THE JOURNAL OF PHYSICAL CHEMISTRY A
A JOURNAL OF THE AMERICAN CHEMICAL SOCIETY

pubs.acs.org/JPCA

Article

NoisET: Noise Learning and Expansion Detection of T-Cell Receptors

Published as part of The Journal of Physical Chemistry virtual special issue "Jose Onuchic Festschrift".

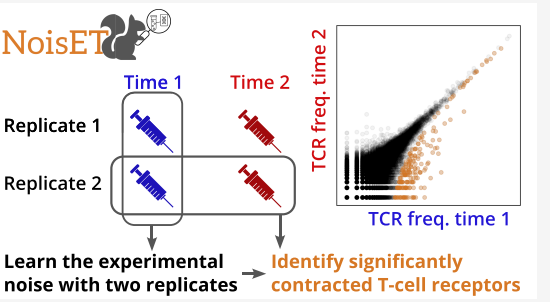
Meriem Bensouda Koraichi, Maximilian Puelma Touzel, Andrea Mazzolini, Thierry Mora,*,# and Aleksandra M. Walczak*,#

Cite This: *J. Phys. Chem. A* 2022, 126, 7407−7414

Read Online

ACCESS | Metrics & More | Article Recommendations

ABSTRACT: 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 review methods for accounting for both biological and experimental noise and 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. We test the package on different repertoire sequencing technologies and data sets. We review how such approaches have been used to identify responding clonotypes in vaccination and disease data. Availability: *NoisET* is freely available to use with source code at github.com/statbiophys/NoisET.



I. INTRODUCTION

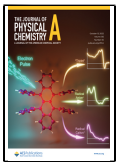
Cells of the adaptive immune system, T and B lymphocytes, recognize molecules foreign to our body and protect us against pathogenic threats. These cells also have the ability to eliminate cells that harbor anomalies, such as cancer cells. Lymphocytes perform this discrimination task between potentially dangerous and normally functioning "self" molecules using specialized receptors on their surface that constantly sample and bind molecules in our organisms. Each cell has one type of receptor and the system relies on a large diversity of a repertoire of different receptors expressed on over 10^9 different B or T cells to protect the organism against infections.[1−4]

The composition of the repertoire contains information about past infections and conditions. Reading this information requires quantitatively understanding the natural repertoire dynamics. Upon recognition of a pathogenic molecule, the recognizing cell proliferates making many cells with the same receptor, forming a clone, which enables fast infection clearance. New cells are constantly produced and introduced into this diverse repertoire. Additionally to specific stimulation, cells also undergo random divisions. Each cell has a finite lifetime and clones can go extinct if all the cells of that clone die. Together these processes define a natural dynamics of the repertoire, which leads to a constantly changing set of different cells present at different frequencies.

High-Throughput Repertoire Sequencing (RepSeq) of T and B cell receptors (TCR and BCR)[1,5−13] enables us to study the dynamics of lymphocytes at the resolution of single clones, by comparing their concentrations across time points or conditions. To detect biologically relevant clones, one must be able to distinguish true differences in clone frequencies from experimental noise. This variability has three 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.[11,14] This makes it difficult to reliably estimate TCR or BCR clonal frequencies from sequence counts. Second, in the case of cDNA based sequencing, these uncertainties are not solely due to different sample preparation but have a more fundamental, biological source. mRNA is produced in bursts,[15−19] which adds a natural longtailed noise to the sequencing read distribution. Third, 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.

Puelma Touzel et al.[20] developed a statistical model to identify responding clones using sequence counts in longitudinal RepSeq data. This model captures features of a

Received: July 15, 2022
Revised: September 15, 2022
Published: September 30, 2022



Downloaded via INIST-CNRS on December 15, 2022 at 09:26:03 (UTC).
See https://pubs.acs.org/sharingguidelines for options on how to legitimately share published articles.

ACS Publications
© 2022 American Chemical Society

https://doi.org/10.1021/acs.jpca.2c05002
J. Phys. Chem. A 2022, 126, 7407−7414

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[21] or beta binomial models.[22] 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.

Here we briefly review the ideas behind the method that calibrates the noise and we introduce NoisET (Noise sampling learning and Expansion detection of TCRs), an easy-to-use python package that implements this method and extends it to data sets of diverse origin describing the clonal repertoire response to acute infections. We also review several applications of this approach.

II. METHODS

In order to correctly identify expanding or contracting clonotypes, whether after direct antigenic stimulation or due to random cell division and death, we need to correctly separate biological and experimental noise from the lymphocyte dynamics. The main idea behind the Bayesian probabilistic modeling method implemented in the NoisET software is learning probabilistic distributions describing sampling and experimental noise from empirical frequencies of TCR counts in biological replicate samples from the same individual. In this section we introduce the two types of models implemented in NoisET: the noise model and the response model.

II.A. Modeling Experimental Noise. TCR sequencing (TCRseq) methods, depending if they are based on DNA or RNA input, produce data with different characteristics. For example, RNA-based methods allow for the usage of unique molecular identifiers (UMI) to limit PCR amplification bias and sequencing errors. Non-UMI methods are better in capturing rare clones which motivates their frequent usage.[14] During this first step, NoisET learns at the same time the exponent of the underlying power-law TCR frequency distribution, $\rho(f)$[5] and the parameters of error model between the empirical abundance of one specific TCR clone $\hat{n}$ and its true frequency f: $P(\hat{n}|f)$. NoisET has also the power to learn these distributions constraining the size of the clones that we want to take into account for the analysis.

