, P. et al. (2007), *Molecular Immunology.*
8. Ralph, D.K. and Matsen, F.A. (2016), *PLoS computational biology.*
9. Vander Heiden, J.A. et al. (2014), *Bioinformatics.*
10. Yaari, G., Uduman, M. and Kleinstein, S.H. (2012), *Nucleic Acids Research.*