NoisET: Noise learning and Expansion detection
of T-cell receptors with Python

Meriem Bensouda Koraichi,[1] Maximilian Puelma Touzel,[2] Thierry Mora,[1, *] and Aleksandra M. Walczak[1, *]

[1]*Laboratoire de physique de l'École normale supérieure, CNRS, PSL University,*
Sorbonne Université, and Université de Paris, 75005 Paris, France
[2]*MILA, University of Montreal, Montreal, Canada*

High-throughput sequencing of T- and B-cell receptors makes it possible to track immune repertoires across time, in different tissues, in acute and chronic diseases and in healthy individuals. However quantitative comparison between repertoires is confounded by variability in the read count of each receptor clonotype due to sampling, library preparation, and expression noise. We present an easy-to-use python package *NoisET* that implements and generalizes a previously developed Bayesian method. It can be used to learn experimental noise models for repertoire sequencing from replicates, and to detect responding clones following a stimulus. The package was tested on different repertoire sequencing technologies and datasets.

Availability: *NoisET* is freely available to use with source code at `github.com/statbiophys/NoisET`.

INTRODUCTION

High-Throughput Repertoire Sequencing (RepSeq) of T and B cell receptors (TCR and BCR) [3] enables to study the dynamics of lymphocytes at the resolution of single clones, by comparing their concentrations across timepoints or conditions. To detect biologically relevant clones, one must be able to distinguish true differences in clone frequencies from experimental noise. This variability has two sources. First, laboratories use various sequencing and sample preparation protocols using either gDNA or cDNA (with or without unique molecular identifiers), with different outcomes in terms of amplification bias and errors [2, 4]. This makes it difficult to reliably estimate TCR or BCR clonal frequencies from sequence counts. Second, one must translate immune information contained in a few milliliters of blood to the whole repertoire. To describe these sources of variability, one needs a probabilistic approach.

Touzel *et al.* [9] developed a statistical model to identify responding clones using sequence counts in longitudinal RepSeq data. This model captures features of a repertoire response to a single, strong perturbation (e.g. yellow fever vaccination), giving rise to a fast transient response dynamics. The method was proposed as an alternative to commonly used tests such as Fisher's exact test [1] or beta binomial models [8]. Its main innovation is to account for the different sources of biological and experimental noise in the clone count measurements in a Bayesian way, allowing for a more reliable detection of expanded or contracted clones. This note introduces *NoisET* (Noise sampling learning and Expansion detection of TCRs), an easy-to-use python package that implements this method and extends it to datasets of diverse origin describing the clonal repertoire response to acute infections.

FEATURES

NoisET has two main functions: (1) inference of a statistical null model of sequence counts and variability, using replicate RepSeq experiments; (2) detection of responding clones to a stimulus by comparison of two repertoires taken at two timepoints. The second function requires a noise model, which is given as an output of the first function. Both functions require two lists of sequence counts associated to each TCR or BCR present in the repertoires: from replicate experiments for the first function (Fig. 1a), and from repertoires before and after the stimulus for the second function (Fig. 1b).

When using the first function, the user must pick the type of noise model, which describes how the sequence count in the RepSeq sample depends probabilistically on its true frequency in the blood. Choices are: a Poisson distribution, a negative binomial distribution, or a two-step model [9]. Once the parameters have been learnt (Maximum Likelihood Estimation optimization algorithm), a generation tool can be applied to qualitatively check the agreement between data and model for replicates (Fig 1a). We also successfully learnt a null model from gDNA data [8], which is included in the package example notebook.