For each TCR clone, the likelihood to sample $\hat{n}$ reads from the first biological replicate and $\hat{n}'$ reads from the second biological replicate is

$$\mathbb{P}(\hat{n}, \hat{n}'|\Theta) = \int_{f_{\min}}^{1} \mathrm{d}f \rho(f|\Theta) P(\hat{n}|f, \Theta) P(\hat{n}'|f, \Theta) \quad (1)$$

where Θ are the parameters of the noise model which define the error model $P(\hat{n}|f)$, $f_{\min}$ corresponds to the minimum clonal frequency for each individual, and $\rho(f)$, which is the clonal frequency prior known to be a power-law distribution $\propto f^{\alpha}$.[5,23] NoisET learns the parameters of the noise model Θ by maximizing the log-likelihood of the observed TCR counts, $\hat{n}$, $\hat{n}'$, from the two biological replicates:

$$\Theta^* = \underset{\Theta}{\operatorname{argmax}} \prod_{i=1}^{N_{\mathrm{obs}}} \mathbb{P}(\hat{n}_i, \hat{n}_i'|\Theta). \quad (2)$$

Since in RepSeq samples we only partially sample an individual's repertoire, the likelihood in eq 1 needs to be modified. We condition the likelihood on observing that specific clone in at least one of the two replicates: $\mathbb{P}(\hat{n} + \hat{n}' > 0)$. The modified likelihood becomes

$$\mathbb{P}(\hat{n}, \hat{n}'|\hat{n} + \hat{n}' > 0) = \mathbb{P}(\hat{n}, \hat{n}', \hat{n} + \hat{n}' > 0) / \mathbb{P}(\hat{n} + \hat{n}' > 0)$$

We can also choose to learn the noise model only on clones having a size larger than a certain threshold. In this case the likelihood in eq 1 becomes

$$\mathbb{P}(\hat{n}, \hat{n}'|\hat{n} > \hat{n}_{th}, \hat{n}' > \hat{n}_{th}) = \mathbb{P}(\hat{n}, \hat{n}', \hat{n} > \hat{n}_{th}, \hat{n}' > \hat{n}_{th}) / \mathbb{P}(\hat{n} > \hat{n}_{th}, \hat{n}' > \hat{n}_{th})$$

To take into account different possible sources of noise due to the various RepSeq method, NoisET gives the choice of three different probabilistic distributions to learn the biological and experimental noise in the measured TCR abundances, $P(\hat{n}|f)$:

- The Poisson distribution, $P(\hat{n}|f) = \mathrm{Poisson}(fN_r)$. In this case, the noise parameters Θ are the exponent α of the clone-size distribution $\rho(f) = Cf^{\alpha}$ and the minimum clonal frequency in eq 1, $f_{\min}$. N_r is the total number of reads in the sample.
- The negative binomial distribution:

  $$P(\hat{n}|f) = \mathrm{NegBin}(\hat{n}; N_r f, N_r f + a(N_r f)^b)$$

  where $\mathrm{NegBin}(n; x, \sigma)$ is a negative binomial of mean x and variance σ. In this case, $\Theta = (\alpha, a, b, f_{\min})$ with α, $f_{\min}$ being the same parameters as described for the Poisson distribution, and a and b being the parameters of the negative binomial distribution.
- The negative binomial combined with a Poisson distribution:

  $$P(\hat{n}|f) = \sum_{m_i}^{\infty} P(\hat{n}|m_i) P(m_i|f)$$

  with $P(m_i|f) = \mathrm{NegBin}(m_i; fM, fM + a(fM)^b)$ and $P(\hat{n}|m_i) = \mathrm{Poisson}(m_i N_r/M)$. A clone of size f appears in a sample containing M T-cells on average as fM cells. To account for over dispersion, the number of cells associated with a specific clone is m and follows a negative binomial of mean fM and variance $fM + a(fM)^b$. For each clone the empirical abundance read in the biological sample is distributed according to a Poisson distribution with mean mN_r/M. For this model $\Theta = (\alpha, a, b, M, f_{\min})$.

While the mathematical framework is the same, when applied to identifying expanding clonotypes NoisET uses noise parameters inferred at both time points, contrary to the approach taken in refs 24 and 20. Experimental conditions at both time points can vary, and it is important to use both sets of parameters Θ to have the correct form of $P(\hat{n}|f, t_1)$ and $P(\hat{n}|f, t_2)$. The exponent of the power-law α and $f_{\min}$ in eq 1 are the learned values inferred at the time point for which sequencing depth is the larger.

II.B. Detecting Responding TCR clones. To account for differential expression after antigenic stimulation, NoisET implements the approach of previous work[24,20] that introduced a selection factor s defined as the log-fold change between a clone's frequency at time $t_1 f(t_1) = f$, and the frequency at time

https://doi.org/10.1021/acs.jpca.2c05002
J. Phys. Chem. A 2022, 126, 7407−7414

t_2, $f(t_2) = fe^s$. A prior is assumed over the variable s, $P(s|\gamma, \bar{s}) = \gamma \exp(-|s|/\bar{s})/(2\bar{s}) + (1-\gamma)\delta(s)$, with $0 \leq \gamma \leq 1$ the fraction of responding clones and $\bar{s} > 0$ being their typical effect size. The likelihood associated with observing a clone with empirical abundances $\hat{n}_1$ at time t_1 and $\hat{n}_2$ at time t_2 integrating the prior knowledge over the log-fold change s is the following:

$$\mathbb{P}(\hat{n}_i(t_1) = \hat{n}_1, \hat{n}_i(t_2) = \hat{n}_2|\gamma, \bar{s}) = \iint df_1 \rho(f_1) ds P(s|\gamma, \bar{s}) P(\hat{n}_1|f_1) P(\hat{n}_2|f_1 e^s) \quad (3)$$

The parameters $(\gamma, \bar{s})$, are learned by maximizing the likelihood of the count pair data taken at two given time points:

$$(\gamma^*, \bar{s}^*) = \underset{(\gamma,\bar{s})}{\operatorname{argmax}} \prod_{i=1}^{N_{obs}} \frac{\mathbb{P}(\hat{n}_i(t_1), \hat{n}_i(t_2)|\gamma, \bar{s}, \Theta(t_1), \Theta(t_2))}{\mathcal{Z}(\gamma, \bar{s})} \quad (4)$$

where $\mathcal{Z}(\gamma, \bar{s})$ is a normalization factor accounting for the probability to observe TCR clone counts in both analyzed samples, and $\Theta(t_1)$, $\Theta(t_2)$ being the noise parameters learned at both time points t_1 and t_2 with NoisET. These two parameters were then used to compute the posterior $\mathbb{P}(s|\hat{n}_i(t_1), \hat{n}_i(t_2))$:

$$\mathbb{P}(s|\hat{n}_1, \hat{n}_2) = \frac{\mathbb{P}(\hat{n}_1, \hat{n}_2|s, \gamma, \bar{s}) P(s|\gamma, \bar{s})}{\mathbb{P}(\hat{n}_1, \hat{n}_2)} \quad (5)$$

The knowledge of the log-fold change posterior (5) is used to discriminate expanded or contracted clones from the bulk between t_1 and t_2. In analogy with p-values, we define $p = \mathbb{P}(s \leq 0|\hat{n}_1, \hat{n}_2, \gamma, \bar{s}, \Theta(t_1), \Theta(t_2))$, the probability corresponding to the null hypothesis of no expansion. If $p <$ threshold, the clone is classified as expanded. When looking at contraction, we use the same method reversing times t_1 and t_2 and looking at significant expansions from t_2 to t_1. The value of the threshold can be chosen by the user. In all the results presented in this review, the threshold was set to 0.05, however no threshold was applied when identifying contracting clones. Another threshold on the median of the $\mathbb{P}(s|\hat{n}_1, \hat{n}_2)$ distribution can be applied to select for clones that are greatly expanded.

The output of NoisET detection of responding clones is the list of statistical properties of the true log-fold change variable s called according to the posterior $P(s|\hat{n}_1, t_1, \hat{n}_2, t_2)$ learned from data after learning the noise and differential model (eqs 2 and 4). These statistics are mathematically defined in Table 1 and are the values of s that defines the first quantile $s_{1,low}$, the median of the posterior $s_{2,med}$, the value of s that defines the third quantile $s_{3,high}$, the mode of the posterior s_{max}, the average of the posterior $\bar{s}$ and the p-value-like value defined as $P(s \leq 0|\hat{n}_1, \hat{n}_2)$.

Table 1. Mathematical Definition of Statistical Properties of the Hidden Variable s, the Log-Fold Change of Counts of a Given Clone, Computed from the Posterior Distribution $P(s|\hat{n}_1, t_1, \hat{n}_2, t_2)$, Learned from the Noise and Differential Model[a]

Feature	Description
$s_{1,low}$	$\int_{-\infty}^{s_{1,low}} \mathbb{P}(s\|\hat{n}_1, \hat{n}_2)\, ds = 0.025$
$s_{2,med}$	$\int_{-\infty}^{s_{2,med}} \mathbb{P}(s\|\hat{n}_1, \hat{n}_2)\, ds = 0.5$
$s_{3,high}$	$\int_{-\infty}^{s_{3,high}} \mathbb{P}(s\|\hat{n}_1, \hat{n}_2)\, ds = 0.975$
s_{max}	$\underset{(s)}{\operatorname{argmax}} \mathbb{P}(s\|\hat{n}_1, \hat{n}_2)$
$\bar{s}$	$\int_{-\infty}^{+\infty} s\mathbb{P}(s\|\hat{n}_1, \hat{n}_2)\, ds$
$1 - \mathbb{P}(s > 0)$	$\int_{-\infty}^{0} \mathbb{P}(s\|\hat{n}_1, \hat{n}_2)\, ds$

[a]The output of NoisET when detecting significantly expanded clones consists of the list of clones that are detected to have respectively increased or decreased in term of abundance associated with these specific s characteristics.

III. RESULTS

III.A. Features of the Software. NoisET has two main functions: (1) inference of a statistical null model of sequence counts and variability, using replicate RepSeq experiments, as described by the models presented in section II.A; (2) detection of responding clones to a stimulus by comparison of two repertoires taken at two time points, as described by the models presented in section II.A. 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 with each TCR or BCR present in the repertoires: from replicate experiments for the first function (Figure 1a left), and from repertoires before and after the stimulus for the second function (Figure 1a right). In addition, NoisET has features for detecting the time points to be compared, to simulate natural immune repertoire dynamics, and to estimate diversity.

All functions are described in a README and notebooks available on the Github repository (https://github.com/statbiophys/NoisET). A tutorial explains the different functions of NoisET.

1. Detecting the Peak Moment of the Response. When more than two time points are available, and when the time scales of the dynamical response of the TCR repertoire to an acute infection are not known, it is difficult to know which pairs of time points in longitudinal data can be informative about responding clonotypes. A method based on Principal Component Analysis (PCA) of longitudinal trajectories was first used in refs 26 and 25 to identify the peak of the response (Figure 1b). It uses the first two PCA components of the 1000 most abundant TCR clonotype frequencies normalized by their maximum postinfection values. The clustered trajectories identify different modes of clonal abundance dynamics. NoisET includes a feature for performing this PCA on trajectories as a preliminary step to pick the best time points.