To use the second function to detect responding clonotypes, the user provides, in addition to the two datasets to be compared, two sets of experimental noise parameters learnt at both times using the first function. When replicates are not available for each time point or donor, a common null model may be used for both timepoints. This should be done with caution, since even if both samples are produced with the same technology for the same donor, the sequencing depth and distribution of clone frequencies may vary between timepoints. Finally the user provides two thresholds: one for the posterior probability above which a clone is labeled as responding, and one for the median log-fold frequency difference above which
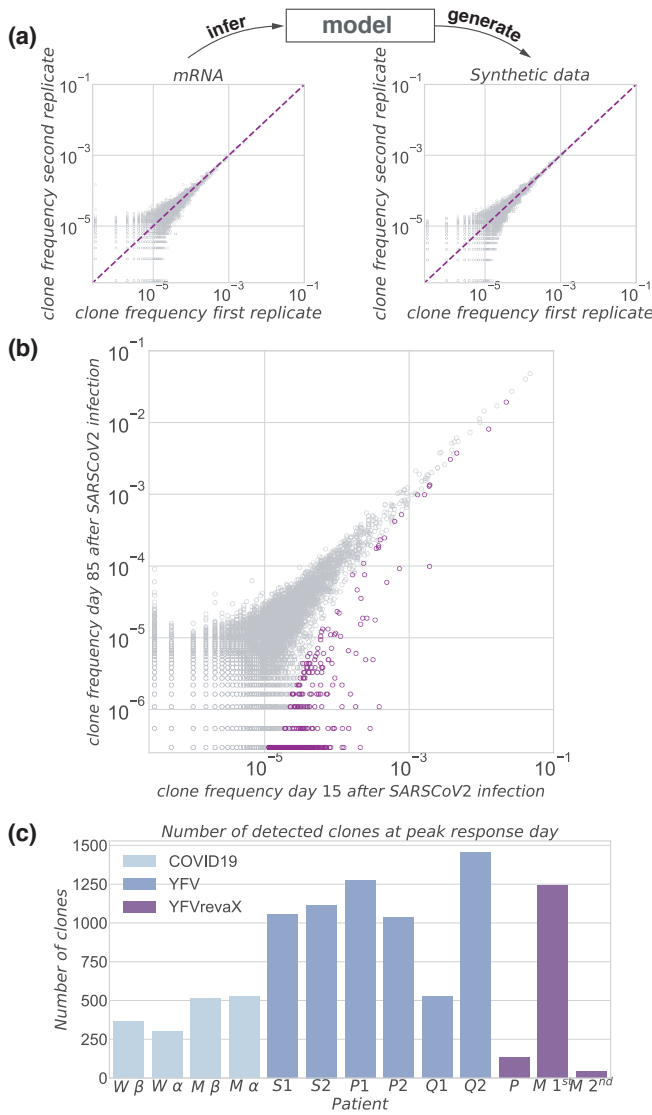
FIG. 1: **(a)** Scatter plots of sequence counts from two biological replicates from Pogorelyy *et al.* [7] (left). *NoisET* learns a statistical model of sequence frequencies and observed counts from these data (with Negative Binomial sampling noise model), which can then be used to generate realistic synthetic data (right). **(b)** Scatter plot of contracted clones from day 15 to day 85 after a mild COVID-19 infection [6]. Clones detected as contracting by *NoisET* are shown in purple. **(c)** Number of responding clones detected by *NoisET* (using a two step noise model) for 3 studies: donors M and W (with both α and β TCR chains) in response to COVID-19 between days 15 and 85 [6]; 6 twin donors (S1 through Q2, only β chain) between days 0 and 15 following yellow-fever vaccination [7]; and yellow-fever first (M) and second vaccination (M and P) [5].

of responding clonotypes detected by NoisET applied to three different datasets revealing COVID-19 and yellow fever vaccine TCR response dynamics.

All functions are explained in a well-documented README and notebooks displayed on the Github repository.

DISCUSSION

NoisET is designed as an easy-to-use package to learn the noisy statistics of sequence counts and to detect responding clones to a stimuli as reliably as possible. It captures the experimental and biological noise for both RNAseq and gDNAseq replicate technologies. Although the package has been tested on diverse datasets, choosing and using the adequate statistical null model should be done with caution. Among the different types of noise model offered, the negative binomial noise model is recommended to start the analysis as its running time is shorter than the two step model, while retaining the ability to account for arbitrary noise amplitudes. So far, *NoisET* has been used to study the short time scale dynamics for acute infections, but could also be used to compare bulk repertoires with selected repertoires derived from functional or cultured assays [1]. For longer time scales, the dynamics of lymphocyte populations should be modeled to best describe slow global repertoire changes that cannot be attributed to a single stimulus.

Ackowledgements

---

* Corresponding authors. These authors contributed equally.

[1] Balachandran *et al.*(2017) Identification of unique neoantigen qualities in long-term survivors of pancreatic cancer. *Nature*, **551**(7681), 512–516.

[2] Barennes *et al.*(2020) Benchmarking of T cell receptor repertoire profiling methods reveals large systematic biases. *Nature Biotechnology* online ahead of print.

[3] Benichou *et al.*(2012) Rep-Seq: uncovering the immunological repertoire through next-generation sequencing. *Immunology*, **135**(3), 183–191.

[4] Heather *et al.*(2017) High-throughput sequencing of the T-cell receptor repertoire: pitfalls and opportunities. *Briefings in Bioinformatics*, **19**(4), 554–565.

[5] Minervina *et al.*(2020) Primary and secondary anti-viral response captured by the dynamics and phenotype of individual T-cell clones. *eLife*, **9** e53704.

detection is allowed. The output is a CSV file containing a table of putative responding clones. The result is illustrated in Fig. 1b, which shows contracted clones (purple points) detected from day 15 to day 85 from a mild COVID-19 infection [6]. Fig. 1c reports the number

[6] Minervina *et al.*(2021)   Longitudinal high-throughput TCR repertoire profiling reveals the dynamics of T-cell memory formation after mild COVID-19 infection. *eLife*, **10** e63502.

[7] Pogorelyy *et al.*(2018)   Precise tracking of vaccine-responding t cell clones reveals convergent and personalized response in identical twins. *Proceedings of the National Academy of Sciences*, **115**(50), 12704–12709.

[8] Rytlewski *et al.*(2019) Model to improve specificity for identification of clinically-relevant expanded T cells in peripheral blood. *PLOS ONE*, **14**(3), e0213684.

[9] Touzel *et al.*(2020) Inferring the immune response from repertoire sequencing.   *PLOS Computational Biology*, **16**(4), e1007873.