2. Learning the Noise Model. When learning a noise model from replicates, 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.[20] Once the parameters have been learned (Maximum Likelihood Estimation optimization algorithm), a generation tool can be applied to qualitatively check the agreement between data and model for replicates (Figure 1a). We also successfully learned a null model from gDNA data,[22] which is included in the package example notebook.

https://doi.org/10.1021/acs.jpca.2c05002
J. Phys. Chem. A 2022, 126, 7407−7414

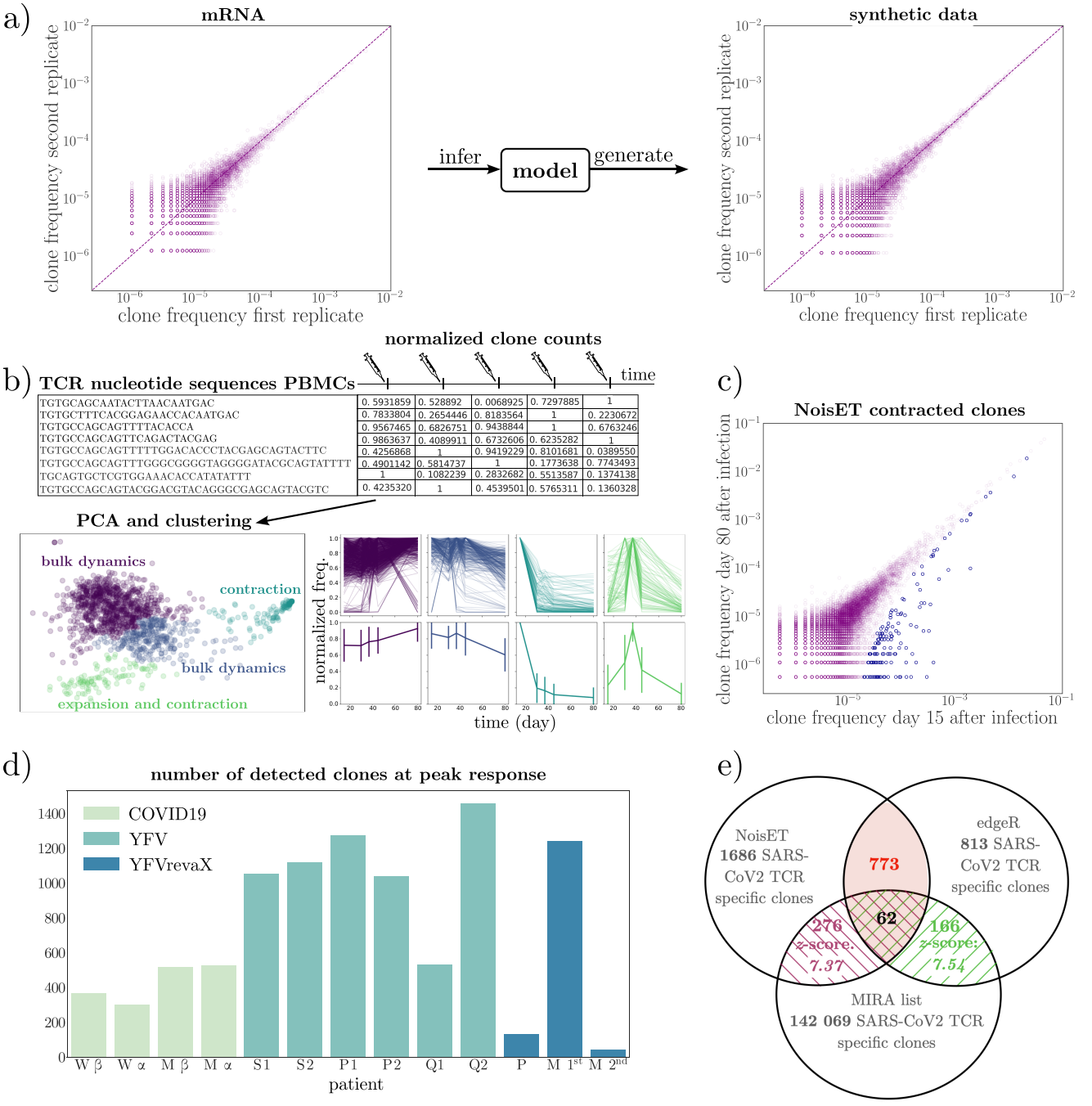


Figure 1. (a) Scatter plots of sequence counts from two biological replicates from ref 24 (left). NoisET learns a statistical model of sequence frequencies and observed counts from these data (here with negative binomial sampling noise model), which can then be used to generate realistic synthetic data (right). (b) PCA (Principal Component Analysis) performed on the matrix composed of the normalized clone frequencies of the top 1000 clones present at every time point of the longitudinal data set. The clustering of the data projected on the two first principal components enables us to understand different kinds of dynamics for four clusters of clones here (bottom left). The number of clusters can be adjusted in NoisET and should be tested. In this example, this preanalysis of the longitudinal data set enabled us to find a significant contracting dynamical trend between day 15 to day 85 and a significant expansion trend between day 15 and day 37 following a mild COVID-19 infection[25] (bottom left). The top plots show the individual trajectories in each trend, the bottom plots show the average with standard variation error bars. (c) Scatter plot of contracted clones from day 15 to day 85 after a mild COVID-19 infection.[25] Clones detected as contracting by NoisET are shown in blue. (d) The 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 a SARS-CoV-2 infection between days 15 and 85 post infection;[25] 6 twin donors (S1 through Q2, only β chain) between days 0 and 15 following yellow-fever vaccination;[24] and yellow-fever first (M) and second vaccination (M and P).[26] (e) Venn diagram showing the overlap between the number of called responding TCR clones by both NoisET and edgeR after a mild COVID-19 infection.[25] The z-scores and p-values of the common clones found by both methods and the MIRA database. Plots a−c are standard NoisET output.

3. Detecting Responding Clones. To use the second function to detect responding clonotypes, the user provides, in addition to the two data sets to be compared, two sets of experimental noise parameters learned at both times using the

https://doi.org/10.1021/acs.jpca.2c05002
J. Phys. Chem. A 2022, 126, 7407−7414

first function. When replicates are not available for each time point or donor, a common null model may be used for both time points. 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 time points. 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 detection is allowed. The output is a CSV file containing a table of putative responding clones. The result is illustrated in Figure 1c, which shows contracted clones (purple points) detected from day 15 to day 85 from a mild COVID-19 infection.[25]

Compared with software introduced in ref 20, NoisET allows for conditioning on TCR clones sizes in the analysis, and for using a Poisson or negative binomial distribution for the experimental noise model.

4. Generating trajectories. Using NoisET, one can also generate *in-silico* RepSeq samples, and their neutral dynamics following the stochastic population dynamics developed in ref 27 and in ref 28. The function takes as input the noise model method (negative binomial or Poisson), the noise model parameters at both time points, the number of reads at both time points, the duration of the simulations, and the values of τ and θ describing the global stochastic population dynamics. The neutral dynamics for each clone is defined by

$$\frac{dn}{dt} = \left[-\frac{1}{\tau} + \frac{1}{2\theta} + \frac{1}{\sqrt{\theta}}\eta(t)\right]n(t)$$

where $n(t)$ is the true somatic abundance for a clone belonging to an individual repertoire.

5. Diversity Estimator. Learning the noise model is also helpful for computing diversity estimates which are known to be sensitive to sampling noise.[23] NoisET includes a diversity estimator $D_0 = N_{obs}/(1 - P(\hat{n} = 0, \hat{n}' = 0))$, where N_{obs} is the number of clones observed in both replicates used to learn the experimental noise and $P(\hat{n} = 0, \hat{n}' = 0)$ is the learned fraction of nonsampled clones from the repertoire. This value is expected to be close to 1. Evidently, the larger is N_{obs}, the deeper the sequencing is and so the diversity estimate is expected to be more trustworthy, assuming comparable quality of data generation.

III.B. Applications of NoisET. The method on which NoisET is based has been applied in two published studies identifying clones involved in yellow fever vaccination[24] and SARS-CoV2 responses.[25] In both cases, the analysis was performed on longitudinal TCR RepSeq cDNA data sets and from several different time-points, we were able to identify the peak of the response (expansion or contraction) thanks to the trajectory PCA method[26] now encoded in NoisET. Figure 1d reports the number of responding clonotypes detected by NoisET applied to these data sets, as well as to data from a secondary Yellow-Fever vaccination study.[26]

In the yellow fever vaccination study, TCR repertoires of three pairs of identical twins were sequenced.[24] In each donor, 600 to 1700 responding TCR clones were identified. The TCR response was highly personalized even among twins. Analyzing the clonotypes that the method identified as responding, we were able to show that while the responding TCRs were mostly private, they could be well-predicted using a classifier based on sequence similarity. Using the a posteriori distribution, different types of dynamics were found in different TCR subsets: CD4+ cells contract faster than CD8+ cells.

TCR cDNA-based repertoire response identified groups of CD4+ and CD8+ T cell clones that contract after recovery (∼15 days after the onset of symptoms) from a SARS-CoV-2 infection.[25] A secondary response peak of the response was identified ∼40 days after the onset of symptoms. This secondary peak was also seen in other SARS-CoV2 studies,[29] however it did not correspond to known tetramer probes. Analyzing repertoire data for the same individuals taken a year and two years before the SARS-CoV2 infection, we showed that T-cell clones detected as reacting to SARS-Cov-2 were present one year before the SARS-Cov-2 infection. A network analysis revealed that these pre-existing cells that could confer immunity were specific to a SARS-CoV2 epitope with a one amino acid mutation compared to a common cold coronavirus. This observation raised the question of the correlation between the presence of cross-reactive T cells before infection and mildness of the disease. The detected reactive T-cell clones were also found in memory subpopulations at least three months after the infection.

As mentioned in section 4 the noise learning feature of NoisET has also been used to learn the natural dynamics of TCR repertoires based on gDNA and cDNA data in the absence of direct antigenic stimulation.[28] This study considered the TCRβ repertoires of 9 people and showed that the dynamics of all people, regardless of age is constrained by the power law exponent of the frequency distribution. The exponent itself is given by the ratio of the deterministic turnover time scale and the stochastic noise time scale. The reproducibility of this ratio is a very strong constraint, not directly encoded by the model but learned from the statistics of the data, that implies strong amplitudes of environmental antigenic fluctuations compared to the mean fitness of lymphocyte clones. This parameter regime translates into a very susceptible dynamical system since the mean of the clone size distribution diverges. This property allows the repertoire to maintain a large number of cells, even if the source disappears or becomes very small. While the ratio is constrained, the repertoire turnover time scale shows a strong dependence on the age of the individual, with clear signatures of aging in the physical sense: turnover time scales grow linearly with the biological age of the individual from ∼10 years for 20-year olds to ∼40 years for 60-year olds. This time scale gives us an estimate of how likely we are to find clones in the repertoire after a certain number of years, depending on the person's age.

III.C. Comparison with Existing Software. The need to characterize experimental noise has been well recognized in the sequencing community. EdgeR is a package used to analyze a variety of data produced with HTS (High Throughput Sequencing) that includes read counts.[30] This software has been mostly used for differential gene expression analysis, differential splicing, and bisulfite sequencing. Applied to lymphocyte repertoires, the EdgeR package enables using statistical tests to identify TCR clones expanded after an acute infection.

We compare EdgeR and NoisET detected clones assumed to respond to SARS-Cov2 antigen, based on TCR data from Adaptive Biotechnologies (https://clients.adaptivebiotech.com/pub/covid-2020). We can validate the responding clones using the MIRA data set from the same group for which the reactivity to SARS-Cov2 antigen was validated experimentally

https://doi.org/10.1021/acs.jpca.2c05002
J. Phys. Chem. A 2022, 126, 7407−7414

https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7418734/. To count the overlap between the responding clonotypes called by each software and the TCR MIRA database, we used the AtrieGC software (https://github.com/mbensouda/NoisET_tutorial/), which enables a rapid comparison of the two lists of amino-acids. In order to ascribe statistical significance to our results, we compare the numbers of overlapping TCRs called by EdgeR and NoisET to overlapping TCRs between lists of 1000 randomly sampled clones from the experimental samples and the MIRA list. Given the mean and standard deviation of overlapping clones, we quantify the performance of the two softwares using a z-score. The conclusion is drawn in a Venn diagram in Figure 1e. For this specific task of recognizing SARS-Cov2 TCR clones NoisET (p-value of 5.10^{-13}) performs similarly to edgeR (p-value of 2.410^{-14}) with the benefits of better understanding of the data, better knowledge of the log-fold change statistics, and the possibility to generate synthetic data. We note that the MIRA database is nonexhaustive so both NoisET and edgeR may have identified truly responding SARS-Cov2 TCR clones that are not included in the MIRA database.

IV. DISCUSSION AND CONCLUSIONS

High-throughput sequencing of immune repertoires is poised to revolutionize systems immunology as well as precision medicine. In particular, there is a growing interest in identifying T-cell receptors that respond to acute infections and vaccine challenges, based on experiments that probe repertoires before and after an antigenic challenge. Due to experimental and biological noise, identifying the response simply based on differences in counts before and after the challenge is not reliable. The commonly used solution is to prune these estimates using statistical tests, which are not tailored to account for these specific sources of noise.

In our previous work, we provided a computational method that accounts for the different biological and experimental sources of noise in the clone count measurements in a Bayesian way, allowing for a more reliable detection of expanded or contracted clones. However, while the proof-of-principle algorithm explored the applicability of the method, it did not provide a user-friendly tool, which limits its wide use by the community of immunologists and clinicians. Here, we described a new computational tool, NoisET (https://github.com/statbiophys/NoisET), a python package with a command-line interface that implements the method for characterizing the noise and identifies statistically significant responding clones. The tool is applicable to data sets of diverse origin describing the clonal repertoire response to acute infections and nonstimulated long-term dynamics.

NoisET is designed as an easy-to-use package to learn the noisy statistics of sequence counts and to detect responding clones to a stimulus 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 data sets, choosing and using the adequate statistical null model should be done with caution.

Among the different types of noise models 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.[21] 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.[27,28]

The Bayesian approach encoded in NoisET results in a more reliable way to account for uncertainty than statistical estimates that are also less interpretable. The detection of responding clones based on the fold change of empirical abundances was not optimal without a robust interpretation of the details of the noise model. Errors in noise identification also propagate to erroneous calling of clonotypes.

From a more general perspective, NoisET and the methods behind it combine many years of the study of gene expression noise.[15−19] NoisET strongly exploits the intermittency of mRNA production and the heterogeneity of mRNA counts in individual cells.

As we briefly discussed, NoisET has been applied to identify SARS-CoV-2-specific T-cell receptors and in the future can be used to study and understand the heterogeneity of SARS-Cov-2 vaccine response. It has potential application uncovering responding T-cell receptors to acute infections and vaccine response.

While the method is generally applicable to T cells and B cells,[31,32] due to the somatic hypermutations occurring in B cells upon proliferation, care must be taken when preparing B cell data input and interpreting the model. One possibility is to collapse the sequences into lineages and consider the dynamics of a lineage in the periphery. However, while this is a reasonable first approximation, more work is needed to correctly account for the complexity of B cell repertoires. For this reason we discuss existing applications to T cells. Nevertheless, the conceptual ideas behind noise calibration as implemented in the NoisET software apply.

More broadly, the noise inference using the first module of NoisET has also been used to learn the natural dynamics of T cell repertoires in the absence of specific antigenic stimulation.[28] For all individuals studied we found a universal constraint on the dynamics, which translated into a susceptible dynamical system that can easily maintain a large number of diverse cells. If the same type of dynamics holds for coarse grained B cell repertoires, which remains to be seen, it would point to universal laws that constrain clone size distributions and govern repertoire dynamics.

AUTHOR INFORMATION

Corresponding Authors

Thierry Mora − *Laboratoire de physique de l' École normale supérieure, CNRS, PSL University, Sorbonne Université, and Université de Paris, Paris 75005, France*; Email: tmora@phys.ens.fr

Aleksandra M. Walczak − *Laboratoire de physique de l' École normale supérieure, CNRS, PSL University, Sorbonne Université, and Université de Paris, Paris 75005, France*; orcid.org/0000-0002-2686-5702; Email: awal-czak@phys.ens.fr

Authors

Meriem Bensouda Koraichi − *Laboratoire de physique de l' École normale supérieure, CNRS, PSL University, Sorbonne Université, and Université de Paris, Paris 75005, France*

Maximilian Puelma Touzel − *MILA, University of Montreal, Montreal H2S 3H1, Canada*

https://doi.org/10.1021/acs.jpca.2c05002
J. Phys. Chem. A 2022, 126, 7407−7414

Andrea Mazzolini − *Laboratoire de physique de l' École normale supérieure, CNRS, PSL University, Sorbonne Université, and Université de Paris, Paris 75005, France;* orcid.org/0000-0003-3194-2052

Complete contact information is available at: https://pubs.acs.org/10.1021/acs.jpca.2c05002

Author Contributions

#T.M. and A.M.W. contributed equally.

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

This work was partially supported by the European Research Council Consolidator Grant no. 724208 and ANR-19-CE45-0018 "RESP-REP" from the Agence Nationale de la Recherche.

■ REFERENCES

(1) Robins, H. S.; Campregher, P. V.; Srivastava, S. K.; Wacher, A.; Turtle, C. J.; Kahsai, O.; Riddell, S. R.; Warren, E. H.; Carlson, C. S. Comprehensive assessment of T-cell receptor β-chain diversity in $\alpha\beta$ T cells. *Blood* **2009**, *114*, 4099−4107.

(2) Mora, T.; Walczak, A. M.*Systems Immunology*; CRC Press, 2018; pp 183−198.

(3) Lythe, G.; Callard, R. E.; Hoare, R. L.; Molina-París, C. How Many TCR Clonotypes Does a Body Maintain? *J. Theor. Biol.* **2016**, *389*, 214−224.

(4) DeWitt, W. S.; Lindau, P.; Snyder, T. M.; Sherwood, A. M.; Vignali, M.; Carlson, C. S.; Greenberg, P. D.; Duerkopp, N.; Emerson, R. O.; Robins, H. S. A public database of memory and naive B-cell receptor sequences. *PloS one* **2016**, *11*, No. e0160853.

(5) Weinstein, J. A.; Jiang, N.; White, R. A.; Fisher, D. S.; Quake, S. R. High-throughput sequencing of the zebrafish antibody repertoire. *Science* **2009**, *324*, 807−10.

(6) Boyd, S. D.; Marshall, E. L.; Merker, J. D.; Maniar, J. M.; Zhang, L. N.; Sahaf, B.; Jones, C. D.; Simen, B. B.; Hanczaruk, B.; Nguyen, K. D.; et al. Measurement and clinical monitoring of human lymphocyte clonality by massively parallel VDJ pyrosequencing. *Science translational medicine* **2009**, *1*, 12ra23−12ra23.

(7) Benichou, J.; Ben-Hamo, R.; Louzoun, Y.; Efroni, S. Rep-Seq: uncovering the immunological repertoire through next-generation sequencing. *Immunology* **2012**, *135*, 183−191.

(8) Six, A.; Mariotti-Ferrandiz, M. E.; Chaara, W.; Magadan, S.; Pham, H.-P.; Lefranc, M.-P.; Mora, T.; Thomas-Vaslin, V.; Walczak, A. M.; Boudinot, P. The past, present, and future of immune repertoire biology-the rise of next-generation repertoire analysis. *Frontiers in immunology* **2013**, *4*, 413.

(9) Robins, H. Immunosequencing: applications of immune repertoire deep sequencing. *Current opinion in immunology* **2013**, *25*, 646−652.

(10) Georgiou, G.; Ippolito, G. C.; Beausang, J.; Busse, C. E.; Wardemann, H.; Quake, S. R. The promise and challenge of high-throughput sequencing of the antibody repertoire. *Nature biotechnology* **2014**, *32*, 158−168.

(11) Heather, J. M.; Ismail, M.; Oakes, T.; Chain, B. High-throughput sequencing of the T-cell receptor repertoire: pitfalls and opportunities. *Briefings in bioinformatics* **2018**, *19*, 554−565.

(12) Minervina, A.; Pogorelyy, M.; Mamedov, I. T-cell receptor and B-cell receptor repertoire profiling in adaptive immunity. *Transplant International* **2019**, *32*, 1111−1123.

(13) Rubelt, F.; Busse, C. E.; Bukhari, S. A. C.; Bürckert, J.- P.; Mariotti-Ferrandiz, E.; Cowell, L. G.; Watson, C. T.; Marthandan, N.; Faison, W. J.; Hershberg, U.; et al. Adaptive Immune Receptor Repertoire Community recommendations for sharing immune-repertoire sequencing data. *Nature immunology* **2017**, *18*, 1274−1278.

(14) Barennes, P.; Quiniou, V.; Shugay, M.; Egorov, E. S.; Davydov, A. N.; Chudakov, D. M.; Uddin, I.; Ismail, M.; Oakes, T.; Chain, B.; et al. Benchmarking of T cell receptor repertoire profiling methods reveals large systematic biases. *Nature biotechnology* **2021**, *39*, 236−245.

(15) Elowitz, M. B.; Levine, A. J.; Siggia, E. D.; Swain, P. S. Stochastic gene expression in a single cell. *Science* **2002**, *297*, 1183−1186.

(16) Ozbudak, E. M.; Thattai, M.; Kurtser, I.; Grossman, A. D.; van Oudenaarden, A. Regulation of noise in the expression of a single gene. A. *Nat. Genet.* **2002**, *31*, 69−73.

(17) Cai, L.; Friedman, N.; Xie, X. S. Stochastic protein expression in individual cells at the single molecule level. *Nature* **2006**, *440*, 358−362.

(18) Taniguchi, Y.; Choi, P. J.; Li, G.-W.; Chen, H.; Babu, M.; Hearn, J.; Emili, A.; Xie, X. S. Quantifying E. coli proteome and transcriptome with single-molecule sensitivity in single cells. *Science* **2010**, *329*, 533−538.

(19) Hornos, J. E.; Schultz, D.; Innocentini, G. C.; Wang, J.; Walczak, A. M.; Onuchic, J. N.; Wolynes, P. G. Selfregulating gene: an exact solution. *Phys. Rev. E* **2005**, *72*, 051907.

(20) Puelma Touzel, M.; Walczak, A. M.; Mora, T. Inferring the immune response from repertoire sequencing. *PLOS Computational Biology* **2020**, *16*, No. e1007873.

(21) Balachandran, V. P.; Luksza, M.; Zhao, J. N.; Makarov, V.; Moral, J. A.; Remark, R.; Herbst, B.; Askan, G.; Bhanot, U.; Senbabaoglu, Y.; et al. Identification of unique neoantigen qualities in long-term survivors of pancreatic cancer. *Nature* **2017**, *551*, 512−516.

(22) Rytlewski, J.; Deng, S.; Xie, T.; Davis, C.; Robins, H.; Yusko, E.; Bienkowska, J. Model to improve specificity for identification of clinically-relevant expanded T cells in peripheral blood. *PLoS One* **2019**, *14*, No. e0213684.

(23) Mora, T.; Walczak, A. M. How many different clonotypes do immune repertoires contain? *Current Opinion in Systems Biology* **2019**, *18*, 104−110.

(24) Pogorelyy, M. V.; Minervina, A. A.; Touzel, M. P.; Sycheva, A. L.; Komech, E. A.; Kovalenko, E. I.; Karganova, G. G.; Egorov, E. S.; Komkov, A. Y.; Chudakov, D. M.; Mamedov, I. Z.; Mora, T.; Walczak, A. M.; Lebedev, Y. B. Precise tracking of vaccine-responding T cell clones reveals convergent and personalized response in identical twins. *Proc. Natl. Acad. Sci. U. S. A.* **2018**, *115*, 12704−12709.

(25) Minervina, A. A.; Komech, E. A.; Titov, A.; Koraichi, M. B.; Rosati, E.; Mamedov, I. Z.; Franke, A.; Efimov, G. A.; Chudakov, D. M.; Mora, T.; et al. Longitudinal high-throughput TCR repertoire profiling reveals the dynamics of T-cell memory formation after mild COVID-19 infection. *Elife* **2021**, *10*, No. e63502.

(26) Minervina, A. A.; Pogorelyy, M. V.; Komech, E. A.; Karnaukhov, V. K.; Bacher, P.; Rosati, E.; Franke, A.; Chudakov, D. M.; Mamedov, I. Z.; Lebedev, Y. B.; et al. Primary and secondary anti-viral response captured by the dynamics and phenotype of individual T cell clones. *Elife* **2020**, *9*, No. e53704.

(27) Desponds, J.; Mora, T.; Walczak, A. M. Fluctuating fitness shapes the clone-size distribution of immune repertoires. *Proc. Natl. Acad. Sci. U. S. A.* **2016**, *113*, 274−279.

(28) Koraichi, M. B.; Ferri, S.; Walczak, A. M.; Mora, T. Inferring the T-cells repertoire dynamics of healthy individuals. *bioRxiv* **2022**, 490247.

(29) Weiskopf, D.; Schmitz, K. S.; Raadsen, M. P.; Grifoni, A.; Okba, N. M.; Endeman, H.; van den Akker, J. P.; Molenkamp, R.; Koopmans, M. P.; van Gorp, E. C.; et al. Phenotype and kinetics of SARS-CoV-2-specific T cells in COVID-19 patients with acute respiratory distress syndrome. *Science immunology* **2020**, *5*, No. eabd2071.

(30) Robinson, M. D.; McCarthy, D. J.; Smyth, G. K. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *bioinformatics* **2010**, *26*, 139−140.

(31) Altan-Bonnet, G.; Mora, T.; Walczak, A. M. Quantitative immunology for physicists. *Phys. Rep.* **2020**, *849*, 1−83.

(32) Chakraborty, A. K.; Košmrlj, A. Statistical mechanical concepts in immunology. *Annu. Rev. Phys. Chem.* **2010**, *61*, 283−303.

Recommended by ACS

EPIFANY: A Method for Efficient High-Confidence Protein Inference

Julianus Pfeuffer, Oliver Kohlbacher, *et al.*

JANUARY 24, 2020
JOURNAL OF PROTEOME RESEARCH READ

Inflect: Optimizing Computational Workflows for Thermal Proteome Profiling Data Analysis

Neil A. McCracken, Amber L. Mosley, *et al.*

MARCH 04, 2021
JOURNAL OF PROTEOME RESEARCH READ

Systematic Partitioning of Proteins for Quantum-Chemical Fragmentation Methods Using Graph Algorithms

Mario Wolter, Christoph R. Jacob, *et al.*

FEBRUARY 16, 2021
JOURNAL OF CHEMICAL THEORY AND COMPUTATION READ

New Interface for Faster Proteoform Analysis: Immunoprecipitation Coupled with SampleStream-Mass Spectrometry

Henrique dos Santos Seckler, Neil L. Kelleher, *et al.*

MAY 27, 2021
JOURNAL OF THE AMERICAN SOCIETY FOR MASS SPECTROMETRY READ

Get More Suggestions >

https://doi.org/10.1021/acs.jpca.2c05002
J. Phys. Chem. A 2022, 126, 7407−7414